Amin Modabberian*, Hoang Khac Nguyen, and Kai Zenger

Mean Value Modeling of Maritime Diesel Engines

Abstract: This work aims to develop a first principles mean value model that can be further used for advanced control design on different types of maritime diesel engines to increase efficiency and reduce fuel consumption. A nonlinear model is initially derived and linearised for a control design. These models are simulated and results are compared for further control design approaches.

Keywords: diesel engines, modeling, mean value modeling

***Corresponding Author: Amin Modabberian:** Aalto University, E-mail: amin.modabberian@aalto.fi
Hoang Khac Nguyen: Aalto University, E-mail: hoang.kh.nguyen@aalto.fi
Kai Zenger: Aalto University, Email: kai.zenger@aalto.fi

1 Introduction

As the European emission standards have become more stringent, diesel engines are becoming obsolete especially within the automotive industry [1]. However, diesel engines will be used in maritime transportation for the upcoming decades. Emission minimization of NO_x, CO_2 and particulate matter is one of the major areas in maritime diesel engine control applications. Adhering to emission conditions set by International Maritime Organization (IMO) [2], current diesel engines requires efficient control, which itself require accurate modeling. In this work, a first principles mean value model for diesel engine airpath and engine dynamics are presented, that is, creating a nonlinear and linear models to examine the state variables of the system, which are intake and exhaust manifold pressures, compressor power and engine speed. The airpath model consists of intake and exhaust manifold pressures, turbocharger power, engine speed and fuel injection ratio. Each dynamics is modeled and simulated separately utilizing MATLAB's Simulink. Some of the models are achieved empirically due the nonlinearity of system. On the other hand, such nonlinear models are required for mapping complex chemical and combustion reactions. Mean value model itself is a valid approach. However, it is not suitable for cylinder wise control, since it does not take the pulsating nature of the engine airpath in to account [3][4].

In general, modeling does not possess a single way approach for all dynamic systems. However, numerous valid and widely used approaches exist in the literature[5–9]. Each system and its features, such as physical quantities mentioned above, requires a careful and a systematic approach, which affects the control design and behavior of controls. In other words, the desired efficiency in control design of the engine is highly dependent on the generated mathematical models of the system. Mean value models are also part of the more accurate models to be developed for cylinder-wise control of the engine, which is a key challenge nowadays and in the years to come [10].

The developed mean value model is presented in section 2. Control design and simulation results are presented in sections 3 and 4, respectively.

2 Modelling

The mathematical model in this work is based on the VEBIC's engine provided by Wärtsilä [11]. The topology is depicted in Figure 1. The engine airpath is controlled by turbocharger, which consists of turbine and compressor. The compressor increases the incoming air pressure to a desired level in addition to increasing the temperature levels. The air temperature is then decreased with inter-cooler to best suit the combustion reaction. The amount of air mass in the cylinders depends on the fuel injection. Therefore, the intake manifold pressure has to be controlled.

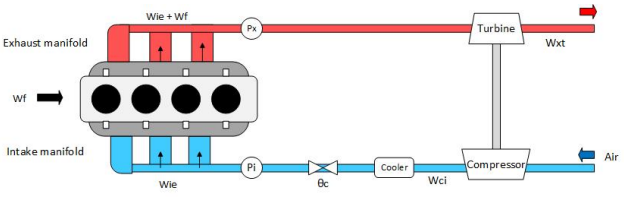


Fig. 1. Topology of the diesel engine

The injected fuel is mixed with heated air mass in cylinder and ignited with a spark. This reaction pushes the piston down. The piston is attached to the crankshaft. Hence, the lateral motion of the piston is converted to rotational motion of the crankshaft. The piston pushes the exhaust gas out of the cylinder toward the turbine. The exhaust gas causes the turbine blades to spin. Since the turbine is attached to the compressor with a shaft, it is possible to control the turbine with the compressor or the intake manifold pressure with turbine speed.[12]

2.1 Airpath model

Mass flows

The conversation of air mass is equal to mass flows through the intake and exhaust manifold [13]. The air mass flow through intake manifold is the difference between the air mass flows into the compressor, W_{ci} $(Kg \cdot s^{-1})$, and cylinders, W_{ie}

$$\dot{m}_i = W_{ci} - W_{ie} \tag{1}$$

The air mass flow though the exhaust manifold is the difference between exhaust mass flow into the turbine, W_{xt}, and sum of air mass flow into the cylinders with fuel mass flow into the cylinders, W_f,

$$\dot{m}_x = W_{ie} + W_f - W_{xt} \tag{2}$$

Manifold pressure

The pressures of intake and exhaust manifolds can be presented with the first law of thermodynamics and ideal gas law [13]. Intake manifold pressure is thus defined as

$$\dot{p_i} = \frac{R_i T_i}{V_i} \dot{m}_i \tag{3}$$

and exhaust manifold pressure is defined as

$$\dot{p_x} = \frac{R_x T_x}{V_x} \dot{m}_x \tag{4}$$

where, R $(J \cdot mol^{-1} \cdot K^{-1})$, T (K) and V (m^3) are the intake and exhaust manifold gas constant, temperature and volume respectively. By substituting (1) and (2) into (3) and (4) respectively, the pressure of intake and exhaust manifold become

$$\dot{p_i} = \frac{R_i T_i}{V_i} (W_{ci} - W_{ie}) \tag{5}$$

$$\dot{p_x} = \frac{R_x T_x}{V_x} (W_{ie} + W_f - W_{xt}) \tag{6}$$

Air flow through the cylinders can be presented as a function of engine speed, ω_e $(rad \cdot s^{-1})$, volumetric efficiency of the engine, η_v, and intake manifold pressure, p_i (Pa)

$$W_{ie} = \frac{\eta_v(\omega_e)\omega_e p_i V_d}{2\pi v R_i T_i} \tag{7}$$

where, V_d (m^3) is engine displacement volume and v is number of revolution per engine cycle. η_v is as defined in [14] as

$$\eta_v = a_{v1} + a_{v2}\omega_e + a_{v3}\omega_e^2$$

Values of constants a_{v1}, a_{v2} and a_{v3} are approximated.

The exhaust gas flow through the turbine can be presented with the orifice equation

$$W_{xt} = \frac{p_x}{\sqrt{R_x T_x}} A_t \Psi\left(\frac{p_a}{p_x}\right) f(u_{uvg}) \tag{8}$$

where, A_t (m^2) is the area of turbine nozzle and Ψ is the pressure ratio correction factor, which is defined as

$$\Psi\left(\frac{p_a}{p_x}\right) = \begin{cases} \sqrt{\frac{2\gamma}{\gamma-1}\left(\left(\frac{p_a}{p_x}\right)^{\frac{2}{\gamma}} - \left(\frac{p_a}{p_x}\right)^{\frac{\gamma+1}{\gamma}}\right)} & \text{, if } \frac{p_a}{p_x} > \left(\frac{2}{\gamma+1}\right)^{\frac{\gamma}{\gamma-1}} \\ \gamma^{\frac{1}{2}}\left(\frac{2}{\gamma+1}\right)^{\frac{\gamma+1}{2(\gamma-1)}} & \text{, if } \frac{p_a}{p_x} \leq \left(\frac{2}{\gamma+1}\right)^{\frac{\gamma}{\gamma-1}} \end{cases}$$

Here, γ is specific heat ratio and it is defined as $\gamma = c_p/c_v$, where c_p $(J \cdot Kg^{-1}K^{-1})$ is specific heat at constant pressure and c_v $(J \cdot Kg^{-1}K^{-1})$ is specific heat at constant volume and p_a is ambient pressure. $f(u_{vgt})$ is the control signal and it is defined as

$$f(u_{vgt}) = b_1 + b_2 u + b_3 u^2 \tag{9}$$

where u_{vgt} is the control signal of variable geometry turbocharger (VGT), and b_1, b_2 and b_3 are approximated parameters [4]. Mass flow through the compressor is function the compressor isentropic efficiency η_c

$$W_{ci} = \frac{\eta_c P_c}{c_p T_a \left(\left(\frac{p_i}{p_a}\right)^{\mu} - 1\right)} \tag{10}$$

where T_a is ambient temperature and $\mu = (\gamma - 1)/\gamma$.

Power

Power of turbocharger is defined as power transferred between the turbine, P_t (kW), and compressor, P_c

$$\dot{P_c} = \frac{1}{\tau_c}(\eta_m P_t - P_c) \tag{11}$$

where, τ_c (s) is turbocharge time constant and η_m is the turbine mechanical efficiency. P_t is function of exhaust gas mass flow, ambient and exhaust pressures

$$P_t = \eta_t W_{xt} c_p T_x \left(1 - \left(\frac{p_i}{p_a}\right)^{\mu}\right) \tag{12}$$

where T_x presented as function of air-to-fuel ratio(λ)[14]

$$T_x = T_i + a_{t1}\lambda^{a_{t2}} + a_{t3} \tag{13}$$

Here, a_{t1-3} are tuning coefficients.

2.2 Fuel-path model

The mean acceleration of the crankshaft of the engine can be described with the second law of Newton

$$J\dot{\omega}_e = \Sigma M \tag{14}$$

where J $(Kg{\cdot}m^2)$ is the moment of inertia of the engine and ΣM is sum torques effecting the crankshaft. These torques are indicated engine torque, M_e (Nm), friction torque, M_f, and load torque, M_l.

The equation of indicated engine torque is modeled under the assumption that indicated thermal efficiency is depended on engine speed and air-to-fuel ratio, λ,

$$M_e = \frac{W_f Q_{hv}}{\omega_e}\eta_i(\omega_e, \lambda) \tag{15}$$

where Q_{hv} $(MJ{\cdot}kg^{-1})$ is lower heating value of the fuel, W_f is the fuel mass flow and $\eta_i = (a_1 + a_2\omega_e + a_3\omega_e^2)(1 - a_4\lambda^{a_5})$ [14]. W_f is used as a control signal.

Friction torque M_f is caused by phenomena such as friction of crankshaft bearings or resistance between the piston rings and cylinder wall and is proportional to the friction mean effective pressure, f_{mep} (Pa), [15]

$$M_f = \frac{f_{mep} V_d}{2\pi v} \tag{16}$$

where f_{mep} is defined as

$$f_{mep} = C_1 + 48\frac{N_e}{1000} + 0.4S_P^2$$

Here, C_1 (Pa) is a constant, N_e (RPM) is the engine speed and S_p $(m{\cdot}s^{-1})$ is the mean piston speed. The load torque can be defined as an external input. Thus, the mean acceleration of the crankshaft becomes

$$J\dot{\omega}_e = M_e - M_f - M_l \tag{17}$$

2.3 Linearization and analysis

A linearised model is needed to better understand the behavior of the model around operating points $(p_{i0}, p_{x0}, P_{c0}, \omega_{e0},\ f(u_{vgt0}))$. The mean value model of this work is linearised utilizing the Taylor's approximation. Detailed calculation have been left out for clarification.

Intake manifold pressure

$$\begin{aligned}\Delta\dot{p}_i = \frac{R_i T_i}{V_i}\Bigg[&\Bigg(-\frac{\eta_c P_{c0}\mu(\frac{p_{i0}}{p_a})^{\mu-1}(\frac{1}{p_a})}{c_p T_a((\frac{p_{i0}}{p_a})^{\mu}-1)^2} \\ &- \frac{\eta_v(\omega_{e0})\omega_{e0}V_d}{2\pi v R_i T_i}\Bigg)\Delta p_i \\ &+ \left(\frac{\eta_c}{c_p T_a((\frac{p_{i0}}{p_a})^{\mu}-1)}\right)\Delta P_c \\ &+ \left(\frac{p_{i0}V_d}{2\pi v R_i T_i}(a_{v1} + 2a_{v2}\omega_{e0} + 3a_{v3}\omega_{e0}^2)\right)\Delta\omega_e\Bigg]\end{aligned} \tag{18}$$

Exhaust manifold pressure

$$\begin{aligned}\Delta\dot{p}_x = \frac{R_x T_x}{V_x}\Bigg[&\left(-\frac{A_t\frac{d\Psi}{dp_{x0}}(\frac{p_a}{p_{x0}})}{\sqrt{R_x T_x}}f(u_{vgt0})\right)\Delta p_x \\ &+ \left(\frac{\eta_v(\omega_{e0})\omega_{e0}V_d}{2\pi v R_i T_i}\right)\Delta p_i \\ &+ \left(\frac{p_{i0}V_d}{2\pi v R_i T_i}(a_{v1} + 2a_{v2}\omega_{e0} + 3a_{v3}\omega_{e0}^2)\right)\Delta\omega_e \\ &+ \Delta W_f + \left(-\frac{p_{x0}A_t\Psi(\frac{p_a}{p_{x0}})}{\sqrt{R_x T_x}}\right)\Delta f(u_{vgt})\Bigg]\end{aligned} \tag{19}$$

Compressor power

$$\begin{aligned}\Delta\dot{P}_c = \frac{1}{\tau}\Bigg[&-\Delta P_c \\ &+ \Bigg(\frac{\eta_m\eta_t A_t\frac{d\Psi}{dp_{x0}}(\frac{p_a}{p_{x0}})c_p T_x}{\sqrt{R_x T_x}}f(u_{vgt0}) \\ &\left(\mu(\frac{p_a}{p_{x0}}^{\mu}) + (1 - (\frac{p_a}{p_{x0}})^{\mu})\right)\Bigg)\Delta p_x \\ &+ \eta_m\eta_t\frac{p_{x0}A_t\Psi(\frac{p_a}{p_{x0}})}{\sqrt{R_x T_x}}c_p T_x\left(1 - (\frac{p_a}{p_{x0}})^{\mu}\right)\Delta f(u_{vgt})\Bigg]\end{aligned} \tag{20}$$

Engine speed

$$\Delta\dot{\omega}_e = \frac{1}{J}\Bigg[\Bigg(\frac{(a_2+2a_3\omega_{e0})(1-a_4\lambda^{a_5})\omega_{e0}W_{f0}Q_{hv}}{\omega_{e0}^2} - \frac{(a_1+a_2\omega_{e0}+a_3\omega_{e0}^2)(1-a_4\lambda^{a_5})W_{f0}Q_{hv}}{\omega_{e0}^2}\Bigg)\Delta\omega_e + \Bigg(\frac{(a_1+a_2\omega_{e0}+a_3\omega_{e0}^2)(1-a_4\lambda^{a_5})Q_{hv}}{\omega_{e0}}\Delta W_f\Bigg) + \Delta load\Bigg] \quad (21)$$

With (18)-(21), the state-state representation of the linearised MIMO-system (Multiple Input, Multiple Output) becomes

$$\begin{aligned}\dot{x}(t) &= A\mathbf{x}(t)_{4\times1} + B\mathbf{u}(t)_{2\times1} + E\mathbf{d}(t)_{1\times1}\\ y(t) &= C\mathbf{x}(t)_{4\times1} + D\mathbf{u}(t)_{2\times1}\end{aligned} \quad (22)$$

where $A \in \mathbb{R}^{4\times4}$, $B \in \mathbb{R}^{4\times2}$. $E \in \mathbb{R}^{4\times1}$, $C \in \mathbb{R}^{2\times4}$ and $D \in \mathbb{R}^{2\times2}$. E is external disturbance, which in the case of this systems is the engine load torque.

3 Control design

Before designing a proper controller, two-way interaction of the MIMO-system has be to analyzed. Ideal case is to have as minimal internal coupling as possible. Such features of the system can be examined with the relative gain array (RGA) method [16]

$$RGA(G(s)) = \Lambda(G(s)) \triangleq G(s) \times (G(s)^{-1})^T \quad (23)$$

where $G(s)$ is the transfer function matrix of the system and $\times$ is the Hadamard product (point-by-point product). In this work the transfer function of the system is

$$G(s) = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s)\end{bmatrix}, \Lambda(G(s)) = \begin{bmatrix}\varepsilon & 1\\ 1 & 0\end{bmatrix} \quad (24)$$

Although the engine speed is one of the inputs for intake manifold pressure as shown in (18), there is only a minor coupling between the channels. Value of ε is close to zero. The Λ-matrix is anti-diagonal due to order of the states, when defining the transfer function.

The results of RGA-analysis enable the separate control design of pressure and engine speed. In this work a separate PID-controller was utilized for each physical quantity in both nonlinear and linear model. The controller outputs are fuel injection and control of variable geometry turbocharger u_{vgt}. Similar results in both nonlinear and linear model mean that advanced control systems can be applied on the linearised system.

4 Simulation results

4.1 Nonlinear and linear open-looped system

The open-loop airpath and fuel-path models were initially simulated without controllers using step functions for inputs W_f and u with MATLAB's Simulink. According to the RGA-analysis, there is little coupling between the engine speed and intake manifold pressure. This can be seen in Figures 2 and 3, where increase of u at time 200 ms does not have any effect on the engine speed, and increase of W_f at 400 ms can be seen as a small peak in pressure curve. Major changes in the curves are caused by increase of engine load at 600 ms. All models

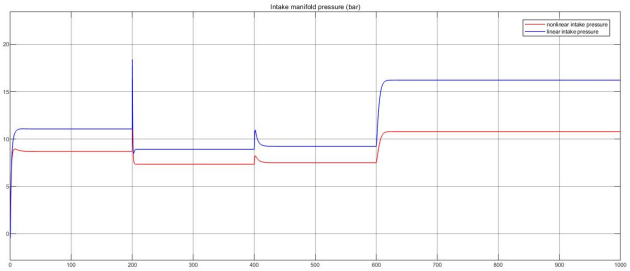


Fig. 2. Nonlinear (red) and linear (blue) intake manifold pressure of an open-looped system.

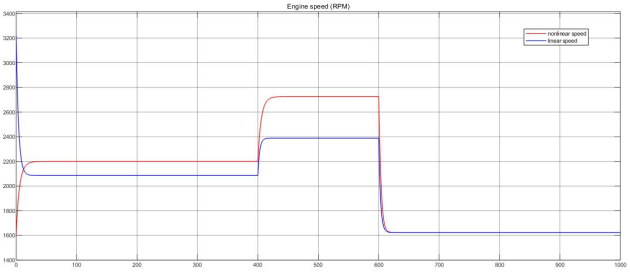


Fig. 3. Nonlinear (red) and linear (blue) engine speed of an open-loop system.

are Lyapunov stable as well as BIBO-stable (Bounded Input, Bounded Output). The deviation between the final values of nonlinear and linear pressure curves means that the models have different operating points.

4.2 Nonlinear controlled system

Simulation results of the nonlinear airpath and engine speed model are depicted in Figures 4 and 5. Both models are controlled with separate PID-controllers, where the input signals are W_f and u_{vgt} for airpath and engine speed model respectively. Intake manifold pressure

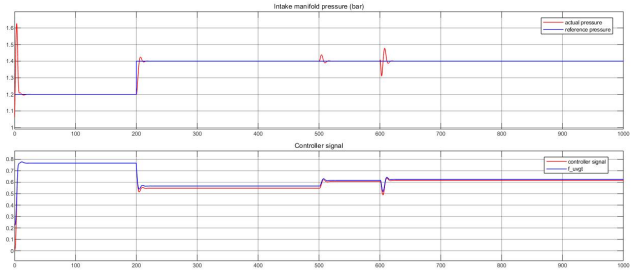


Fig. 4. Nonlinear intake manifold pressure (red) with respect to controller signal.

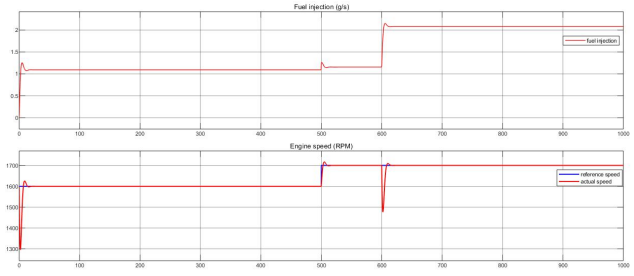


Fig. 5. Nonlinear engine speed (red) with respect to fuel injection.

is inversely proportional to the control signal as shown in Figure 4 at 200 ms. Increase of speed and external engine load can be seen as small peaks at 500 ms and 600 ms respectively. The increase of intake manifold pressure does not affect the engine speed. Increase of load causes the speed to drop, which is stabilized by fuel injection. Both models maintain the desired values despite the changes and disturbances affecting the system.

4.3 Linear controlled system

Simulation results of linearised, closed-loop intake manifold pressure and engine speed system are presented in Figures 6 and 7 respectively. As observed in section

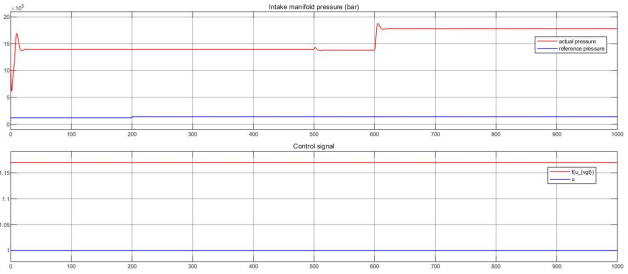


Fig. 6. Linear intake manifold pressure with respect to controller signal.

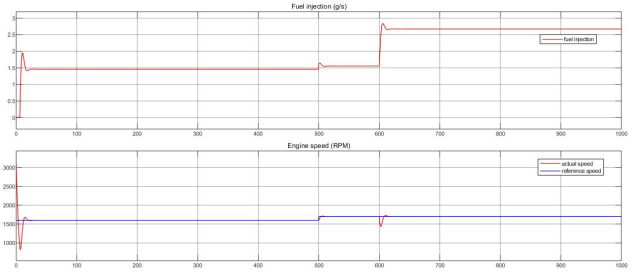


Fig. 7. Linear engine speed with respect to fuel injection.

4.1, the airpath model is strongly dependent on operating points. The controller signal is at its peak during the entire simulation time and for this reason the pressure does not drop to its reference value. This means that the linear airpath model has different operating point than the nonlinear model. Engine speed is stabilized within a short time and follows the reference value well, which is seen as overlapping curves in Figure 7. Major difference to the nonlinear engine speed model is a slight increase of fuel injection at 600 ms.

5 Conclusions

The purpose of this work was to design a first principles mean value model for the airpath and fuel-path of a maritime diesel engine to reduce fuel consumption and thus, increase the efficiency. To achieve this, a nonlinear and linearised model were simulated with PID-controllers.

Initially, simulations were performed without any controllers and the results indicated that the outcome of airpath model depended heavily on the operating points. This problem occurred again in the closed-loop simulations. Nonlinear models provided sufficient results. However, advanced control approaches cannot be performed on nonlinear systems.

Proper operating points can be found with methods, such as pole placement or numerical approximation. This is a challenging problem due to complexity of the simulated systems. Only then, the model can be further developed with other control algorithms and utilized for more accurate models used for cylinder-wise control the diesel engines.

It will be left as an open question, how this situation should be observed and the problem removed.

6 Acknowledgement

In doing this work the co-operation with Wärtsilä Oyj Apb is greatly appreciated.

References

[1] European Commission. Emission in the automotive sector. URL https://ec.europa.eu/growth/sectors/automotive/environment-protection/emissions_en.

[2] International Maritime Organization. Nitrogen oxides (nox) - regulation 13. URL http://www.imo.org/en/OurWork/Environment/PollutionPrevention/AirPollution/Pages/Nitrogen-oxides-(NOx)-%E2%80%93-Regulation-13.aspx.

[3] J. Aalto-Setälä. Adaptive wastegate control of combustion engines, 2013.

[4] S. Samokhin. *Adaptive control of conventional and hybrid marine diesel engines subject to uncertain or time-varying dynamics*. Phd thesis, Aalto University, 2018.

[5] A. G. Stefanopoulou, I. Kolmanovsky, and J. S. Freudenberg. Control of variable geometry turbocharged diesel engines for reduced emissions. In *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No.98CH36207)*, volume 3, pages 1383–1388 vol.3, June 1998. 10.1109/ACC.1998.707043.

[6] M. Jankovic and I. Kolmanovsky. Robust nonlinear controller for turbocharged diesel engines. In *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No.98CH36207)*, volume 3, pages 1389–1394 vol.3, June 1998. 10.1109/ACC.1998.707047.

[7] M. Jung and K. Glover. Calibratable linear parameter-varying control of a turbocharged diesel engine. *IEEE Transactions on Control Systems Technology*, 14(1):45–62, Jan 2006. ISSN 1063-6536. 10.1109/TCST.2005.860513.

[8] A. Plianos and R. Stobart. Dynamic feedback linearization of diesel engines with intake variable valve actuation. In *2007 IEEE International Conference on Control Applications*, pages 455–460, Oct 2007. 10.1109/CCA.2007.4389273.

[9] L. Kocher, E. Koeberlein, K. Stricker, D. G. Van Alstine, B. Biller, and G. M. Shaver. Control-oriented modeling of diesel engine gas exchange. In *Proceedings of the 2011 American Control Conference*, pages 1555–1560, June 2011. 10.1109/ACC.2011.5991425.

[10] Gabriel Turesson. *Model-Based Optimization of Combustion-Engine Control*. PhD thesis, Lund University, 06 2018.

[11] Vebic's engine laboratory starts operating. URL https://www.univaasa.fi/en/news/vebicin_moottorilaboratorio_kaynnistyi/.

[12] L. Guzzella and C.H. Onder. *Introduction to Modeling and Control of Internal Combustion Engine Systems*. Number 2nd ed. Springer-Verlag Berlin Heidelberg, 2010. ISBN 978-3-642-10775-7.

[13] Johan Wahlström and Lars Eriksson. Modeling of a diesel engine with vgt and egr capturing sign reversal and non-minimum phase behaviors. Technical Report 2882, Linköping University, Vehicular Systems, 2009. URL http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-18484.

[14] L. Guzzella and A. Amstutz. Control of diesel engines. *IEEE Control Systems Magazine*, 18(5):53–71, Oct 1998. ISSN 1066-033X. 10.1109/37.722253.

[15] J.B. Heywood. *Internal Combustion Engine Fundamentals*. Automotive technology series. McGraw-Hill, 1988. ISBN 9780071004992. URL https://books.google.fi/books?id=O69nQgAACAAJ.

[16] Sigurd Skogestad and Ian Postlethwaite. *Multivariable Feedback Control: Analysis and Design*. John Wiley & Sons, Inc., USA, 2005. ISBN 0470011688.

Mingzhang Wu, Janne Koljonen and Timo Mantere*

Addressing Resource Allocation Issues in Cloud Computing Environment with Ant Colony Optimization

Abstract: Cloud computing is a fast growing and attractive paradigm in information technology, since it allows using resources on-demand wherever and whenever needed. The use of dynamic cloud resource allocation allows immediate accommodation to unpredictable demands and improvement in the return on investment as for the computing infrastructure. The cloud resources allocation optimization model is one of the core parts in cloud computing. However, despite the recent growth of the research in the cloud computing area, several problems with the process of resource allocation remain unaddressed. Cost and performance are two important but contradictive objectives in the cloud resources allocation process. Cost-performance trade-off constitutes a challenging multi-objective optimization problem in cloud resources allocation. In this paper, a new optimization model is proposed to solve this multi-objective optimization problem effectively. An ant colony optimization algorithm that optimizes the Quality of Service (QoS) and the response time in a simulated CloudSim environment that models five servers of varying characteristics. Experimental results demonstrate the effectiveness of the designed algorithms. Ant colony algorithm shows mostly higher performance than the round robin and greedy assignment algorithms that were used as benchmarks.

Keywords: ant colony optimization, cloud computing, CloudSim, cost-performance, resource allocation, trade-off problem

***Corresponding Author: Timo Mantere:** University of Vaasa, School of Technology and Innovations, E-mail: timo.mantere@uva.fi

Mingzhang Wu: E-mail: mingzhang.wu@gmail.com

Janne Koljonen: University of Vaasa, School of Technology and Innovations, E-mail: janne.koljonen@uva.fi

1 Introduction

1.1 Cloud computing

In the recent years, information technology (IT) has been integrated into our daily life more and more. The major applications are build up based on network and internet technologies. We are now in an era of “big data” with rapid growth on the number of transactions, information, and data. However, low cost, fast speed, and efficient computing are desired. The traditional network and local computation capacity are unable to meet these needs. Instead, distributed network technologies are developed to enable the utilization of distributed computing resources from the internet. How to integrate and distribute the resources, such as servers, over the internet give new research topics to be considered.

Cloud computing, as a new emerging information and communication technology concept, has been an interesting topic recently. There are many definitions of cloud computing. Cloud computing is a result of the convergence of several technologies, such as, (1) hardware, (2) internet technologies, (3) distributed computing, and (4) systems management. The main advantage of cloud computing is providing computing resources based on the public utility model (compare to water, electricity, gas, and telephony) to enhance reliability, scalability, and performance [1].

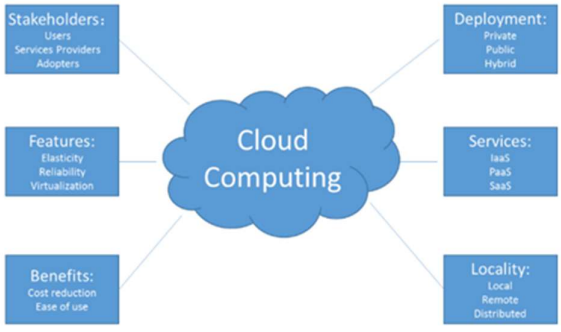


Fig. 1. A holistic view of cloud computing [2]

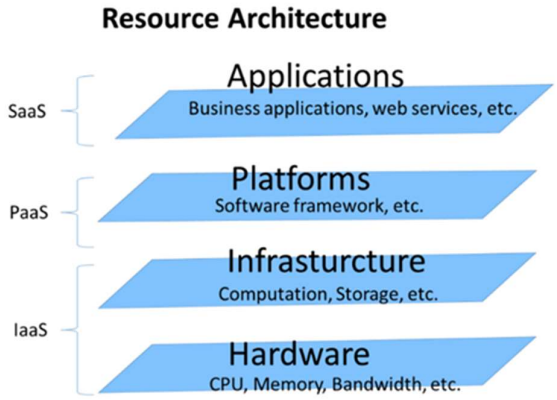


Fig. 2. Network topology of virtual resources in cloud computing

From the technical perspective, cloud computing is the integration of many aspects of technologies, such as (1) virtualization, (2) utility computing, and (3) distributed computing, among others.

From the business perspective, cloud computing is a new business model. It enables (1) sharing information among users, (2) buying resources on-demand without large investments, (3) selling capacity to many users, and subsequently (4) improving the return on investments due to better rates of capacity use. Furthermore, (5) investing on the latest, high-performance infrastructure should give a business advantage to the service provider.

Cloud computing model has changed and will affect many companies' business model and operational status, not only for the IT industry. Cloud computing can provide three types of services: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). A summary of cloud computing properties are given in Figs. 1 and 2.

The National Institute of Standards (NIST) has emphasized the elasticity feature of computing resources in their definition of cloud computing [3], which is largely accepted and frequently cited. It defines cloud as follows: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." The terminology used in this study is clarified in Table 1.

This article is based in major parts on the first author's master's thesis [4].

1.2 Ant colony optimization

Ant Colony Optimization (ACO) is a heuristics proposed by Marco Dorigo [5] in the early 1990s. It was inspired by the observation of the collaboration activities of ants searching for food. Ants can gradually find out the shortest path between a food source and the nest of the colony. ACO is suitable for many optimization problems that can be modeled as a graph, including resource assignment. However, in its original form, ACO modifies the edges of a graph, not the nodes as will be done in this paper.

Table 1. Terminologies used in this research

Cloud	IT cloud consisting of physical resources. It is managed by the cloud services providers.
Cloud services providers	They fully control the resources and how requests are assigned by customers. They are the entity of resources allocation assignment.
Customers	They are users of cloud and the requesting entity. In the context of the algorithm, the customer is also referred to as an ant.
Requests	The requests for resources are specifically virtual machines (VM). The customer may have preferences where its request should be allocated. The final resource allocation, however, is determined by the cloud services providers.

The weakness of many optimization methods is their inability to handle more than one objective. In addition, these methods often employ local and greedy (e.g., hill-climbing) approaches, which are prone to find only a local optimum [6]. Instead, ACO is considered as a part of the family of evolutionary algorithms that use multiple parallel trials and stochastic search to improve the probability to find the global optimum. How ACO is implemented in this study, is reported in Section 3.

ACO has previously been applied successfully to a number of benchmark combinatorial optimization problems (see Section 2 for more details). In this study, it is proposed how to use ACO to solve the multi-objective model for Cost-Performance trade-off problem (CPTOP). The concept of ACO-based multi-objective CPTOP model is designed and tested using *CloudSim*, which is an extendable discrete-event simulation toolkit that enables modeling and simulation of cloud computing environments and the application-provisioning environment.

2 Related work

Resource allocation for clouds has been studied with a very wide scope in the literature. The problem of determining an optimal allocation of the requests to a pool of resources is NP-hard (non-deterministic polynomial-time hard) problem. Nevertheless, many optimization strategies may be used to solve it efficiently. In particular, several heuristic algorithms have been proposed by researchers for optimal allocation of cloud resources [7].

Fidanova [8] proposed an adaptive resource allocation algorithm in cloud computing environment.

That paper used adaptive min-min scheduling and list scheduling, but it was used in a static manner [9].

In Foster et al. [10], the authors proposed an optimal virtual machine placement algorithm for minimizing the cost that cloud customer have to pay to the cloud service provider, when they need virtual machine from cloud computing environment access as a part of the cloud service. In Chaisiri et al. [11], the authors described the multi-objective mechanism for scheduling applications that take various cost constraints and the availability of resources into account.

In Frincu and Craciun [12], the focus is more on the resource allocation strategy in selecting the cloud provider, but the approach is static as for selecting the data center from the distributed environment where the global data center is available, with taking care of timing parameter as in [9]. Chimakurthi [7] propose an energy-efficient mechanism to minimize the number of servers to be used for hosting the services and allocating the cloud resources to the applications.

The paper by Hua et al. [13] propose an ant colony optimization algorithm for resource allocation, in which all the characteristics in cloud are considered. It has been compared with a genetic algorithm and a simulated annealing algorithm, proving that it is suitable for computing resource search allocation in cloud computing environment.

Omara et al. [14] propose an optimization solution to the allocation of shared resources to minimize the estimated cost and enhance virtual machine configuration. Banerjee et al. [15] propose optimization method by using modified Ant Colony Framework to optimize the scheduling throughput to the service for all the diversified requests using different resource allocators available. Wei et al. [16] suggest a deadline and budget constraint cost-time optimization algorithm for scheduling dependent subtasks by using game theory.

In [17], the cost-performance tradeoff in cloud IaaS was addressed, where the problem has been formulated as a multi-objective optimization. The proposed model was built based on a fine-grained charging model and a normalized performance model. The implementation using genetic algorithms and the experimental results proved the effectiveness of the proposed model.

3 Experimental setup

As mentioned earlier, the optimal cloud resource allocation problem will be studied. Resource allocation in a cloud is understood here as the allocation of virtual machines (VM) to physical resources. The cloud network is clearly dynamic, so rather than allocating according to the physical resources of a node, it should be done with respect to the instantaneously available free resources of a node. The result of the optimization is an assignment of VM-node pairs [18] and the related performance metrics.

3.1 ACO-based Multi-objective CPTOP Model

The master-slave architecture is a mature architecture with a single master server or job tracker and several slave servers, which has been widely used in cloud computing like in Google's MapReduce and Hadoop. Fig. 3 shows a typical scenario of the network topology of virtual resources in cloud computing, which is based on the master-slave architecture.

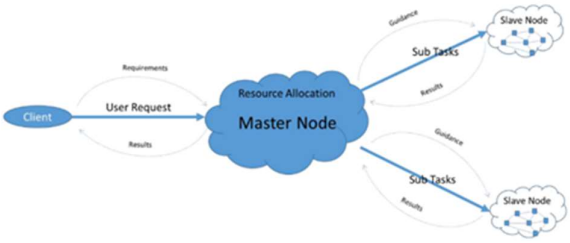


Fig. 3. Cloud computing resource architecture

In the master-slave architecture, a request is first submitted to a master node in the cloud platform by the user. Then the request is divided into several executable tasks in the master node and the generated tasks are distributed to different slave nodes.

After receiving the assigned sub tasks, the slave nodes will find appropriate resources. The tasks are executed in the slave nodes separately with the guidance of the master node, and the results are returned to the master node. The results include information about processing abilities, characteristics (number of CPU cores, amount of main memory, etc.), and cost.

Finally, the distributed results are combined together in the master node and sent to the requesting user. Furthermore, the master node is responsible for monitoring the all the steps and re-executing the failed tasks.

In this paper, the role of ACO is simulate several generations of artificial ants that search for the optimal solution. Every ant of a generation builds a solution step-by-step going through several probabilistic decisions. In general, ants that find a good solution mark their paths through the decision space by putting some amounts of pheromone on the edges of the path.

The pheromone attracts the ants of the next generations, and the result is that they search the solution space near the previous good solutions. In addition to the pheromone values, the ants are usually guided by some problem-specific heuristic for

evaluating the trial solutions.

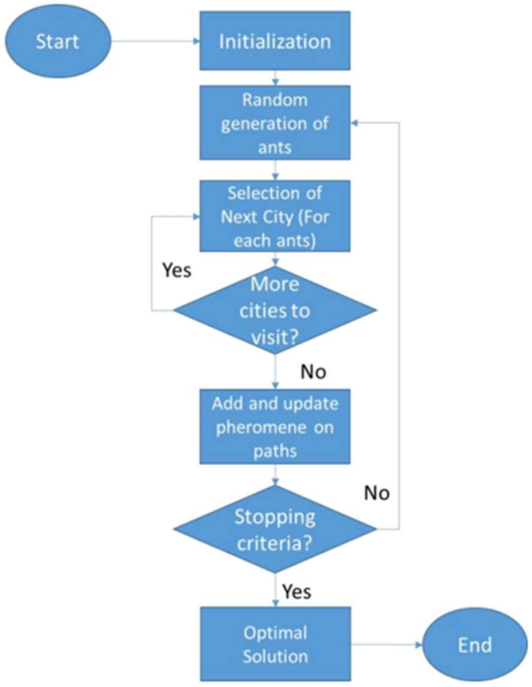


Fig. 4. Flowchart of Ant Colony Optimization

In order to apply ACO to tackle Cost-Performance trade-off problem (CPTOP), the problem should be transformed into a Travel Salesman Problem (TSP) in Fig. 4. Moreover, the separate objectives of cost and performance should be integrated into a single objective function.

An analysis of a cloud computing platform reveals several characteristics that are in common with the standard ACO: slave nodes being analogous to food locations, master nodes to nests, and resource allocation to foraging activity. For each of the slave nodes, ACO needs to calculate the free cloud resources. If the available resource exceeds the user's requirements for accomplishing the sub tasks, then this slave node should be allocated to the respective sub task. If the remaining resource is insufficient for the minimal requirement of the user, another appropriate slave node is searched for.

When the slave node sends back the results, the pheromone values are updated and saved. The master node will send new guidance according to the returned information to other slave nodes. This method can optimize the final results. The search for suitable slave node activity is conducted in a certain range to decrease the cost and increase the performance.

3.2 CloudSim

CloudSim provides many ways for managing and utilizing the resources, such as virtual machine (VMs), datacenter, and so on. It supports the research and development of cloud computing in testing the performance of a newly developed application service in a controlled and easy to set-up environment. CloudSim offers: (1) support to modeling and simulation of a large cloud computing infrastructure, (2) a self-contained support data center, service agent, scheduling and allocation strategy platform.

The framework and architecture of CloudSim consist of four main layers:

- *SimJava layer* supports several core functionalities required for simulation, such as queuing and processing of events, creation of system components (services, host, data center, broker, VMs), and management of the simulation clock.
- *GridSim* layer includes libraries that support high-level software components for modeling multiple grid infrastructures, including networks and associated traffic profiles, and fundamental grid components.
- *CloudSim* layer provides support for modeling and simulation of virtualized cloud-based data center environments including dedicated management interfaces for VMs, memory, storage, and bandwidth.
- *CloudSim stack* is the top layer, and it includes the user code that defines basic entities for hosts (number of machines, their specification, and so on), applications (number of tasks and their requirements), VMs, number of users and their application types, and broker scheduling policies.

3.3 Optimization

An important step in defining an algorithm for the resource allocation problem is defining the objective of the optimization. The objectives for the customers and the cloud service providers are different. In a simplified view, the objective of the customer is to maximize the performance of resources with a fixed cost. For the cloud service provider, the total amount of resources is fixed, and the objective is to add as many customer requests to the cloud as possible.

The more detailed rationale is as follows:

- As for the customers: When customers send requests to the cloud services provider for execution of their tasks, they seek to reduce their costs by transferring their operation to the cloud environment. Then the best offer is selected and the corresponding resources will be allocated to run the task. Moreover, they want to receive as good service as possible (within the limits of the cost). From the customers' perspective, they are selfish, because the customers are not concerned about the other customers in cloud.
- As for the cloud services providers: They want to

increase the profits obtained from the limited resources through the increase of income from hosting more users while minimizing the cost (investments) by optimal assignment of customers' requests to the resources.

Cost and performance are two competing objectives in cloud resources allocation. It is a NP-hard and multi-objective optimization problem without a unique solution. The number of possible solutions grows exponentially with respect to the number of resources and customers.

The optimization goal is to find, in some respect, the best trade-off between cost and performance. There are two challenges:

– A multi-optimal approach seems infeasible due to the hardness of the problem. Solutions optimal with respect to different criteria will tend to be vastly different, and there is no way to find a trade-off by interpolation due to the discrete nature of the resource assignment.
– It is difficult to define performance and quality from a system perspective. In the dynamic resource allocation, requests are assigned one by one, and simple heuristics would give no guarantee of fairness in performance. In this paper, the minimum level of performance is therefore determined by the Service Level Agreement (SLA) of each request.

In the ACO, the movements of the ants are controlled by probabilities that are products of two parts. The first is an assignment probability that is proportional to the attractiveness of a match from the customer point of view (called visibility in [5]). The second is a memory of the best past assignments represented by the fictitious pheromone trail. As long as the SLA of a request can be fulfilled the attractiveness is non-zero, otherwise it is zero.

The probability of transition to another (including current) node is evaluated only for feasible assignments:

$$P_{ij} = \begin{cases} \frac{(\tau_{ij}(t))^{\alpha}\eta_{ij}(t))^{\eta}}{\sum(\tau_{ij}(t))^{\alpha}((\eta_{ij}(t))^{\beta}} \\ 0, \end{cases} \quad (1)$$

where the pheromone $\tau_{ij}(t)$ and the attractiveness $\eta_{ij}(t)$ dependent on time t. The pheromone density changes in each cycle, while the attractiveness in every move within a cycle.

Cost can be defined in terms of idle capacity, that is, unoccupied capacity that cannot be assigned to another VM due to limitations in some other resource type. The cost will depend on the applications, or, in other words, the distribution of demands of arriving requests.

The cost for a cloud service providers can be expressed in the degree of infrastructure utilization or, equivalently, return on investments. The service providers wish to allocate jobs to resources in a best-fit manner, so that an allocated customer occupies no more than the necessary minimum.

The greedy principle from the cloud provider's perspective is that the more VMs can be allocated, the higher the utilization and the return on investment is. As a metric for system efficiency, the energy of the relative free resources is used. The optimization then follows the principle of minimum energy. The system energy is defined as:

$$E = \sum_{i=1}^{n}(C_i - \sum_{j=1}^{v_i} r_{ij})^2 \quad (2)$$

where C_i is the capacity of the server i and r_{ij} is the VM capacity requirement of VM j on node i. The total energy sums over the squared free resources of all servers.

It is important to performance features into account while allocating resources, since it allows providing the customers high Quality of Service (QoS), with the best response time as an example, and to meet the Service Level Agreement (SLA) established. Indeed, it is not easy to handle efficiently resource allocation processes in Cloud, since the applications deployed in Cloud obey non-uniform usage patterns, and the cloud allocation architecture needs to provide different scenarios of resource allocation to satisfy the demands and provide quality [19].

Now the actual algorithm is an adaptation of the ACO algorithm for solving the Traveling Salesman Problem (TSP), described in [5]. The assignment problem is modeled as a complete graph on the set of n nodes.

Initially, the ants are distributed between the nodes in a round-robin fashion. They could also originate from a source node (a nest), but this is not necessary, as the algorithm only performs a single iteration in each cycle. The ants could also be distributed randomly, which would affect the order of assignments. In the example below, however, this has no or little effect.

The ants move according to a matrix of transition probabilities, where self-loops are allowed, so that an ant may request its job to be assigned to the node it originally occupies. As opposed to the TSP, where the attractiveness is fixed, the system state changes with assignment of a new job (the property of the ant, or customer). Therefore, after each move, the transition probabilities change and must be recalculated. The attractiveness of a server to a given customer decreases when resources are assigned to another customer. As a measure of attractiveness, the (scaled) available CPU processing power of the server is used.

Next, the system cost is calculated according to (2). This energy could be a composite measure that includes other resources, such as RAM as suggested in [1]. However, now the energy is based only on the CPU processing power.

The deposited amount of pheromone, $\Delta\tau$ on each edge is now dependent on the system cost, rather than the trail of a single ant as in the TSP. This quantity is given by:

$$\Delta\tau = Q/c_k \tag{3}$$

where Q is a scaling constant and c_k the cost in cycle $k \in \{1, 2, ..., N\}$. Since c_k can be zero, a maximum limit on $\Delta\tau$ is set to one. This limit is rather arbitrary, and it is an additional system parameter that affects the convergence properties of the algorithm.

The cost is used to update matrix **P**, first by multiplying all previous pheromone levels p_{ij} by the evaporation constant $(1 - \rho)$, and then by adding $\Delta\tau$ onto edges describing the assignments used in the iteration.

The minimum cost and the corresponding assignment obtained so far is recorded after each cycle. Matrix **A** and the vectors of the free node resources and assignments are restored to their initial values, corresponding to not yet assigned jobs. Table 2 summarizes the ACO algorithm.

Table 2. Resource allocation algorithm

Algorithm (Resource Allocation).
Given matrices of server capabilities S and VM requirements V.
STEP 0: (initialize)
Let J be a list of initial node assignments, and set algorithm parameters α, β, τ_0, ρ and K, the number of cycles. Set the matrix of free resources to $A = S$ and the matrix of pheromone concentrations $P = (\tau_0)$, the matrix where all entries equals τ_0. Set $c_{min} = \infty$
STEP $k = \{1, 2, \ldots, N\}$: (iterate)
while $k < K$ (the number of cycles) **do**
Randomly select a node i and a customer request,
Divide a customer request into tasks (ants). For each ant j:
Calculate the probabilities p_{ij} of assignment based on A and P, subject to the technical constraints (Eq. (3.1));
By simulation, select a move of the ant j, assign the job to the selected target node if resources are available; assign resources, and update A (end).
When all ants have been moved once:
Calculate the cost c_k, (defined by the energy (Eq. (3.2))) of this assignment, and update matrix P by $\Delta\tau$;
If $c_{min} < c_k$, let $c_{min} = c_k$ and $J_{min} = J_k$, Reset J, $A = S$, and let $P = (1 - \rho)P$; (end)
Delete customer request (end)
end;
Output c_{min} and J_{min}, the optimal assignment.

4 Experiments

To test the algorithm, the small cluster setup described in [18] was used. The simplicity of this scenario with five servers having different characteristics and a single type of virtual machine (VM) makes the manual comparison with other assignment schemes straightforward. The algorithm as such should be easily extended to larger and more general cases.

To compare the algorithm with other assignment schemes, the results with the round-robin and a customer greedy heuristic schemes were both tested. In the CloudSim experiments, the goal was to keep things as simple as possible apart from the hosts and VMs. Only one user, one datacenter and one broker were created and initiated. The VMs represent the ants, and the cloudlets jobs assigned to the VMs.

The three assignment strategies (round-robin, greedy, and ACO) were implemented in CloudSim using Java. The round-robin assignment was implemented on the basis of a project in Github [20]. The number of cloudlets was set to 10 as the code in [20] also used.

The number of ants (VMs) is set equal to the number of nodes. The properties of the host servers in the cluster are listed in Table 3, and of the virtual machines in Table 4.

Table 3. Host servers specification: MIPS and RAM capacities

Host ID	Core	CPU (MIPS)	Memory (RAM)
0	1	1000	2048
1	2	500	2048
2	2	300	2048
3	1	2000	2048
4	2	300	2048

Table 4. Virtual machine specification: requirements on MIPS and RAM

VM ID	Core	CPU (MIPS)	Memory (RAM)
0-4	1	300	512

Tables 5, 6, and 7 show the results of the three algorithms. The reported metric for each of the hosts is the percentage of free capacity, calculated as:

$$1 - \frac{\text{occupied capacity}}{\text{total host capacity}}. \tag{4}$$

Two versions of the metrics are calculated: taking and not taking into account the number CloudSim Processing Elements (PEs), i.e., cores.

The cost as defined in Equation (2), that is, the sum of the squared entries. The round-robin and the greedy algorithms are deterministic, whereas the ACO algorithm is stochastic. ACO may therefore give a different result at each run, and not even a reasonable convergence is guaranteed. This depends on the random number generator.

In the round-robin scheme, the VMs are simply distributed one at each node, and the relative free capacity is shown in Table 5. The cost is 1.3725 and taking the number of processors into account, the

energy is 2.2025.

Since the round-robin assignment scheme is deterministic and not an optimization method, it is likely to perform poorly when there are VMs with different requirements. In this example, however, the round-robin assignment is good, under the energy metric.

Table 5. Efficiency of round-robin assignment

Host ID	No. VM	Free Cap. w/o PEs	Free Cap. with PEs
0	1	0.7	0.7
1	1	0.4	0.7
2	1	0.0	0.5
3	1	0.85	0.85
4	1	0.0	0.5
Cost (Energy)		1.37	2.20

The greedy assignment method lets each customer choose server according to the largest amount of available processing capacity. These results are shown in Table 6. The cost (energy) in this case is 3.65. The same value is achieved when taking the number of processors into account, since there are zero VMs in all multi-core hosts (Host IDs 1, 2, and 4, compare to Table 3). The greedy scheme is essentially what would be expected from a single iteration of the algorithm.

Table 6. Efficiency of greedy assignment

Host ID	No. VM	Free Cap. w/o PEs	Free Cap. with PEs
0	1	0.7	0.7
1	0	1.0	1.0
2	0	1.0	1.0
3	4	0.4	0.4
4	0	1.0	1.0
Cost (Energy)		3.65	3.65

The ACO algorithm applied to the same problem gave the assignments shown in Table 7. It can be seen that the lower capacity nodes (1, 2, and 4; see Table 3 for the Host specifications) are assigned VMs, but not node 3. The minimum energy obtained is 1.32. After reaching the minimum energy, the algorithm was run for up to N = 10,000 without showing any further improvement. Taking the number of processors into account, the energy is 2.15 for this policy.

Table 7. Efficiency of ACO assignment

Host ID	No. VM	Free Cap. w/o PEs	Free Cap. with PEs
0	2	0.4	0.4
1	1	0.4	0.7
2	1	0.0	0.5
3	0	1.0	1.0
4	1	0.0	0.5
Cost		1.32	2.15

The parameter values used are α = 0.5, β = 0.5, ρ = 0.1 and υ_0 = 0.1. The cut-off limit for the inverse of the cost was set (rather arbitrarily) to 1. Elaborating on the system parameters would probably influence the convergence of the algorithm significantly, but it has not been studied in detail here.

The convergence performance (Fig. 5.) shows one run trace of the algorithm. This case shows an initial guess that already better than the cost of the greedy assignment (compare to Table 6). After that ACO finds a local minimum in two more iterations. The figure shows a case of an optimization with a particular (lucky guess) seed, which converged very fast. Normally, such an optimization would show a jagged curve stretching much longer on the x-axis. The figure is intended to show the actual ideal convergence rather than the convergence performance in general.

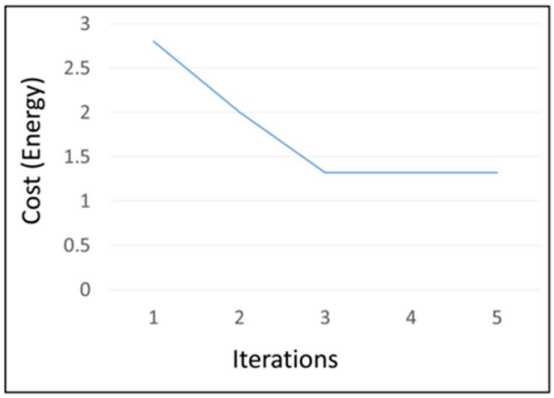


Fig. 5. Ideal convergence of ACO

Fig. 6. shows the energy for each of the three assignment strategies for an increasing number of standard size VMs as defined in Table 4. The ACO algorithm described in this study (green line) has lower energy than the other two, although the round-robin strategy (blue line) is close to optimal. The greedy algorithm is evidently the worst of these three with practically any number of VMs.

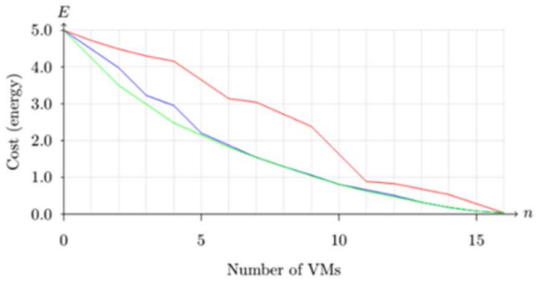


Fig. 6. Comparison of the efficiency of the algorithms: round-robin (blue), greedy (red), and ACO (green).

5 Conclusions

Cloud computing enables a more efficient way to use utility computing resources and services. Users can access the computing resources with virtualized technologies and pay only for the resource accessed while getting the level of quality of service (QoS) wanted. In this paper, a modified ant colony optimization algorithm for cost and performance trade-off optimization problem is developed to encourage the formation of solutions to achieve the cost minimization.

An optimal assignment was found by minimizing a combined energy function that measures cloud provider's costs. However, the test setup was relatively limited and simplified, due that a full-scale simulation of cloud computing services are complicated. The benefit for the cloud provider is to maximize the possibility to add further VMs to the existing cloud infrastructure without performance degradation or delays.

Three assignment policies were simulated and tested in CloudSim: round-robin, a customer greedy heuristic, and an optimized allocation implemented as an Ant Colony Optimization algorithm. When comparing the three assignment policies, the round-robin can be said to be both simple and efficient. The greedy assignment, where a customer can choose to allocate a VM to the host with the freest capacity was rather expensive. However, in this implementation, each processor had the same MIPS, so the results might be different in an environment with more versatile set of resources available.

Finally, the way that ACO uses resources differed from the round robin method in this case only so that one extra VM was assigned to host number 0, instead of host number 3. Nevertheless, this little change makes ACO to obtain the best energy function values. However, in this simple setup the difference is only slight. The further tests with varying numbers of VMs validated the mutual order of the three algorithms; ACO consistently outperforms the simple round-robin method slightly, while the round-robin method outperforms the simple greedy method significantly.

By using this dynamic optimization, the new request will be given to some host, and in the same time, an already assigned VM can in principle be re-assigned, but this case has not been tested in this study. Instead, to reach an optimal solution, the algorithm starts afresh in each iteration. It converges after a (random) number of iterations, and this converged result is then the assignment policy.

These experiments have some limitations. Actually, the implementation of ACO in CloudSim makes a solution 'all at once', not just a list of nodes like in CloudSim. Therefore, it is recommended to develop an ACO variant that could find an optimal policy with a more dynamic situation, where VMs are created and terminated all the time.

References

[1] Asha N., Rao G. R., A Review on Various Resource Allocation Strategies in Cloud Computing, International Journal of Emerging Technology and Advanced Engineering (IJETAE), 2013, 3(7).

[2] European Commission, The Future of Cloud Computing – Opportunities for European Cloud Beyond 2010, Public Report, European Commission, 2010.

[3] Mell P., Grance T., The NIST definition of cloud computing (version 15), 2009, retrieved from https://www.nist.gov/sites/default/files/documents/itl/cloud/cloud-def-v15.pdf.

[4] Wu M., Addressing Resources Allocation Issues in Cloud Computing Environment, M.Sc. thesis, University of Vaasa, Vaasa, Finland, 2016.

[5] Dorigo M., Maniezzo V., Colorni, A., Ant system: optimization by a colony of cooperating agents, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 1996, 26(1), 29–41.

[6] Afshar A., Kaveh A., Shoghli O. R., Multi-objective optimization of time-cost-quality using multi-colony ant algorithm, Asian Journal of Civil Engineering (Building and Housing), 2007, 8(2), 113–124.

[7] Chimakurthi L., Power efficient resource allocation for clouds using ant colony framework, 2011, arXiv preprint arXiv:1102.2608.

[8] Fidanova S., ACO algorithm for MKP using various heuristic information, In: Dimov I., Lirkov I., Margenov S., Zlatev Z. (eds.), Proceedings of the International Conference on Numerical Methods and Applications (20–24 August 2002, Borovets, Bulgaria), Springer, Berlin Heidelberg, 2002, 438–444.

[9] Parikh K., Hawanna N., Haleema P.K., Jayasubalakshm R., Virtual Machine Allocation Policy in Cloud Computing Using CloudSim in Java, IJGDC, 2015, 8(1), 145–158.

[10] Foster I., Kesselman, C. (Eds.), The Grid 2: Blueprint for a new computing infrastructure, 2nd ed., Morgan Kaufmann, 2003.

[11] Chaisiri S., Lee B. S., Niyato, D., Optimal virtual machine placement across multiple cloud providers, In: Proceedings of the IEEE International Conference on Services Computing (21–25 September 2009, Bangalore, India), IEEE Asia-Pacific, 2009, 103–110).

[12] Frincu M. E., Craciun C., Multi-objective meta-heuristics for scheduling applications with high availability requirements and cost constraints in

multi-cloud environments, In: The Proceedings of 4th IEEE/ACM International Conference on Utility and Cloud Computing (5–7 December 2011, Melbourne, Australia), IEEE, 2011 267–274.

[13] Hua X. Y., Zheng J., Hu W. X., Ant colony optimization algorithm for computing resource allocation based on cloud computing environment, Journal of East China Normal University (Natural Science), 2010, 1(1), 127–134.

[14] Omara F. A., Khattab S. M., Sahal R., Optimum Resource Allocation of Database in Cloud Computing. Egyptian Informatics Journal, 2014, 15(1), 1–12.

[15] Banerjee S., Mukherjee I., Mahanti P. K., Cloud computing initiative using modified ant colony framework, World academy of science, engineering and technology, 2009, 56, 221–224.

[16] Wei G., Vasilakos A. V., Zheng Y., Xiong N., A game-theoretic method of fair resource allocation for cloud computing services, The journal of supercomputing, 2010, 54(2), 252–269.

[17] Kong S., Li Y., Feng, L. (2012). Cost-performance driven resource configuration for database applications in IaaS cloud environments, In: Ivanov I., van Sinderen M., Shishkov B. (eds.), Cloud Computing and Services Science (18–21 April 2012, Porto, Portugal), Springer, New York, 2012, 111–129.

[18] Lee H. M., Jeong Y. S., Jang, H. J., Performance analysis based resource allocation for green cloud computing, The Journal of Supercomputing, 2014, 69(3), 1013–1026.

[19] Sagbo K. A. R., Houngue, P., Quality architecture for resource allocation in cloud computing, In: Service-Oriented and Cloud Computing, Springer, Berlin Heidelberg, 2012, 154–168.

[20] AnanthaRajuC, CloudSim Example with Round Robin Data center broker & Round Robin Vm Allocation Policy with Circular Hosts List, 2015, retrieved from https://github.com/ AnanthaRajuC/CloudSim-Round-Robin.

Mika Karaila

Automaation tulevaisuus – Tekoälyn ja ihmisen vuorovaikutuksia

Abstract: Uudet pilvipalvelut ja tekoäly tulevat, mitä se tarkoittaa automaatiossa. Tarvitaanko ihmistä prosessinohjauksessa, kuka tekee päätökset. Teollisen Internetin uudet pilvipalvelut on toteutettu koneoppimisen ja tekoälyn avulla. Paperiradan katkoanalyysi kertoo jo hyvällä todennäköisyydellä katkoherkkyden ja sen todennäköisen aiheuttajan. Nämä uudet tuotteet perustuvat erilaisiin koneoppimisen algoritmeihin ja tarjoavat aivan uudenlaisia sovelluksia, joiden avulla asiakkaat voivat ajaa konettaan. Muita uusia mahdollisuuksia tekoäly tuo ongelmien ratkaisuun erilaisissa olosuhteissa, joissa ns. Expert – tekoälypohjainen avustaja voi jutella käyttäjän kanssa ja hakea erilaisia vaihtoehtoja tai jopa ehdottaa ongelmaan jotain löytämäänsä ratkaisua.

Nämä ylläolevat esimerkit ovat uuden sukupolven pilvipohjaisten automaatioratkaisujen toimivia esimerkkejä. Havainnollistavia kuvia ja lyhyitä videoita sekä mahdollisesti myös live-demo on mahdollista näyttää esityksen aikana.

Avainsanat: AI, assistant, avatar

*Kirjoittaja: Research Director, E-mail: mika.karaila@valmet.com

1 Johdanto

Perinteinen automaatio on muuttunut yksikkösäätimistä ja perinteisistä PID-säätimistä monimuuttuja säätimiin ja muihin mallipohjaisiin säätöihin. Automaatioaste on erittäin korkea nykyaikaisissa tehtaissa ja seuraava nähtävissä oleva kehitys vie ratkaisuja pilvipohjaisiin / Edge-tason ratkaisuihin. APC-säätöjen lisäksi haetaan sopivaa tasoa ja toteutusta, jonka avulla automaatiosta osa voisi olla hajautettuna tehdaan ulkopuolelle. Näissä ratkaisuissa tulee vastaan tietoturva ja tiedon omistajuus. Näihin löytyy ratkaisuja ja sopimuksia, joita ei tässä artikkelissa kuitenkaan käsitellä tämän enempää.

Tämän artikkelin kohteena on juuri uudet ratkaisut, jotka pohjaituvat tekoälyyn ja miten se kehittyy automaation mukana. Ihmisen vaikutus ja päätöksen teko tulee olemaan oleellinen osa. Mallipohjainen ja APC-säädöt toimivat matemaattiset suunnitellun mukaisen kiinteän ratkaisun pohjalta. Kun tätä verrataan tekoälyyn pohjautuvaan säätöön, joka on voitu ensin opettaa ja validoida kerätyn datan perusteella on huomioitava seuraavat asiat:

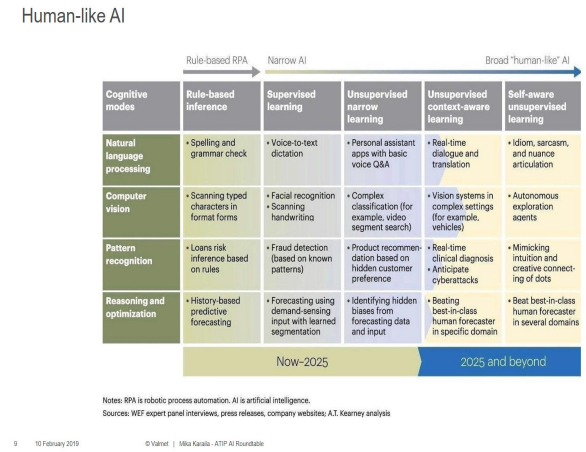
- Säätö perustuu opetusdataan, jos siinä on puutteita säätö ei voi toimia ko. tilanteissa.
- Jos ympäristö (prosessi) muuttuu, ei opetusdata ole enää oikeaa.
- Muut epäjatkuvuuskohdat ja anomaliat, joista ei ole ollut opetusdataa pitää huomioida säädössä, jotta voidaan varmistaa ettei säätö toimi väärin näissä tilanteissa.

Kun yllä olevat tiedostetaan on hyvin ymmärrettävää että vastaavalla tavalla kuin autonomisen auton kanssa ihmisen on koskettava ohjausrattia n. 10 sekunnin välein tarvitaan automaation tekoälyssä myös ihmistä, jonka näkymys ja laaja tietotaito on varsinainen päätöksen tekijä prosessin ohjauksessa.

Toisaalta kun ajatellaan miten tekoäly voi auttaa ihmistä prosessin ohjauksessa pitää huomioida, että tekoälyllä on seuraavat hyvä puolet:

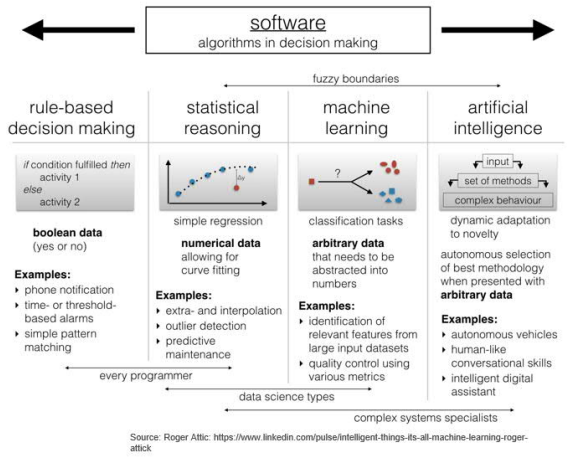
- Ihmistä parempi muisti eli usean vuoden ajalta kerätty data sekä omalta tehtaalta ja mahdollisesti muilta tehtailta on enemmän kuin ihminen todennäköisesti voi muistaa.
- Eri tavoitteiden tarkastelu oikein toteutettuna voi hakea paremman optimitilanteen tai tavoitteen kuin on ihmisen tiedossa. Joissain tilanteissa ihminen näkee 1. tason tavoitteen, mutta algortimi voi hakea optimaalista 2. tason tavoitetta, jossa on huomioitu sekä raaka-aineet että energian käyttö. Lisäksi monessa tilanteessa operaattori ajaa ensisijaisesti häiriötöntä tuotantoa (todennäköisesti mm. paperin tuotannossa hieman parempaa laatua, koska ei halua ratakatkoja vaikka tekoäly pystyy ajamaan enemmän kierrätyskuidulla tuotettua paperia).
- Yleisesti voidaan todeta että tekoäly toimii paremmin suppean alueen sanaston ja ongelman käsittelyssä, jossa ulkoisia tekijöitä on pystytty rajoittamaan

Ihmismäisen tekoälyn toteuttamisessa tarvitaan luonnollisen kielen ymmärtämistä sekä sen ”puhumista”, tätä selventää alla oleva Kuva 1 Ihmisen kaltainen tekoäly.



Kuva 1 Ihmisen kaltainen tekoäly

Yleisesti miten tekoälyn algoritmit voivat auttaa päätöksen tekemisessä ja sen ohjelmoinnin vaativuus on kuvattu eri tasoisina Kuva 2 Tekoäly ja algoritmit päätöksen tekemisen apuna



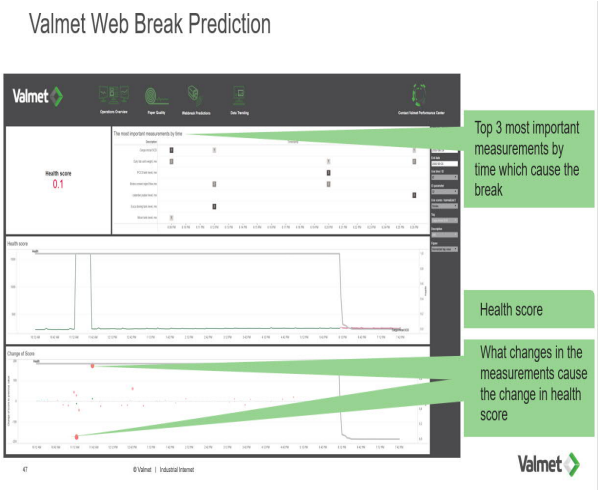
Kuva 2 Tekoäly ja algoritmit päätöksen tekemisen apuna

2 Tekoäly automaatiossa ihmisen apuna

Tekoälyn liittäminen automaatioon tai sen käyttö eri tavoin voidaan toteuttaa käyttämällä valmiita tekoäly algortimeja ja työkaluja:

– Katkoanalytiikka on koostettua monesta eri algoritmista, jotka hakevat prosessin herkkyyttä ja siihen vaikuttavia muuttujia.
– Valmet Expert, joka perustuu ihmisen keräämään tietoon ja olemassa oleviin asiakastapauksiin, ongelmasta ja sen ratkaisusta.

Katkoanalytiikassa on ensin valittava ne mittaukset ja säädöt, joiden perusteella data-analyysi voidaan tehdä ja kouluttaa. Näiden muuttujien hakeminen ja siivoaminen oli ensimmäinen työläs vaihe ennen ennustemallin rakentamista. Vasta tämän jälkeen tekoälyalgoritmi voidaan kouluttaa opetusdatan avulla, jotta ne löytävät erilaisia katkoja. Näiden tunnettujen katkojen tietojen perusteella saadaan viritettyä neuroverkko sille tasolle että ennustemalli pystyy varoittamaan operaattoria 1-2 tuntia ennen mahdollista katkoa tilanteen, Kuva 3 Katkoanakytiikan käyttöliittymä. Malli pystyy näin koulutettuna antamaan operaattorille ne prosessimuuttujat (root cause parameters), jotka vaikuttavat eniten ko. katkon ennusteeseen.



Kuva 3 Katkoanakytiikan käyttöliittymä

Yllä kuvattu paperikoneen katkoanalytiikka ei ole kuitenkaan yleispätevä. Siitä on tehty oma versio pehmopaperikonetta varten. Systeemidynamiikan takia näistä on tehty erilliset, jotta ennusteen tarkkuus on saatu tarpeeksi tarkaksi. Lisäksi ko. mallit eroavat myös opetusdatan osalta. Tiedon kerääminen ja sen tallettaminen on tehty kuitenkin samalla tavalla. Data on normalisoitu, joten samat algoritmit ovat käytettävissä muiden asiakkaiden vastaavien koneiden ennustemallien tekemiseen, mutta opetusdata on aina konekohtaista. Vastaavaa ennustelmallia voidaan todennäköisesti soveltaa myös johonkin muuhunkin prosessiin kunhan löydetään selkeä kaupallinen tarve, jotta kehitystyö kannattaa tehdä.

Avustava tekoäly, joka voidaan tuoda eri käyttöliittymiin, joko Web-selaimen avulla chatbot-tyylisenä tai VR-ympäristöön Avatar:ina on yleisti käsitelty kirjassa [1] ja artikkelissa [2]. Yleensä chat-bot on jo käytössä ns. ensimmäisenä ongelman ratkaisijana 24/7 online-support kanavissa sekä yksinkertaisessa

myyntikanavassa. Näissä ihminen voi ottaa alun jälkeen keskustelun haltuun, jolloin asiakastyytyväisyys saadaan pidettyä suhteellisen korkeana. Chat-bot hoitaa alkuruuhkan ja asiakas saa heti alkuun nopean vasteen. Kuva 4 Ihmisen kaltainen "chat-bot" on yksi edistyneempi proto-tyyppi, jossa kysymys – vastaus pohjainen tekoäly osaa vastata muutamaan paperikoneen huoltoon liittyvään kysymykseen. Tämä Q&A pattern oli nopeasti kirjoitettu käsin tehty kuvaus asiantuntijan ohjeista. Kuva 5 Valmet Expert avatar: ihminen tai tekoäly avustajana on yksinkertaisempi grafiikaltaan, mutta kollaboraatio toisen ihmisen kanssa on tehokkaampaa kuin tekoälyn kanssa. Taustajärjestelmä on kuitenkin sama, mutta käyttöliittymä saadaan kytkettyä näin tarpeen mukaan ja muutettua sopivaksi ko. käyttäjäryhmälle.

Kuva 4 Ihmisen kaltainen "chat-bot"

Kuva 5 Valmet Expert avatar: ihminen tai tekoäly avustajana

Avatar voidaan laittaa suoraan oikeaan kohtaan virtuaalisessa ympäristössä ja tarvittaessa se voi osoittaa huollettavaa kohdetta joko laser-osoittimella tai kädellä. Kohde voidaan myös korostaa muista laitteista värillä tai jollain muulla tehosteella. Tämän lisäksi voidaan avata liittyvää dokumentaatiota tai ehdottaa aikaisempia huoltotoimenpiteitä, joista ihminen voi valita sopivimman. Näin lopullinen päätöksentekijä on ihminen.

3 Toteutuneet tekoälypohjaiset sovellukset

Katkoanalytiikka on pilotoitu ja valmisteltu tuotteeksi, jota voidaan käyttää paperikoneen sekä pehmopaperikoneen katkojen ennustamiseen. Tämän hetken rajoitteena on saada ensin luotettava ennustemalli viritettyä. Tämä vaatii siis opetusdataa ja sen painokertoimien korjaamisen data-analyytikon toimesta. Alkutilanteessa ennusteen tarkkuus oli vain hieman yli 20%. Kuitenkin kun malli on saatu tehtyä ja se on validoitu muutamaan kertaan sen luotettavuus on yli 50%. Näihin mallin antamien juurisyiden tulkintaa tarvitaan ensin prosessiasiantuntijaa auttamaan asiakasta ymmärtämään mitä ko. tilanteessa pitää tai kannattaa tehdä, jotta ennustettu ratakatko voidaan välttää. Kaikkia ratakatkoja ei kuitenkaan voida ikinä estää, sillä osa katkoista johtuu siitä yksinkertaisesta syystä että kone tai prosessin osa vaatii oikeasti huoltoa esim. koneen telojen pesua tai vaihtoa.

Avustava tekoäly, joka ymmärtää ihmisen puhetta on vielä ns. proof-of-concept tasolla. Avustava ”Valmet Expert”-avatar saadaan vastaamaan kysymyksiin ja sen tietämyskanta on vielä suhteellisen rajattu. Rajoitteena on vielä ns. monitaitoisen (multi-skill) tekoälyn tietämyksen rakentaminen. Jos M-Files tai Salesforce korttien (ticket) tekstistä voidaan luoda alustava tietämyskanta ja siihen liitetään takaisinkytkentä (feedback channel), jonka kautta ihminen voi luokitella vastauksen oikeellisuuden (rating) niin tietämyskanta saadaan paremmaksi käytön mukaan. Tämä on seuraavan tekoälyyn liittyvän tutkimuksen kohde, josta yritetään rakentaa jonkinlainen proof-of-concept prototyyppi. Jos tämä onnistuu niin tämä mahdollistaa normaalin organisaation toimintatavan ja palveluprosessin kytkemisen yhteistoimintaan tekoälyn kanssa. Tällä voi olla vaikutuksia toiminan tehokkuuteen ja laatuun. Toisaalta, jos takaisinkytkennän kautta tekoälyn tarkkuutta ja oikeellisuutta ei saada paranemaan tekoälystä ei ole kunnollista apua eikä hyötyä.

Tämän takia on oleellista kokeilla uusia luonnollista kieltä tukevia tauastajärjestelmiä niiden tekninen valmius paranee koko ajan. Oleellista kehityksessä on

se kuinka helppoa niitä on soveltaa. Soveltamisessa tarvitaan API-rajapintaa sekä ajoaikaista dynaamisuutta. API-rajapinnan avulla sanastoa ja taitoja voidaan päivittää ja muuttaa takaisinkytkennän kautta. Ajoaikainen muunneltavuus liittyy suoraan API-rajapintaan, osa järjestelmistä on tällä hetkellä staattisia eli suurin osa dialogista on etukäteen suunniteltu. Niissä on muutamia ajoaikaisia muuttujia, mutta suurin osa rakenteesta on staattista. Kuitenkin nähtävissä on että näistä saadaan tulevaisuudessa dynaamisempi, koska toteutuskielet tukevat mm. template rakenteita ja ovat tulkattavia tai ajoaikaisesti käännettäviä.

Tämän lisäksi jos käytetään kognitiivisia menetelmiä ja kerätään operaattoreiden tietämystä erilaisista tavoista ajaa prosessia ja yhdistetään nämä muuhun kerättyyn dataan voidaan saada aikaan vieläkin parempaa tietämystä. Tämän avulla ikääntyvän operaattorisukupolven tietämystä ja osaamista voitaisiin siirtää opastavalle tekoälylle. Tosin tämä tietämys on erittäin riippuvaista henkilökunnasta ja kulttuurista. Ongelma voi olla vielä, että tietämys on erittäin painottunut ko. tehtaalle ja sen järjestelmään että prosessiin. Tosin tämä voi olla asiakkaalle kuitenkin juuri haluttu tietämys, jotta toimintatavat perityvät sukupolvelta toiselle. Tätä prototyyppiä ei ole vielä kokonaisuudessaan päästy kokeilemaan.

4 Tulevaisuuden näkymät

Nämä ensimmäiset tekoälyyn pohjautuvat ratkaisut ovat avaamassa automaatiolle täyden uusia mahdollisuuksia ja mielenkiintoisen evoluution aikaa, kun useita tehtaita ja erilaisia analytiikka sovelluksia tullaan toteuttamaan tekoälypohjaisesti.

Tekoälyn kouluttaminen ja opetusdatan päivittäminen ovat tällä hetkellä ihmisen vastuulla. Tulevaisuudessa nämä saadaan todennäköisesti suurilta osin automatisoitua tai mallien adaptiivisuus toimii ilman erillistä kouluttamista. Tämä mahdollistaa edelleen tekoälyn soveltamista useampaan kohtaan, kun ns. ”rutiinityö” voidaan jättää koneen tehtäväksi.

Tärkeää tässä kehityksessä on pitää ihminen mukana, jolloin tekoälyn tekemät mallit voidaan varmentaa ja samoin päätöksen tekeminen on edelleen ihmisellä. Väärin koulutettu tekoäly tekee vääriä ehdotuksia ja muuttunut ympäristö, johon koulutusdata ei saa tarvittavaa tarkkuutta ei voi säätää prosessia. Siksi ihminen tarvitaan tekemään ja vastaamaan varsinaisesta prosessin ohjaamisesta ja erilaisista huoltotoimenpiteistä laitoksissa.

Viitteet

[1] The next generation of AI assistants in enterprise https://www.oreilly.com/ideas/the-next-generation-of-ai-assistants-in-enterprise2007, 2018

[2] The Rise of a New Generation of AI Avatars https://singularityhub.com/2019/01/15/the-rise-of-a-new-generation-of-ai-avatars/#sm.001t70u7m14uvfsmsph1g9qp6wolg

Henri Pettinen, Marko Elo

Utilizing multifunctional display computer as a local gateway in industrial IoT use cases

Abstract: The Industry 4.0 paradigm emphasizes the increased level of automation and data exchange in manufacturing technologies. Hence, industrial companies have to deal with vast amounts of data gathered from the production. Usually it is not practical to send all the measured data outside the local premises for further processing. Onboard processing, temporary storing, and appropriate forwarding of the data are key operations to be handled by an industrial gateway.

CrossControl produces computers for industrial use, with integrated displays. Most common use case for such display computers are in heavy industrial vehicles, with connectivity to the onboard vehicle control bus. By utilizing a dedicated connectivity module, the onboard display can have a wireless connection as well, for on-site device and backoffice connectivity.

This conference paper for Automaatiopäivät23 introduces key parts of the Productive4.0 project which addresses the general theme of Industry 4.0. The paper addresses the use cases of machine and fleet management offered as an industrial Software as a Service (SaaS). The pilot implementation of CrossControl's gateway device is presented with the focus being on the software stack, information flow architecture and the graphical user interface. Moreover, applying Service Oriented Architecture (SOA) on gateway and edge level is also discussed.

Keywords: Industry 4.0, Internet of Things, IoT, Productive4.0, Multifunctional display, Arrowhead framework, Graphical user interface, Gateway, SaaS, SOA, CPS.

***Corresponding Author: Henri Pettinen:** CrossControl Oy, E-mail: henri.pettinen@crosscontrol.com

Second Author: Marko Elo: CrossControl Oy, E-mail: marko.elo@crosscontrol.com

1 Introduction

In a continuously digitizing world, industrial machines have reached the point where they are capable of harvesting data and efficiently exchanging it with other systems over wireless media. This brings a vast number of benefits, one being, for example the improvements in the overall situational awareness at the work site. The very same trend is prevailing in every business sector and everyday life. Internet of Things (IoT) is the umbrella concept for the physical objects being able to sense their surroundings and then sharing that information with peer systems and other surrounding infrastructure.

In industrial domain, the collected data provides valuable assets for multiple stakeholders, including but not limiting to the end users, system integrators, maintenance service providers, OEMs, and even insurance brokers. However, the data is usually valuable for hostile parties as well, whose interest is to attack system vulnerabilities and engage industrial espionage or cause disruption in general. The characteristics of Internet-based connections force companies to apply hardening approaches to their digitized production environments, setting tougher security requirements also to gateway and edge devices in IoT concepts.

Productive4.0 is a European-wide innovation project involving more than 100 partners from 19 European countries. As an ARTEMIS Innovation Pilot Program (AIPP), it aims for providing industry-driven digitalization and internet connectivity use cases for multiple industrial domains, with proof of concept implementations conducted by project partners organized as co-creation teams.

CrossControl Oy is specialized in control and information system solutions for industrial vehicles operating in demanding environments. The company is involved in Productive4.0 project, with one of the main contributions in a use case focusing on offering machine and fleet management as an industrial service.

This paper is partitioned as follows: Chapter 2 introduces CrossControl product platform supporting the scope of IoT concepts. Chapter 3 focuses on describing the key principles behind the Arrowhead framework. Chapter 4 is dedicated for enlightening two Productive4.0 use cases, and Chapter 5 describes downstream details by CrossControl's implementation

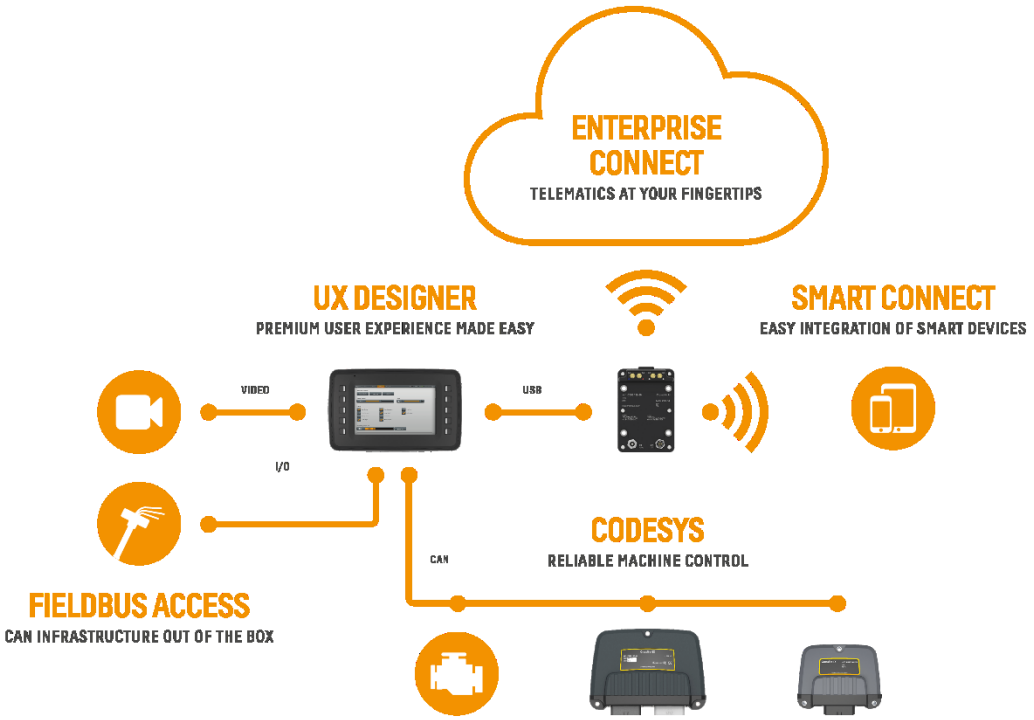


Fig. 1. LinX application platform comprises various premade software and hardware components.

in the use case context. Chapter 6 outlines the upcoming steps of the use case, along with a few development ideas. Chapter 7 concludes the paper.

2 CrossControl platform

This chapter introduces the relevant parts of CrossControl product platform in an IoT-featured application context, with specific match to requirements driven by Productive4.0.

2.1 Overview

CrossControl's CCpilot display product line is capable of fulfilling divergent functions needed in the industrial field. They can be used as an onboard instrumentation display, human machine interface (HMI), video monitor, electronic manual, or as an operator's connected task manager, for instance.

The software development kit (SDK) and runtime library for CrossControl displays, called as LinX, provides the tools and protocols for applications operating in onboard data collection, monitoring, visualization, logging, and exchange with the cloud. Functionality provided by LinX is illustrated in figure 1. The SDK comes with virtual machines for each different target in the CCpilot product line, facilitating and accelerating application development and deployment. The SDK contains all the display specific environment variables and configurations, enabling developers to cross-compile software applications for the display.

2.2 CCpilot VS

CCpilot display, model name VS, is involved in the targeted Productive4.0 use case. The VS is based on an ARM Quad Cortex-A9 CPU and 12" TFT multi-touch display, running a custom-made Linux distribution as its operating system. It has four CAN ports supporting ISO 11898 CAN 2.0B protocol, an Ethernet interface

Fig. 2 CCpilot VS display computer.

supporting 10BASE-T/100-BASE-TX/1000-BASE-T connection and Auto-MDIX, total of four USB ports, and a light sensor.

2.3 CrossLink AI

CrossLink AI is a communication module, connecting to CCpilot displays via USB. It supports a selection of wireless technologies enabling connectivity with on-site peers and enterprise systems. The add-on module provides WLAN access point, Bluetooth, 3G and GPS functionality for the displays.

3 Arrowhead architecture

Arrowhead framework is an academic approach for a generic and distributed framework for IoT automation applications. The first version was developed in ARTEMIS project Arrowhead, with further evolution taking place in projects EMC2 and FAR-EDGE. The Arrowhead framework is being utilized e.g. in projects Opti, MANTIS, and Productive4.0.

The framework aims to increase the interoperability of various systems and enhance the use of Service Oriented Architecture (SOA) in industrial domain, with Industry4.0-based thinking in mind. In its terminology every internet connected object is abstracted to a **service**. Typical for SOA is that different single functionality providing services can be utilized together in order to achieve functionality of a large software application [1]. Arrowhead framework follows this approach in several ways, e.g. by providing service orchestrator component. The orchestrator dispatches service requests to find the best suitable services to be invoked for a specific query or function call. The service orchestrator and other core components of Arrowhead framework are discussed in more detail later in this paper.

The way Arrowhead increases the level of automation is based on the idea of local clouds, that is, a whole cloud infrastructure on enterprise's own local network. All physical components must be inside a closed industrial network in order to achieve all benefits of the Arrowhead framework. The framework also strives to improve real-time data handling, with data and system security, and with built-in methods for efficient scalability. The local clouds can communicate through defined end points, forming larger clouds, and networks of clouds. These end points create interfaces for every local cloud, establishing a controlled access point for external actors.

Arrowhead core services are the mandatory components in every local cloud infrastructure that must be up and available. This comprises of three components in total: Service registry, Orchestrator and Authorization system. The Service registry stores all new and available services inside the local cloud and information about them. The Orchestrator utilizes this registry and its information to inform the services in local cloud that are willing to consume other services. In other words, the Orchestrator knows all available services and where they reside. By requesting a specific service from the Orchestrator, it returns the direct address to the exploitable service. The Authorization system verifies that a given consumer is allowed to access the addressed service. [2] These core services are required in order to deploy the minimum Arrowhead compliance in the local cloud. More information of further services, including Plant description, Configuration, System registry, Device registry, Event handler, QoS manager, Historian and Gatekeeper, can be found from the Arrowhead documentation.

Arrowhead provides flexibility to add new machines on the work site as they can be automatically registered as Arrowhead services when they first time connect to the local cloud's Service registry. After that, the services provided by the new machine can be discovered by other registered services inside the local cloud or even outside it, if made accessible.

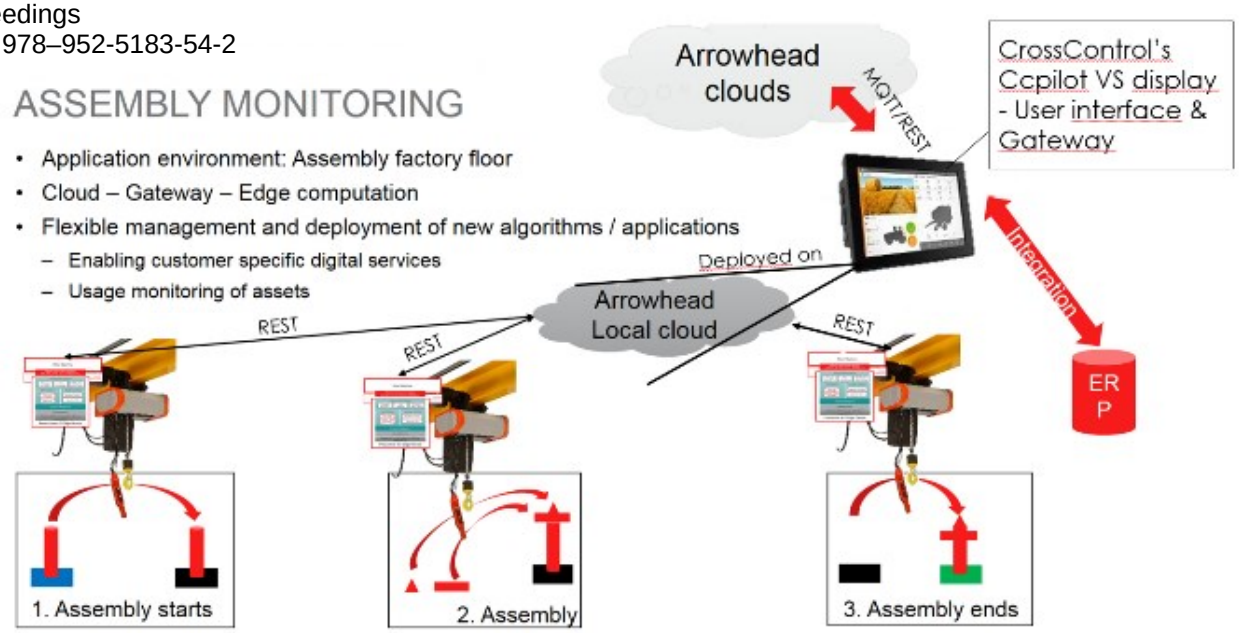


Fig. 3. Demonstration of the lifting business use case

4 Use case descriptions

This chapter introduces two Productive4.0 use cases, involving piloting of the Arrowhead framework in a fleet management context, deployed on CrossControl system platform.

4.1 Lifting business use case

In various assembly tasks there are lifting machines involved to help the process. Assembly, like any other part of manufacturing, requires control and monitoring to improve the process. Further added value can be created by harnessing the collected data to offer customer-specific, differentiated digital services. Such special requirements may cover e.g. field data forwarding to ERP or other backoffice systems, or specific functional safety triggers for onboard systems to react to predefined irregular data values.

This use case centralizes to a chain hoist, and the hoist OEM manufacturer's interest is to develop a method to track the movement of the chain hoist and then recognize the different work phases of the ongoing assembly process. Assembly process is divided into three phases which are the start of the assembly, the assembly itself, and the end of the assembly. These phases can be identified when the load and the vertical movement of the chain hoist are monitored. Figure 3 depicts this assembly monitoring process and clarifies the display computer's role as a gateway in this use case.

The chain hoist has an IoT edge device attached to it. The edge device retrieves sensor data from the hoist and forwards the data to the gateway, within the local network. It takes its power supply straight from the hoist and it can send the measured values to multiple subscribers. The data is conveyed in the JSON message format. Local Arrowhead cloud can be established on the work site and its core services can be ran on the CCpilot VS or some other computer on site dedicated for the gateway role.

The gateway dispatches a defined set of the collected data to the manufacturing company's backoffice systems and provides selected information also for other Arrowhead clouds. The communication is done using HTTP and MQTT protocols, usually over wireless connections.

4.2 Rock crushing use case

Another use case is associated with vibrating screens which are machines primarily utilized in the mineral processing industry for separating material according to size. Because of their high frequency vibration, the screens have to withstand high accelerations. These machines are relatively prone to mechanical part wear and tear, which requires well-planned maintenance activities to avoid downtime and loss of efficiency in operation.

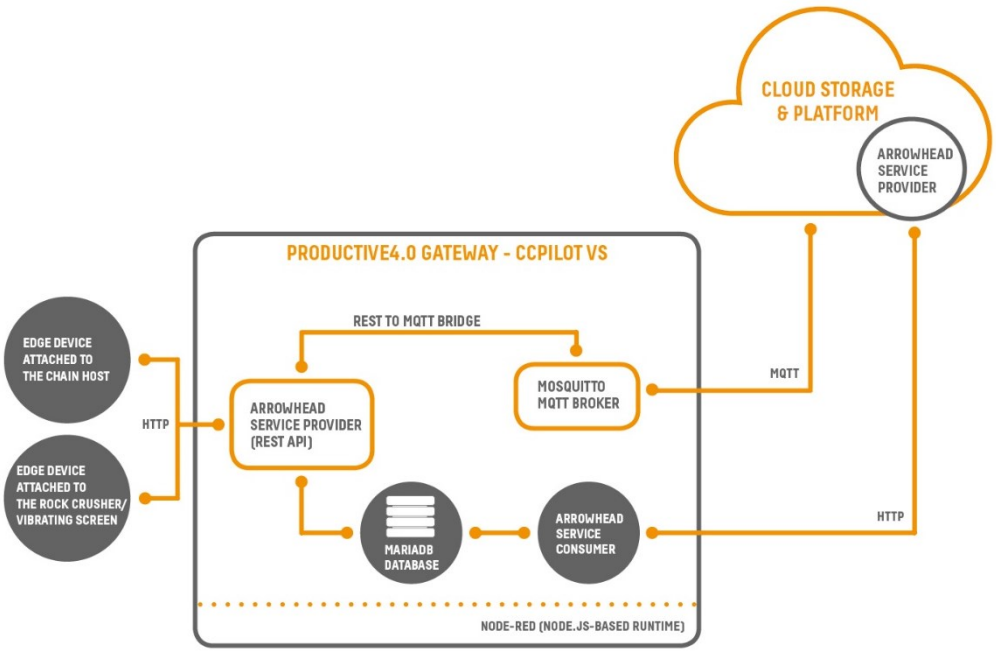


Fig. 4. Data flow in Productive4.0 use cases and gateway device's software and technology stack.

Predictive maintenance practices require data from condition monitoring measurements, including screen movement and acceleration levels. Further, the bearings can also be monitored for excessive noise and heat, therefore requiring multiple types of sensors. Analysis services for the CBM (Condition Based Maintenance) logic can be run at the gateway and accommodated into the Arrowhead framework, enabling also the distribution of warning and alarm messages in the regular communication infrastructure of the pilot.

The data gathered from the condition monitoring shall be analyzed further so that accurate maintenance plans can be generated and the lifetime of the mechanical parts in different environments can be predicted in advance. The optimal solution for data analysis is not always achieved by putting all the computing and data storing to the cloud level. The balance between local processing and cloud computing shall be optimized, leaving the possibility to make adjustments later on when more knowledge of the process is gained.

Further requirements of this use case are being specified, but the general architecture follows the guidelines of the chain hoist use case. The gateway computer, instantiated by CCpilot VS, has a similar role for local data processing, HMI and network connectivity. The display would be probably mounted to the vibrating screen itself, hardening the requirements for its environmental ruggedness.

5 Implementation

This chapter introduces the current state of the implementation for the Productive4.0 lifting business use case.

4.1 Software stack

The outline of the data flow and the software components deployed for CCpilot VS in both Productive4.0 use cases is simplified in figure 4, enabling the gateway functionality of the targeted IoT concept.

The data flow logic is made utilizing a programming tool called Node-RED. It uses a Node.js-based runtime and is run on a web browser. With its interactive user interface consisting of node blocks and wires between them, it enables the design and deployment of the logical data flow subsystem. [3] This solution also supports further customization of the data flow by embedding custom JavaScript code blocks in the process. Node-RED applications running on the VS can exploit Python code blocks as well, through the Node-RED libraries and Python3 interpreter installed on the VS. The core purpose for Node-RED is to convert the messages coming from the edge devices from REST to MQTT. This conversion and transfer component is called as a bridge in figure 4. REST (Representational State Transfer) is a software architectural model for standardizing web services. MQTT (Message Queuing

Telemetry Transport) is a lightweight application layer publish/subscribe protocol laying on top of the TCP/IP protocol stack, with built-in username and password-based authentication.

Mosquitto MQTT broker serves the MQTT clients and conveys their MQTT messages [4]. The MQTT messages can be seen and subscribed to by any actor with access to the MQTT broker. The gateway device is usually required to store data to a local database. MariaDB is a light, open-source SQL-based database, well suited for piloting and testing in an environment of embedded devices typically having limited memory storages [5].

Java runtime environment is also installed on the display in order to develop and deploy the Arrowhead core services and custom Arrowhead-compliant services inside the Java virtual machine. Those services are exploitable also by other local cloud services, or conversely VS can exploit other available Arrowhead services.

4.2 Graphical user interface

The gateway computer is also required to operate as an onboard HMI, with a graphical user interface. The display illustrates select raw measurement signals coming from the sensors, and it can be enabled to acts as a HMI control board. For example, the direction of the movement of the chain hoist can be visualized, including the current load it is lifting. The current version of the UI made for this use case is depicted in figure 5.

Fig. 5. Simple UI for visualizing the measurement data.

CrossControl SDK builds on Qt, with industrially pre-designed graphical widgets for rapid application development. As an open standard, Qt supports cross-platform application development and overall reusability of code.

6 Discussion

The finalization of Arrowhead services at the gateway is in progress, for full implementation of the functionality depicted in figure 4. Data storing as a service will enable the field machines to get a flexible, temporary data storage for the retrieved sensor data before it is conveyed to the cloud. The service consumers will find the storing service from within the local cloud, and in case there would be multiple displays offering the same service, service discovery and load balancing can be applied across multiple gateways. A specific service will be created also for implementing computing for local analytics.

The chain hoist use case enables further data analysis performed at the gateway. For example, specific tasks performed by the chain hoist operator can be recognized and analyzed, and integrated with other available context data, increasing the situational awareness. General task progress can be shown on the display and synchronized with workflow management or ERP in the cloud.

Finally, the security measures need to be considered. As for the time being in this pilot, the data flowing through the gateway is not subject to strong encryption. Industrialization of the concept potentially requires a X.509 certificate based encryption solution, or equivalent.

7 Conclusion

This paper presents an IoT concept architecture, as piloted in Productive4.0 use cases in the scope of fleet management offered as a service. As one of the key components in the use cases, the role of the gateway computer has been discussed in more detail, with requirements for gathering data from various edge devices, data processing and analysis, and dispatching of the data to the cloud and potentially other connected systems.

The use cases also demonstrate CrossControl's industrial display computer running as a gateway and as an onboard HMI in the pilot environment, visualizing select telemetry and dashboard state values of the monitored machine.

8 References

[1] Velte A. T., Cloud Computing: A Practical Approach, New York: McGraw Hill, 2010.

[2] Delsing J., IoT Automation: Arrowhead Framework, Boca Raton: CRC Press, 2017.

[3] Node-RED, March 2019, Available: https://nodered.org/.

[4] Light R. A., Mosquitto: server and client implementation of the MQTT protocol," The Journal of Open Source Software, vol. 2, 2017.

[5] MariaDB, March 2019, Available: https://mariadb.org/.

Industrial IoT applications

Jukka Koskinen*, Petri Tikka ja Hannu Tanner

VTT Technical Research Centre of Finland Ltd

*Corresponding Author: VTT Technical Research Centre of Finland Ltd, Email: Jukka.Koskinen@vtt.fi

Keywords: edge and fog computing, robotics, intralogistics, smart water management, FIWARE

Abstract: In this paper, we present three industrial-oriented IoT applications from industrial robotic, intralogistics and smart water management domains. They utilize Fiware open source IoT platform. We present architecture diagrams of the developed IoT platform applications and share our experiences of the implementation.

1 Introduction

Cloud computing platforms offer new possibilities for data analytics applications. Microsoft Azure and Siemens Mindsphere are examples of commercial cloud computing platforms that offer extensive possibilities for application development and deployment in the cloud, especially in big data storage, analysis and visualization. IoT platforms based on edge computing or fog computing offer possibilities for real-time data streaming, analytics, and visualization in the local networks.

Fiware is an open source platform offering components for analyzing and visualization of data from Internet of Things (IoT) sources. The Orion Context Broker component and the Next Generation Services Interface (NGSI) are the core of Fiware. Orion Context Broker (OCB) provides an NGSI API. OCB publishes context information from entities (virtual real-world objects). Entities can be IoT sensors or application components, etc. OCB publishes context information from context producers (IoT sensors etc.) to context consumers (visualization, analysis components etc.), and supports two-way communication.

Especially SMEs can benefit from Fiware due to its open source approach and ease of deployment of the components. Fiware has a strong European background and is favored by the European Union in many projects funded by EU.

VTT has implemented advanced industrial IoT applications. In this paper, we present three industrial oriented applications from industrial robotic, intralogistics and smart water management domains. They utilize Fiware components.

2 Intralogistics application

Digitalization of the manufacturing industry has created growing interest in digital technologies such as IoT and how the manufacturing floor can be more integrated. The European project Logistics for Manufacturing in SMEs (L4MS) aims to tackle this need by accelerating the automation of intra-factory logistics for SMEs and Mid-Caps. The goal is to reduce the time and installation costs of mobile robots by helping manufacturing SMEs and Mid-caps to develop new, smart intra-factory logistics solutions. The deployment of exceptionally small and flexible logistics solutions requiring no infrastructure change, no production downtime and no in-house expertise aims to make investment in logistics automation more attractive (www.project.l4ms.eu/).

One of the main results that the L4MS project provides is an IoT platform called Open Platform for Innovation in Logistics (OPIL). The OPIL platform will contain the latest navigation, localization, mapping and traffic management services for rapid and cost-effective deployment of logistics solutions. OPIL also enables manufacturing SMEs to conceive highly autonomous, configurable and human-robot logistics solutions according to their business needs.

OPIL architecture consists of three layers: IoT Nodes layer, Cyper-Physical middleware layer and Software systems layer. Components in the IoT node layer interact with the physical world, whereas software applications operate at the level of software systems layer control and monitor OPIL functions. These components communicate with each other by exchanging messages with the middleware layer that decouples components interacting with the outside world, pure software components, and enterprise applications from each other. Components can either

directly operate according to these messages or translate them into a suitable format for internal usage. Here we concentrate on the functionalities of the middleware layer containing the FIWARE components.

Role of Fiware

OPIL utilizes FIWARE in its Cyper-Physical middleware layer, which provides the means for exchanging context information between different modules and functions. This message/context-oriented communication is managed by FIWARE's generic enabler Orion Context Broker (OCB). Communication is managed with HTTP requests. OCB enables the OPIL system to gather, publish, exchange, process and analyze context data. OCB manages information with a publish/subscribe pattern and brokerage of contexts (www.fiware-orion.readthedocs.io/). For maintaining context data OCB uses MongoDB as a database, in which data is stored in Next Generation Services Interface (NGSI) data format. NGSI includes data in entities that hold registrations and subscriptions. The data itself is produced by IoT sources such as sensors, cameras and mobile robots.

Coarsely, mobile robots can be divided into ROS- (Robot Operating System) and non-ROS driven. As a way to enable communication between OCB and mobile robots, OPIL uses FIROS. FIROS is an open source ROS module, which translates ROS messages into FIWARE's NGSI entities. This convention also works from entities to ROS messages (www.wiki.ros.org/firos). In order to avoid duplicate information, a timestamp is also included in the entity. With non-ROS based mobile robots, OPIL uses a custom component to convert ROS messages into a suitable format for the robot.

FIROS is also supplied with sensors in OPIL, but additionally IoT Agent UL 2.0 is used. This agent converts Ultralight 2.0 protocol messages such as HTTP and MQTT into NGSI messages, making it possible for devices using Ultralight 2.0 protocol to communicate with OCB (www.github.com/telefonicaid/iotagent-ul). With device provisioning, IoT Agent can find a specific context entity from Intelligence Data Advanced Solution platform (IDAS).

IDAS is FIWARE's Backend Device Management generic enabler implementation. IDAS implements protocol agents for NGSI entities to allow intercommunication between OCB and generic wireless sensors and actuators. These agents include the above-mentioned IoT Agent UL 2.0 and are part of the IDAS platform (www.forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/IDAS). In OPIL, IDAS is concerned with NGSI context entity management regarding sensors.

Architecture

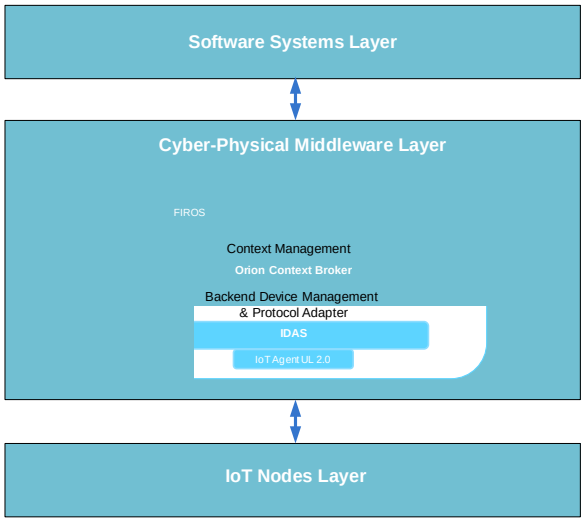


Figure 1. Simplification of OPIL architecture with highlighted FIWARE components.

FIWARE offers a platform with diverse resources to innovate IoT-based technologies. FIWARE's generic enablers are relatively easy to use and understand. In order to get started with OCB and MongoDB, it is necessary to have sufficient hardware. The critical resource regarding MongoDB is RAM memory, as MongoDB performance is related to the amount of available RAM to map database files into memory. Because of the NGSI entity structure, it is up to the user to take care of saving historical data, as entities are always overwritten. In addition, FIWARE is under continuous development, and so some enablers are removed and new ones are brought up. This is both a pro and a con.

3 Robotic application

MIDIH Reference architecture

In the industrial robotic application, the IoT computing platform is based on the MIDIH reference architecture model (http://midih.eu/) (Figure 2). The model provides guidelines for designing of architectures for Cyber Physical System/IoT systems. The functionalities provided by the architecture model are mapped into software components.

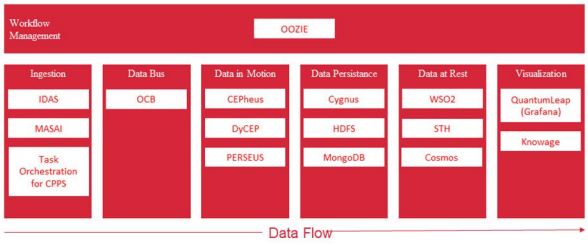


Figure2. MIDIH Industrial IoT and Analytics Platform Architecture (FIWARE4Industry Pipeline) (Benedicto Jesús, 2019, Atos, Internal project document, [Accessed 13.3.2019])

The MIDIH architecture offers two lanes for implementing applications: FIWARE4Industry and APACHE lanes. FIWARE4Industry is an ecosystem that offers FIWARE-based software (open source) components for the manufacturing domain (www.fiware4industry.com). Apache has similar components as the Fiware for developing IoT platforms. APAHCE is also an open source software ecosystem (www.apache.org) and it is supported by the Apache foundation.

The IoT computing platform for the robotics application is based on FIWARE lane.

IoT platform

The IoT platform is running in a virtual machine (Ubuntu 18.04). The architecture is shown in Figure 3. The platform consists of the following Fiware components:

- Orion Context Broker: OCB receives requests from subscribers. It delivers data from sources (sensors) to subscribers/context consumers. (https://fiware-orion.readthedocs.io/en/master/)
- MongoDB: OCB uses MongoDB for storing information about NGSI data entities (registrations, subscriptions). An entity is a virtual representation of real-world objects with attributes.
- Quantum Leap: This component is an OCB subscriber. When a new value arrives to the OCB, Quantum Leap parses and validates the data and stores it in a time-series database. (https://smartsdk.github.io/ngsi-timeseries-api/)
- Crate DB: Crate DB is an SQL Data Base Management System, which supports real time querying and time series data. Crate supports NGSI v2 format. (https://crate.io/)
- Grafana: This is an analytics tool for visualization of time-series data (Figure 4). It supports basic analytics functions such as maximum/minimum, average, smoothing etc. Grafana uses CrateDB (grafana.com)

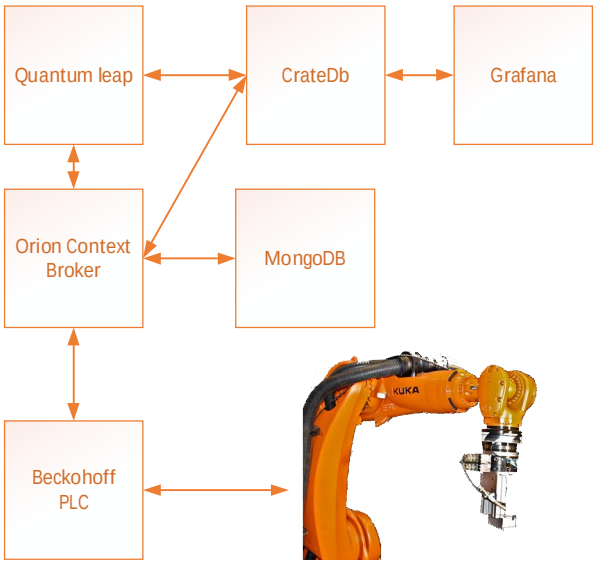


Figure 3. Architecture of the edge computing platform

Figure 4. Visualization of the force sensor data from Grafana

The first implementation was running as Docker containers in Docker desktop application. In this version we had some instability issues. The problem was solved by installing the software from source codes (a virtual machine with Ubuntu 18.04.). With this setup the stable performance was achieved (real time visualization of the sensor data).

4 Smart water management application

The objective of the SWAMP (swamp-project.org) project is to develop an IoT-based smart water management platform for precision irrigation in agriculture. Up to 70% of annual freshwater consumption is related to agriculture, making it one of the fundamental global challenges. The SWAMP platform supports farmers in using water more efficiently, avoiding over- and under-irrigation, as well as decreasing costs and energy consumption. The

platform makes use of measurements made by local weather stations, soil probes, drones, and water monitors, as well as external services such as satellite imagery, to provide the farmer with an optimized plan on how to control each valve and sprinkler head of the irrigation system. With four pilot sites in Europe and Brazil, the project approach is pragmatic and solution-oriented.

The SWAMP Platform development is still ongoing, and the system currently consists of the mainstream components of FIWARE: Orion Context Broker/NGSI-LD Context Broker, IoT Agents, Mosquitto MQTT broker, MongoDB, QuantumLeap, and CrateDB. These components can be deployed with a number of possible configurations, depending on the used IoT devices and communication technologies. The system architecture is categorized into five layers (Figure 5).

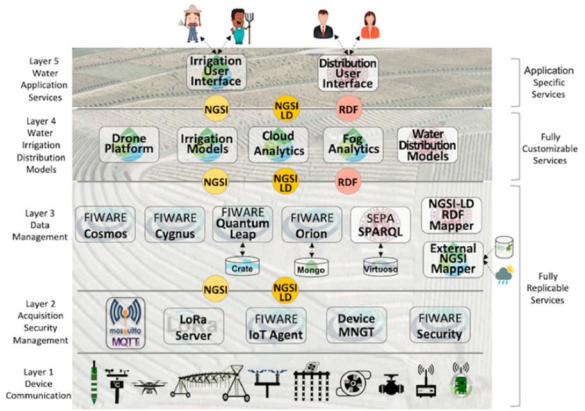


Figure 5. Preliminary Swamp architecture layers.

When the planning of the SWAMP Platform was initiated, a performance analysis for the core FIWARE components was performed. In general, the results suggested that FIWARE should be able to deliver the performance required. The pilots aim for high scalability, which may however require re-engineering of some components as well as special configurations for deployment. The SWAMP Platform is currently under development, and the first version of the system is planned to be launched during autumn 2019.

Jukka Pulkkinen* and Igor Trotskii

Data Strategy in Service Development: Case Study for a Facility Management Service Company Utilizing IoT

Abstract: Digital servitization is a key context for the application of data management, analytics and machine learning in service value creation, thereby contributing to improve competitive position. The recent technology development in the area of Internet of Things and cloud computing provides numerous opportunities for digital servitization, and on the other side, a new approach for service design is needed to utilize these opportunities properly. This paper aims to apply the data strategy framework in service design for facility management services in order to ensure the fulfillment of business requirements. The data strategy framework consists of two phases: from the business requirements to work process and from data to actions. This paper demonstrates the usefulness of a data strategy framework for the development of the facility management service having different types of requirements to manage and analyze the data to reach original business requirements. The main contribution of this paper is to demonstrate how the digital servitization can improve the competitive position of the company by a proper service design using data strategy framework.

Keywords: digital servitization, service design, facility management services, data management, lean service development

***Corresponding Author: Jukka Pulkkinen:** Häme Univeristy of Applied Science, E-mail: jukka.pulkkinen@hamk.fi

Second Author: Häme Univeristy of Applied Science, E-mail: igor.trotskii@hamk.fi

1 Introduction

This paper aims to present a method on how to utilize data in order to improve competitive position on top of the existing service portfolio, applied to one case study for a facility management service company. Servitization is the process in which either

- intangible service components are incorporated to a manufacturing company's existing products or
- pure services are provided to customers.

During the last decades, servitization has been a global trend among many traditional manufacturing companies, with the main interest in creating a competitive advantage in traditional product business. On the other hand, several pure service companies exist, whose offering includes only services without any tangible products, and whose competitive advantages has been based on resilient, in-depth customer relationships. Their services aim to improve the value of their customers' assets. Recently, the competition in this sector has increased due to globalization, digitalization, and newcomers to the local market. Therefore, innovation to improve a competitive position for such local service providers is necessary to ensure the sustainable regional development.

The case study is conducted for facility management services, where the personnel are responsible for operating the building maintenance of multiple buildings owned by their customers. The data management process is applied to the maintenance operation of a building. Actually, the facility management services market is changing in Finland; currently, the market is divided between private and public companies, and there is a trend to move toward private own companies. The winners in the new changing market are the players who can manage their service operation most effectively; therefore, utilizing data for efficiency improvement purposes is a critical success factor. On the other side, the buildings have a lot of data, which could be used to improve the service operation. Naturally, this has not been optimally used yet due to the new technology.

We argue that the data-driven service operation can be used to improve the competitive position by having the right data strategy in place. The competitive position improvement means to improve competitiveness by reducing costs or to improve the value of the customer's assets. Many literature reviews aim to emphasize servitization as one key differentiation factor, especially for manufacturing companies, by bundling the product and services together into one solution. Nevertheless, less attention has been paid to pure service companies, which are not manufacturing

companies by nature, and how they can differentiate themselves by using the data in their market place.

Our research question is: *How is data strategy created and how is datadriven service operation for a pure maintenance company implemented in order to improve the competitive position in their market place?*

In this paper, the data strategy framework is applied to the development of the facility management service. Special attention will be paid to the special business requirements related to facility management services. The purpose of research is to figure out elements which will be used to improve the company's competitive position. The service operation improvement is a new differentiation possibility for facility management service companies.

2 Theoretical background

2.1 Data as a driver for pure maintenance companies

The theoretical background follows the same approach as in Pulkkinen [1], where the service development was shown from the small or medium-sized enterprises' (SMEs) points of view; as a result, the data strategy framework was created. In this paper, the data strategy framework is applied to the pure service companies.

The literature review presented in this chapter is aimed to provide a high-level of understanding of service development and how useful they are in the context of pure service companies. Naturally, improving the competitive position using data differs between product companies and maintenance companies. The core of product companies is their product, and in order to create customer value, they create a solution bundling the products and services together. The solution aims to create benefit for their customer, and this should differentiate them from their competitors and finally to improve their competitive position. The pure service companies do not have the product, and their offering is only services to be provided to their customers. Therefore, their competitive position is based fully on the services and how it creates customer benefit and differentiates them from their competitors.

As described by Pulkkinen [1], the most famous strategic program to improve manufacturing industries' competitive position by using data is the program called Industry 4.0, which originates in Germany and aims to upgrade Germany's industrial capabilities with the help of a smart factory concept [2]. Industry 4.0 has also received much attention in many other countries, where similar programs have been initiated. Industry 4.0 means 4th Generation Industrial Revolution, where "software embedded intelligence is integrated in industrial products and systems" [3]. Thus, Industry 4.0 has been discussed a lot in the literature [3-6] and [7], but these are mainly focused on the industrial manufacturing companies and how to create competitive advantage in their market place. They have very limited experience in the area of pure service companies outside of the manufacturing environment. Actually, we believe that the strong emphasis of Industry 4.0 has moved the focus strongly to manufacturing companies, at the expense of pure service companies and how the data could be used to build benefits for customers and improve the competitive position.

One main reason for the increased amount of data is the development of The Internet of Things (IoT) technology and of ICT technology. This has been the focus of several of the literatures [5, 8] and [9]. A major part of these focus on technical implementation to collect data and neglect the business-level objectives. In addition to this, less focus has been put on service development with the help of IoT and ICT technology. Nevertheless, some papers already exist in this area, discussing how IoT-based solutions are cost effective for improving the competitive position in different areas of service operations [3] and [1].

We argue that the data-driven service operation can be used to improve the service performance in pure service companies by having the right data strategy in their practical implementation. In the current literature, little attention has been paid to pure service companies, which are not the manufacturing companies by nature, and how they can improve their service execution. Therefore, we apply the data strategy framework for a pure service company.

2.1 Data strategy framework

Pulkkinen [1] presented the general data strategy framework. The data strategy consists of two phases:

- Phase 1: from business requirements to work process
- Phase 2: from data to actions

Phase 1 pays a lot of attention to knowledge that is needed to fulfill the business objectives and how this knowledge is utilized in work processes. This knowledge creates the basis for the data-driven service operation; on the other side, the knowledge needs to be directly connected to the business objectives. This phase consists of three steps that are presented in Fig. 1:

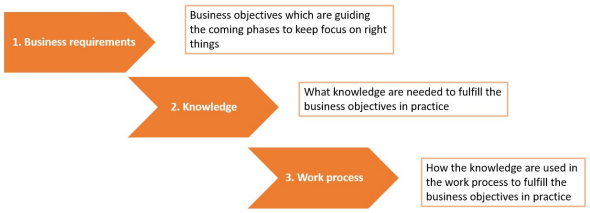


Fig. 1. Three-step model in data strategy framework to create knowledge, starting from business objective and ending with work process. Slightly modified from source [1].

Phase 2 focuses more on technical implementation and restrictions to implement data-driven service operation in practice. Therefore, a lot of attention is paid to data availability and data quality and how to turn data into a positive user experience by using automatic controls, analytics, and machine learning. Phase 2 consists of three phases presented in Fig. 2:

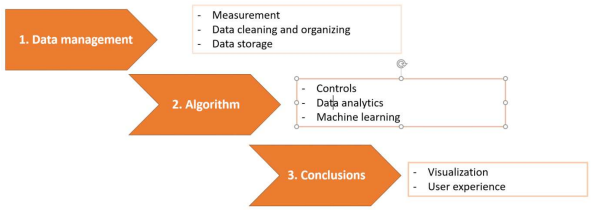


Fig. 2. Three-step model in data strategy framework, from data to conclusion. Slightly modified from source [1].

The presented data strategy framework is a practical method to develop the Proof of Concept (POC) in developing the data-driven services. This way, the feasibility of the solution to fulfill the business requirements can be ensured, and the technical restriction can be identified before the final solution development.

2.2 Lean service development

The lean approach has been applied in many areas, like manufacturing and software development, where different agile methods are popular nowadays. The origin of the lean approach is in Toyota manufacturing and applying the same approach to software development was done later [10] and [11-13]. The lean approach has also been applied in service development [14].

The key idea in the lean approach is to improve efficiency by reducing waste. Poppendieck [15] translated the seven wastes of manufacturing for software development in contrast to operating with a mass production paradigm, which is presented in Table1.

Table. 1. A summary of eliminating waste in manufacturing and software development.

Seven wastes in Manufacturing	Seven wastes of software development
1.Overproduction	1.Extra features
2.Inventory	2.Requirements (e.g. story cards detailed only for current iteration)
3.Extra processing steps	3.Extra steps
4.Motion	4.Finding information
5.Defects	5.Defects
6. Waiting	6.Waiting, including customers
7. Transportation	7.Handoffs

The lean service development combines lean principles, lean software development, and lean service creation, where we start from business objectives following lean principles and using lean service creation methods, and moving toward drastically improved service operation in order improve the competitive position of a pure service company.

3 Methodology

The case study is conducted for facility management service, where the personnel are responsible for operating the building maintenance of multiple buildings owned by their customers. This company is called facility management service provider in the paper.

This study aims to develop an overall understanding of how to create added value with the help of data by the facility management service provider. We accomplish this objective by combining action research consisting of projects related to the facility management services in two cities in Finland. Action research means that the knowledge is created in the context of practice and requires researchers to work with practitioners [16]. To achieve our research objective, researchers in academia are required to participate in real-world cases. Therefore, our researchers have been working in the development project, aiming to create added value for the facility management service providers' customers.

The selected methodology to develop data strategy for the facility management service provider is the data strategy framework according to Pulkkinen et al. [1], where the framework was developed as a result of a case study. The case study was to develop data strategy for SMEs [1], and the result was presented on a general level, fitting to different environments.

4 Results

The results of our research form a data strategy framework for the facility management service provider. The data strategy consists of two phases: the first phase is presented in chapter 4.1 and the second phase in chapter 4.2.

4.1 Data strategy framework, from business requirements to work process

Step 1 – Business requirements: Facility management service is facing a big change in Finland. Big cities have a huge fleet of buildings, and they have had their own department taking care of their facility management services. Now cities have begun to privatize facility management service departments, aiming at cost savings through competitive bidding. The facility management services cover different areas, like technical maintenance, outdoor-area maintenance, facility management, and even energy savings belongs to some companies' offerings. There are different types of companies in the marketplace, like city-owned previous municipal service providers, small and big private-family-owned companies, and big international companies. The offering can vary to some extent, but the core part of their offering is the same, including the areas mentioned above. In addition to this, some companies also have cleaning services, laundry services, and catering services.

The nature of the facility management service is that the company provides services to their customer and their customer owns the buildings that are the object of those services. On the other side, the buildings are big assets to their owner and a bad quality of services may result in big economic losses, even long after the services have been provided. There are many examples of this when bad indoor air quality has ended with a situation where the health of people working in the building is compromised, and in some ultimate cases, the buildings cannot be used for their original purpose anymore. Therefore, the building owners need to carefully select their facility management service provider in order to have the right balance between low cost and high quality. Low-cost service is often emphasized as a short-term target, but the understanding of good quality is spreading among building owners as a result of several bad experiences .

Therefore, the value proposition of facility management service providers consists of minimizing the cost of services and maximizing the facility management value. According to Niemi et.al. [17], the most significant costs for the building owners are heating, including electricity, repairs, administration, and maintenance. The report is related to residential homes and costs may vary between buildings, especially buildings made for different purposes; but we strongly believe the same topics are valid for other buildings, even if the share of the topics may be different.

In order to provide value for their customers, the facility management service provider tries to reduce the costs mentioned above for the building owners. In this research, we focus mainly on the heating and maintenance costs to which the facility management service provider can influence directly. In addition, we will look at the repair costs to which the facility management service provider can indirectly affect.

As a summary, we can state that the value proposition of the facility management service provider is to directly reduce the heating and maintenance costs and indirectly and positively affect repair costs.

Step 2 – Knowledge: Next, we need to create the knowledge needed to reach our goals. First, heating, including the electricity, is a very wide topic, and there are many possibilities to positively affect the heating costs. In this study, we consider two areas: air condition and lighting. Air condition is one big electricity consumer, and it has a direct impact on the heating. The more air is taken out of the building, the more new fresh air is heated, which consumes electricity. In addition to this, reducing unnecessary air conditioning also reduces electricity consumption. The knowledge needed for the right air conditioning is the amount of the people in the room, which is related to the CO2 level. The more people in the room, the higher the CO2 level and vice versa. Therefore, we can state that the knowledge needed for the right air conditioning is the CO2 level in the room. The lighting is also one electricity consumer in the buildings, and the smart lighting systems developed lately have also opened the possibility to reduce electricity consumption. Naturally, the key idea in saving is to reduce lighting when daylight is available, and to make this happen we need to know about the brightness.

The technical maintenance is labor-intensive work, and big cities especially have many buildings requiring several people to take care of their daily maintenance. To gain an understanding of the workload, one big city in Finland has about 20,000 failure messages to manage and repair every year. Therefore, it is obvious that efficient technical maintenance is very important to reduce costs in this area. The good approach to improve efficiency in maintenance tasks is lean service creation presented in chapter 2.2, and applying the seven wastes in manufacturing to lean service creation, we can present following:

Table. 2. A summary of eliminating waste in service creation.

Seven wastes in Manufacturing	Seven wastes of software development	Seven wastes of service creation
1.Overproduction	1.Extra features	1.Extra services not paid by customer
2.Inventory	2.Requirements (e.g. story cards detailed only for current iteration)	2.Services done too often
3.Extra processing steps	3.Extra steps	3.Extra steps
4.Motion	4.Finding information	4.Finding something
5.Defects	5.Defects	5.Don't fulfill customer's expectation
6. Waiting	6.Waiting, including customers	6.Waiting
7. Transportation	7.Handoffs	7.Transportation

Finally, we can state that the needed knowledge to improve efficiency in technical maintenance, while reducing costs at the same time, are extra services not paid by customers, services done too often, extra steps, finding something, waiting, and transportation.

The repair cost is actually the biggest cost element in the study of the building life-cycle costs [17], and we state that Indoor Air Quality (IAQ) indirectly affects this. The reason for our statement is that there are several cases in Finland where bad IAQ has ended with a situation where people cannot use the building anymore, and the owner was forced to make significant repairs to make the building useable again.

IAQ involves many parameters and measurements, and some of those are presented in Table 3. There are certain tolerances to all parameters defined in standard [20], and having parameters inside the tolerances, we can guarantee healthy conditions for the people using the building. So, we can state that the knowledge to avoid unnecessary repairs is the tolerances for the IAQ. On the other side, it is obvious that this is not all of the needed knowledge to avoid unnecessary repairs. and there are several other factors also affecting repairs.

Table. 3 IAQ measurements

Measurement	Description
Temperature	Temperature inside room/building.
CO_2	CO_2 level inside a room/building. Shows room utilization and HVAC system performance.
Humidity	Humidity effects user satisfaction and building condition as in some cases high humidity can cause mold [19].
TVOC	Total Volatile Organic Compounds display purity of the air.
PM2.5/PM10	Small particle measurements show performance of HVAC system and cleaning. Can be thought of as dust level.
Pressure difference	Pressure difference can be used for analyzing air flow inside the building. It is also an important metric in building condition and HVAC system monitoring.

Step 3 – Work process:

In this step, the created knowledge is connected into a work process to reach the original objective defined in step 1, business requirements. First, we can divide the usage of data into two different categories: automatic and manual. Automatic means that traditionally there is a control loop, where a control algorithm automatically controls the process without people. Manual means that people need to be connected to the evaluation of result before conclusion. Nowadays, machine-learning algorithms are taking a bigger role in many environments, and it can help the people make the conclusion, or in some cases, the machine-learning algorithm has even replaced people in the work process.

In our research, the objective to reduce the energy with the help of air conditioning and lighting is a typical example of where automatic control algorithms are used. Nevertheless, this does not mean that people can be completely forgotten. Technical maintenance people need to be aware of such controls, and they need to have the capability to tune and modify the control when needed.

The objective to reduce the maintenance cost through Lean Service Creation is a typical process, where the people are at the center of the process. Therefore, all data needed to optimize the maintenance tasks need

to be presented to maintenance people in the right way and at the right time. In practice, this means that the seven wastes in service creation (extra services not paid by customers, services done too often, extra steps, finding something, waiting, and transportation) need to be analyzed thoroughly, waste by waste. It is then decided how it impacts the work process and the required data is presented to the maintenance people so they can reach a seamless work process with high efficiency.

IAQ is a wide topic and, typically, the problems in IAQ require a comprehensive analysis where several different people and organizations are participating in the evaluation. Therefore, a specific process involving all relevant organizations needs to be defined, and seamless data sharing through the entire process is a very effective method in managing IAQ problems.

As a summary, we can state that the three different objectives, meaning energy savings, reducing maintenance costs, and ensuring IAQ within tolerances, require very different approaches from the work process point of view:

- Energy saving is attained with automatic control, without directly impacting people.
- Reducing maintenance costs requires a new data-driven work process, where the right data presented at the right time to all maintenance people is necessary.
- Problems in IAQ require the involvement of several organizations, and sharing information among them needs to be defined carefully.

4.2 Data strategy framework, from data to actions

Step 1- Data management:

The main goal of the data management phase is to evaluate the data availability for the intended purpose. In the case of a facility management service provider, the data availability is a critical issue, because assets, data sources, data measurements, and even data itself are owned by their customer. Therefore, the facility management service provider is very much dependent on their customer regarding strategically critical elements in their offering. In addition to this, there are typically several technical restrictions and challenges to get high-quality data for the intended purpose. Next, we evaluate our case from this point of view.

Data for the air condition and lighting control are CO2 and brightness correspondingly. They are standard measurements nowadays, and in the case of modern air condition and light control systems, the measurements are already integrated into the system, or they are quite easy to add on as an extra feature. Therefore, technically needed data is a standard feature, but they are tightly integrated to their customer's infrastructure.

Data for the maintenance efficiency improvement is a very wide topic, and it can be divided into two different areas. First, the data from the building and surrounding systems, like building automation systems, needs to be collected. Partly, the data can already exist or some new measurements need to be added. All of this data needs to be collected and stored into data storage, from where it can be used to optimize service execution. The implementation of this can be very expensive and time consuming, requiring deep technical knowledge. Second, the data needs to be collected from or sent to the facility management service provider's own operative IT systems, like ERP, CMMS systems, etc. As a summary, it is a very big effort to make all needed data available for service execution optimization, and it will require several interfaces between different systems. The final solution is integrated into their customer's operative systems and their own IT systems. This can be possible only in the case of a long-term partnership between the facility management service provider and their customer.

IAQ measurements have been developed very fast during the last years. The price has decreased, a battery lifetime has become longer, and the wireless technology is standard for data communication. All of these measurements are easy to install by the facility management service provider, and they are widely used nowadays. The measurement for the temperature, CO2, humidity, and pressure are all proven design technology, but TVOC and PM2.5/PM10 have still more uncertainty regarding trustworthy results.

Step 2 – Controls, analytics, and machine learning:

Regarding energy savings, the needed CO2 and brightness measurements are integrated into the customer's air condition and lighting control systems. A simple controller controls the air condition based on the CO2 levels in the air condition system or in the building automation system. Similarly, the lighting is controlled in a modern light control system. This means that the energy-saving implementation is a small investment, and it can be very difficult to sell this as a service to the customer. Technically, the implementation is very easy and straight forward if air condition and lighting control are implemented with modern technology. The older the technology, the bigger the investment is needed to reach the energy-

saving goals. In our case, the energy savings was 24% in CO2 control compared to constant air conditioning, and 60% in light control compared to continuous lighting. Therefore, reasonable pay-back time can be reached depending on the investment cost. CO2 control implementation needs to take into account the IAQ, because control can have some affect on air quality, and it can even cause some damage to the building in the end. This is the feature that has some value to the customer, and which can be sold as services to the customer on the top of the investment for the energy savings.

Maintenance optimization through lean principle, presented earlier, is a comprehensive approach concluding with new processes to execute service operations. Actually, the scope of our research is to explore the data and how it can be used to improve the service execution efficiency. The measured and stored data is used to create a deviation, and each deviation initiates a work order to eliminate the deviation. The deviation is e.g. an increased pressure difference over a filter in the air condition machine, which indicates the need to replace the contaminated filter. Then, the filter is replaced at exactly the right time, not too early or too late, which is typically the case if it's done according to a time schedule. This is one example of moving from the scheduled maintenance to predictive maintenance, where maintenance action is done based on the real demand and not based on the schedule. Another example is to use a machine-learning algorithm to anomaly detection. This means that the algorithm learns the normal behavior of some measurements, like e.g. water or energy consumption. Then, the algorithm is used to follow these measurements, and it indicates abnormal behavior to further the investigation to the maintenance people. The benefit of this approach is to have the automatic follow-up for the huge amount of measurements, which would be very time consuming and costly to make by people.

IAQ evaluation and deciding corrective actions is a comprehensive process requiring the involvement of several organizations, and a large amount of the data is available for this evaluation. On the other side, the IAQ problems are typically very demanding, and a root cause, as well as corrective actions, requires deep knowledge. Therefore, it is obvious that the analytics can be used to get a better understanding of the situation when a large amount of data is available. The large amount of data has two dimensions: first, the history data for the corresponding building, and second, the data from similar buildings. Nevertheless, the expert knowledge is always needed to make final conclusions, and analytics using the data can only support the expert in gaining a better understanding of the system.

As a summary, we can state that there are very different types of approaches for using data to create value:

- Energy savings is done based on automatic control integrated into the building infrastructure.
- Maintenance operation efficiency improvement based on the new data-driven processes, where simple logic using limit-value or machine-learning algorithms are used
- IAQ problem solving based on the analytics, where analytics have the role to support decision makers

Step 3 – Conclusion:

User experience is very critical to ensure a successful data strategy implementation in practice. This means that the result of analytics & machine learning need to turn into usable information for users. Poor implementation of user interface may easily destroy the value created by analytics & machine learning.

The control for the air condition and the lighting is done automatically without the involvement of people. Therefore, the user interface is not so critical in this case, and technical maintenance people need to have access to the system for maintenance and tuning purposes. On the other side, the energy-saving information is very interesting, and it can be presented to the customer, especially if the solution has been sold as aiming at cost savings.

The user interface plays an important role in data-driven maintenance operations aiming at efficiency improvement. The work orders generated as a result of analytics need to be presented seamlessly to technical maintenance people. The mobile user interface is the best solution for this. This user interface needs to be very clear, precise, and easy to use by the technical maintenance people.

Experts do IAQ problem solving and people from several organizations participate in the evaluation. Therefore, the user interface needs to be more flexible and needs to provide the possibility to analyze the situation using data from different points of view. There are different analytics tools providing a good user interface, like PowerBI and Tableau, developed for this purpose.

5 Discussion

We can state that the data strategy framework presented in Pulkkinen et al. [1] is applicable for the

facility management service provider. Actually, the different areas for creating added value were the reduction of heating, including electricity, and maintenance costs, also positively affecting repair costs. These areas require a very different data management process but nevertheless, we can state that the data strategy framework was applicable to this case study regardless of the different nature of these areas. This can be seen as the strength of the presented data strategy framework.

We want to highlight a few observations discovered during the study. The data quality is a very critical success factor. Our approach was emphasizing that the data is correct from the business requirement point of view, and even this was quite easy to forget during the process. So, one thing we learned is to make it very clear from the beginning, our business objective, and keep this in mind through the whole process. On the other side, the data quality from the technical point of view was also very critical, and it's important to check this from different points of view, like completeness, timelines, validity and accuracy. This is one area which needs further studies to improve the overall data strategy framework process.

In our study, the measurable results were identified for the energy savings, like air condition and lighting optimization. The air condition optimization was done in an old school building, and the lighting optimization at the sheet metal center. Naturally, the results were achieved in these specific environments, and the extendibility of results to other applications need further research. On the other side, our objective was not to figure out scientifically accepted energy-saving opportunities, but only to demonstrate the applicability of the data strategy framework in this context.

One our results was the seven wastes of services derived from the lean manufacturing principle. This was not yet implemented in the real environment and, therefore, this needs further research to demonstrate the validity. Actually, this is the most challenging objective, where several aspects need to be taken into account, like old maintenance traditions , current IT systems used in the company, and how to manage the information in a modern environment, which includes legacy systems, etc. So, the maintenance efficiency improvement can be seen as a bigger challenge in the company to make the digital transformation, which will require strong leadership to manage the whole transformation. Definitively, this is one important research topic in the future, and this would serve the needs of many companies nowadays.

6 Conclusion

Many companies have been adopting a service business strategy, especially in the mature industries, in order to differentiate their offering and enhance customer engagement for decades already. Therefore, servitization has been a global trend among many traditional manufacturing companies, and a great deal of research has been made for service development. However, until now, such research has not been made for the pure service companies, having only service offerings without manufacturing.

Our research focuses on improving the competitive position utilizing data for the facility management service providers, who are providing pure services for their customers. The research was based on action research with the facility management service providers in a changing market. The facility management service provider has a strategy of sustainable growth in the long term, which requires improving their competitive position in their market. The competitive position improvement is created by utilizing the data in their service operation to reduce the heating and maintenance costs and to positively affect repair needs.

The result of the study is that the selected data strategy framework was applicable in the area of the facility management services. The facility management service provider has different areas to improve the competitive position, as mentioned above, and they result in very different data-driven services. However, the framework was applicable to all these areas, emphasizing the strength of the data strategy framework to be applicable in different environments.

Acknowledgements

This work was supported by the European Regional Development Fund.

References

[1] Pulkkinen, J., Jussila, J. Partanen, A., Trotskii, I.: Data Strategy Framework in Servitization: Case study of service development for a vehicle fleet. Rii forum, (2019)

[2] Heng, S.: Industry 4.0-Upgrading of Germany's industrial capabilities on the horizon. (2014).

[3] Rymaszewska, A., Helo, P., Gunasekaran, A.: IoT powered servitization of manufacturing–an exploratory case study. Int. J. Prod. Econ. 192, 92–105 (2017).

[4] Heng, S.: Industry 4.0-Upgrading of Germany's industrial capabilities on the horizon. (2014).

[5] Lee, J., Kao, H., Yang, S.: Service innovation and smart analytics for industry 4.0 and big data environment. Procedia Cirp. 16, 3–8 (2014).

[6] Schmidt, R., Möhring, M., Härting, R.-C., Reichstein, C., Neumaier, P., Jozinović, P.: Industry 4.0 - Potentials for Creating Smart Products: Empirical Research Results. In: International Conference on Business Information Systems. pp. 16–27. Springer (2015).

[7] Wang, S., Wan, J., Li, D., Zhang, C.: Implementing Smart Factory of Industrie 4.0: An Outlook. Int. J. Distrib. Sens. Networks. 12, 3159805 (2016).

[8] Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of Things (IoT): A vision, architectural elements, and future directions. Futur. Gener. Comput. Syst. 29, 1645–1660 (2013).

[9] Abdul-Qawy, A.S., Pramod, P.J., Magesh, E., Srinivasulu, T.: The Internet of Things (IoT): An Overview. Int. J. Eng. Res. Appl. 1, 71–82 (2015).

[10] Ries, E.: The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses. Crown Books (2011).

[11] Poppendieck, M., Poppendieck, T.: Lean Software Development: An Agile Toolkit: An Agile Toolkit. Addison-Wesley (2003).

[12] Abrahamsson, P., Salo, O., Ronkainen, J., Warsta, J.: Agile Software Development of Mobile Information Systems. VTT publication 478, Espoo (2002).

[13] Womack, J.P., Jones, D.T.: Lean Thinking—Banish Waste and Create Wealth in your Corporation. J. Oper. Res. Soc. 48, 1148–1148 (1997).

[14] Sarvas, R., Nevanlinna, H., Pesonen, J.: Lean service creation. The Handbook V1.8. Futurice (2017).

[15] Poppendieck, M.: Principles of lean thinking. IT Manag. Sel. 18, 1–7 (2011).

[16] Huang, H.,B.: What is good action reseach? Why the resurgent interest? Action research. 1(8), 93-109 (2010)

[17] Niemi,J., Hietala, M., Kaleva, H.: ARA-talojen hoitokulut ja kulurakenne. Asumisen rahoitus- ja kehittämiskeskuksen raportteja (2014)

[18] Sosiaali- ja terveysministeriön asetus asunnon ja muun oleskelutilan terveydellisistä olosuhteista sekä ulkopuolisten asiantuntijoiden pätevyysvaatimuksista (2015). https://www.finlex.fi/fi/laki/alkup/2015/20150545#Pidp446002528

[19] Kenneth M. Elovitz, P. E. (1999). UNDERSTANDING WHAT HUMIDITY DOES. *ASHRAE*, (April), 75–81.

Veli-Pekka Pyrhonen* and Matti Vilkko

Improving tracking performance of composite nonlinear feedback controllers via new reset and hold feature of nonlinear functions

Abstract: This paper proposes new reset and hold feature for the nonlinear functions within composite nonlinear feedback (CNF) controllers. Reset and hold feature helps closed-loop control system to accurately track e.g., step input sequences by resetting the initial value of the controlled output whenever an individual step reference changes in value. The new reset and hold feature work independently of the chosen nonlinear function for all CNF controllers. If CNF controllers are used without the revisions proposed in this paper, then the command following ability of the closed-loop control systems may significantly degrade. Furthermore, they may also use excessive amount of control authority, which may result in actuator saturation and other practical problems. The simulation and experimental results show that the closed-loop control systems designed using the refinements of this paper provide better tracking performances both in steady-state and during transients compared with the control systems without them.

Keywords: Control Applications, Constrained Control, Linear Systems, Nonlinear Control, Servo Systems

***Veli-Pekka Pyrhonen:** Tampere University, Tampere, Finland, E-mail: veli-pekka.pyrhonen@tuni.fi

Matti Vilkko: Tampere University, Tampere, Finland, E-mail: matti.vilkko@tuni.fi

1 Introduction

Composite nonlinear feedback (CNF) is relatively well-known control design methodology that attempts to achieve simultaneous fast command following and robustness under limited control authority. The CNF was originally proposed by Lin et al. [1] for a class of second order systems. Chen et el. [2] generalized the results of [1] to cover more general systems with measurement feedback. Multivariable case is studied in [3–4], whereas CNF control for a class of nonlinear systems is studied in [5]. Furthermore, Cheng et al. [6] have generalized CNF control for tracking more general nonstep references, whereas Pyrhonen [7] provided design framework for improving the transient stage of CNF control using general dynamic feedforward set point filters.

The CNF methodology produces feedback controllers that consists of parallel-connected linear and nonlinear parts, which are designed as follows. First, the linear part is designed by placing the dominant pair of the closed-loop poles with small damping ratio, which would result in a step response having short rise time but large overshoot. Then, the nonlinear part is designed such that the damping ratio of the dominant pair smoothly increases, when the control error gradually diminishes. Because of such mechanism, the step response of the closed-loop system maintains short rise time, while the overshoot tendency caused by the linear part is eliminated. Overshoot is eliminated, because CNF controllers are able to use significant amount of control at the late stage of the transient response, which eventually shortens the settling time of closed-loop systems. This is the key property of all CNF controllers, which makes CNF methodology feasible for many servo control applications requiring fast and precise command following, see for example: [8–15].

However, the proposed nonlinear functions of CNF work best in single step experiments, because they are parameterized such that appropriate scaling is obtained when step is changed. Despite of scaling, the nonlinear functions may not work well, if step sequences with varying magnitudes are used as input commands. The reason for performance degradation is the invariable initial condition of the controlled output inside the scaling parameter. That is, when the reference input is changed, the initial condition remains fixed, which means that the scaling parameter is indeed inherently reset, but the reset value may have an offset.

In this paper, new reset and hold feature is introduced for the scaling parameter such that satisfactory command following is enabled when step sequences are used as reference inputs. The material in this paper is organized as follows. In Section 2, the design procedure of the CNF control law with the revised nonlinear function is presented. In Section 3, the performance of the revised nonlinear function is demonstrated by a design example in which the angular position of a rotary servo system is controlled. Finally, in Section 4, some concluding remarks are drawn.

2 CNF control design procedure

Consider the following class of SISO (Single-Input-Single-Output) systems with input nonlinearity

$$\begin{cases} \dot{x} = Ax + B\text{sat}(u) \\ y = C_y x \\ m = C_m x \end{cases}, \; x(0) = x_0, \tag{1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}$, $y \in \mathbb{R}$ and $m \in \mathbb{R}^p$, $p \leq n$, are the state, control input, controlled output and measured output, and where x_0 is an initial condition. The input nonlinearity in (1) is represented by

$$\text{sat}(u) = \min\{u_{\max}, |u|\}\,\text{sgn}(u), \tag{2}$$

where $u_{\max}$ is the saturation limit of the input and sgn denotes the sign function. Furthermore, the following requirements on the system matrices of (1) must be satisfied:

A1: the pair (A, B) is stabilizable;
A2: the triple (A, B, C_y) is invertible and has no invariant zeros at the origin;
A3: the pair (A, C_m) is detectable;
A4: the controlled output y is a subset of m i.e. y is also measured.

Next, a step-by-step design procedure for CNF control is presented. The procedure is partitioned in two separate steps, which are: the design of a linear state feedback part, and the design of a nonlinear state feedback part.

2.1 Design of linear state feedback part

First, assume that $C_m = I$, i.e. that all states of (1) are measured and available for feedback. Furthermore, assume that A1 and A2 are satisfied. Then design a linear full-state feedback law

$$u_L = -K_L x + R_s r, \tag{3}$$

where K_L is the full-state feedback gain and r is the target step reference. The gain K_L must be chosen such that 1) all eigenvalues of the matrix $(A - BK_L)$ have strictly negative real parts, and 2) the closed-loop system $C_y(sI - A + BK_L)^{-1}B$ has small damping ratio. The selected feedback gain K_L results in the following scalar-valued feedforward gain

$$R_s = -[C_y(A - BK_L)^{-1}B]^{-1}, \tag{4}$$

which ensures that the DC-gain for the model-based linear closed-loop system from the target reference r to the controlled output y is one.

2.2 Design of nonlinear state feedback part

First, compute the value of the desired state x_d using

$$x_d \triangleq R_d r = -(A - BK_L)^{-1}BR_s r. \tag{5}$$

Then form a parallel-connected CNF control law

$$u = u_L + u_N, \tag{6}$$

where u_N is the nonlinear feedback component given by

$$u_N = \rho(r, y)K_N[x - x_d] = \rho(r, y)B^T P[x - x_d] \tag{7}$$

with $P = P^T > 0$. The function $\rho(r, y)$ is any nonpositive function locally Lipschitz in y, which is used to smoothly increase the damping ratio of the closed-loop system when its output y approaches the target step reference r. The matrix P can be computed by solving the Lyapunov equation

$$(A - BK_L)^T P + P(A - BK_L) + Q = 0 \tag{8}$$

for a given $Q = Q^T > 0$. The solution P always exists since all eigenvalues of $(A - BK_L)$ are in the left-half complex plane.

The following theorem from [2] provides important stability properties for the closed-loop control system consisting of the system (1) and the CNF control law (6).

Theorem 1. Consider the system (1), the linear feedback control law (3), and the composite nonlinear feedback control law (6). For any $\delta \in (0, 1)$, let $c_\delta > 0$ be the largest positive scalar satisfying

$$|K_L x| \leq u_{\max}(1 - \delta), \; \forall x \in \{x : x^T P x \leq c_\delta\} \triangleq S. \tag{9}$$

Then, the linear closed-loop control system consisting of (1) and (3) tracks a step input r without saturating the actuator provided that $x(0)$ and r and satisfy:

$$\tilde{x}(0) \triangleq (x(0) - x_d) \in S \text{ and } |Hr| \leq \delta u_{\max}, \; \forall t, \tag{10}$$

where

$$H \triangleq -[1 + K_L(A - BK_L)^{-1}B]R_s. \tag{11}$$

Moreover, for any $\rho(r, y)$ as discussed above, the composite nonlinear feedback law in (6) is able to asymptotically track a step input r provided that (10) is satisfied.

Next, a suitable function $\rho(r, y)$ needs to be chosen that is used to increase the damping ratio of the closed-loop system when $e \to 0$. In this paper, the following nonlinear function is used, which was originally proposed by Lin in [1] and later revised by Lan in [16]

$$\rho(e) = -\beta \exp(-\alpha\alpha_0|e|), \; |e| = |r - y|, \tag{12}$$

with

$$\alpha_0 = \begin{cases} 1/|r - y(0)|, & y(0) \neq r \\ 1, & y(0) = r \end{cases}, \tag{13}$$

where α_0 is a scaling parameter. The role of the scaling parameter α_0 is to widen the applicability of fixed ρ for nonunit step references. That is, the first condition in (13) makes the value of α_0 dependent on the size of the given r when the initial condition $y(0) \neq r$. Note that

division by zero is avoided by the second condition of (13) when $y(0) = r$. Note also that the nonlinear function in (12–13) is implementable as A4 is satisfied.

The tuning parameters $\alpha > 0$ and $\beta > 0$ are chosen such that the closed-loop control system satisfies the desired transient performance requirements i.e. short settling time and small overshoot. The function (12) has the following convenient property

$$\lim_{t\to\infty}\rho(e) \to \rho(0) = -\beta. \tag{14}$$

Hence, the parameter β can be selected e.g., to yield the desired damping ratio of the dominant pair at the steady-state situation.

However, the functions that have been proposed so far within the CNF framework do not necessarily work well, if the input command is composed of more than a single step signal e.g., if the input is a step sequence. The main drawbacks with the previously proposed functions are that they are unable to 1) reset the initial condition $y(0)$ when necessary, and 2) hold the reset initial condition to ensure satisfactory performance, when the output of the system is commanded towards different reference values.

To overcome the above-mentioned drawbacks, the scaling parameter α_0 in (13) is supplemented by the following new rule

$$\alpha_0 = \begin{cases} 1/|r - y(0)|, \; y(0) \neq r, & \begin{cases} \Delta r \neq 0 \Rightarrow y(0) = y(t_n) \\ y(0) = y(t_{n-1}), \text{ otherwise} \end{cases} \\ 1, \qquad y(0) = r \end{cases}. \tag{15}$$

The first condition of the rule (15) states that the initial condition $y(0)$ is reset to the current measured value of the controlled output $y(t_n)$ whenever the target step reference changes in value at some time instant t_n. The second condition states that $y(0)$ is hold at the previously set value $y(t_{n-1})$ otherwise. It should be noted that the above rule can be applied to other forms of commonly-used nonlinear functions, see e.g., [2, 11, 16] that are suitable for CNF control. The effectiveness of the rule (15) is demonstrated in subsections 2.3 and 3.3.

Nonetheless, it can be shown that the closed-loop state-error system using the CNF control law (6) is given by

$$\dot{\tilde{x}} = (A - BK_L + \rho(e)BB^TP)\tilde{x}, \; \tilde{x} = x - x_d, \tag{16}$$

because $Ax_d + BR_d = 0$. Proof, see for example [2]. Therefore, the closed-loop eigenvalues can indeed be changed by ρ when e decreases.

Remark 1. When $p < n$, a measurement feedback CNF controller can be designed for (1) if A3 is satisfied. Specifically, if the system (1) can be partitioned as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix} \text{sat}(u), \tag{17}$$
$$y = C_y x, \; m = C_m x$$

then, a reduced-order measurement feedback CNF controller can be designed, which is given by

$$\begin{aligned} \dot{x}_v &= (A_{22} - L_R A_{12})x_v + (B_{21} - L_R B_{11})\text{sat}(u) \\ &\quad + (A_{21} - L_R A_{11} + (A_{22} - L_R A_{12})L_R)m \\ u &= K_L \begin{bmatrix} m \\ x_v + L_R m \end{bmatrix} + R_s r \\ &\quad + \rho(r, y)B^T P \left[\begin{pmatrix} m \\ x_v + L_R m \end{pmatrix} - x_d \right], \end{aligned} \tag{18}$$

where the gain L_R must be selected such that the eigenvalues of $(A_{22} - L_R A_{12})$ have strictly negative real parts. Interested readers may refer e.g., to [2] for full-order measurement feedback case.

2.3 An illustrative example

Consider the following system

$$\begin{cases} \dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \\ y = [1 \quad 0]x, \; m = Ix \end{cases} \tag{19}$$

and the full-state CNF control law

$$u = K_L x + R_s r - \beta \exp(-\alpha\alpha_0 |e|)B^T P(x - x_d), \; e = r - y, \tag{20}$$

$$K_L = [9 \quad 0], \; R_s = 10, \; P = \begin{bmatrix} 20 & 2.5 \\ 2.5 & 1.5 \end{bmatrix}, \; \alpha = 15, \; \beta = 18.5. \tag{21}$$

The scaling parameter α_0 is implemented using (13) and (15) for comparison. The responses of the closed-loop control systems for a step sequence consisting of a downward step and a consecutive upward step have been depicted in Fig. 1. Judging from Fig. 1, the tracking performances of the CNF systems with the original and revised nonlinear functions are the same for the downward step. However, for the upward step, the original CNF cannot hold performance, because the profile of the nonlinear function becomes deteriorated, and hence, it yields over 20% overshoot. Moreover, it also results in a large momentary peak in the controller's output, which could cause actuator saturation, larger overshoot, longer settling time and other practical problems. Conversely, the revised CNF yields fast and strictly monotone response for the upward step, which is result from the new reset and hold feature of (15). The output response using the strictly linear feedback control u_L as in (3) has also been included in Fig. 1. Note that CNF controllers maintain the short rise time of the lightly-damped linear system.

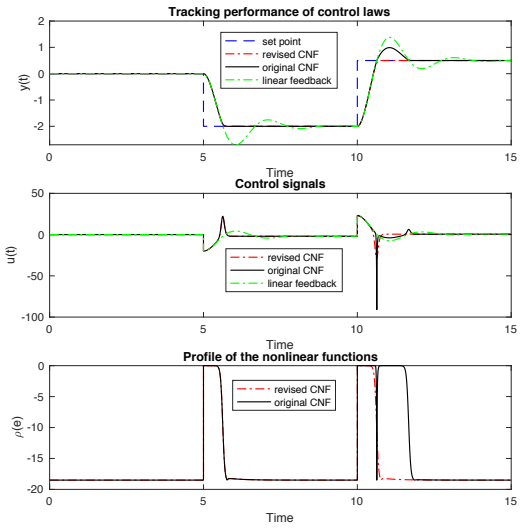


Fig. 1. Tracking performances of closed-loop systems, feedback control signals and profiles of ρ.

3 Design example

In the previous section, the scaling parameter of the nonlinear functions within CNF controllers has been revised to improve the tracking performance of closed-loop control systems in practical design tasks. In this section, the performance of the revised CNF controller is demonstrated by a design example in which the angular position of a rotary servo system is controlled. Here, the rotary servo system is a Quanser QUBE-Servo 2 unit with a metal disc attachment, which is depicted in Fig. 2.

QUBE-Servo 2 unit uses a small direct-drive 18V brushed DC motor (Allied Motion CL40 model 16705) to drive the motor shaft and the attached load to desired positions, or to desired angular velocities. The unit is equipped with an optical relative single-ended rotary shaft encoder (US Digital model E8P-512-118) for accurate angle measurements. Furthermore, the motor is powered by a Pulse-Width Modulation (PWM) amplifier, which receives commands from the integrated Data Acquisition (DAQ) device. The DAQ communicates with PC via USB connection. In this paper, feedback controllers are built in Matlab/ Simulink environment, which has been supplemented by Quanser Real-Time Control (QUARC) software (version 2.5). The fundamental sample time of QUARC has been kept at the default value of 1 ms.

3.1 Mathematical model of DC motor

The mathematical model of the motor with the disc load based on first-principles modeling can be found e.g., in [17]. Here, a device specific model is obtained via open-loop step experiment, and a suitable model will be fitted according to the measured response data. For such a purpose, a square wave that alternates between 1 V and 3 V is fed as an input to the DC-motor. The input is strong enough to overcome static nonlinearities such as friction forces occurring in the motor assembly. The response data from the open-loop experiment is displayed in Fig. 3. According to Fig. 3, the following first-order model from the input voltage to angular velocity

$$G_{u\omega}(s) = \frac{K}{\tau s + 1} \tag{22}$$

with the time constant τ = 0.173 s and the DC-gain K = 24.2 rad/(s V) fits well to the experimental data. To obtain the model from the input voltage to angular position, an integrator needs to be attached to the model (22):

$$G_{u\theta}(s) = \frac{K}{s(\tau s + 1)}. \tag{23}$$

Next, the model (23) is converted to a state-space representation that is suitable for CNF control design:

$$\begin{bmatrix} \dot{\theta} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1/\tau \end{bmatrix} \begin{bmatrix} \theta \\ \omega \end{bmatrix} + \begin{bmatrix} 0 \\ K/\tau \end{bmatrix} \text{sat}(v_m), \tag{24}$$

where θ is the angular position, ω is the angular velocity and v_m is the control voltage, which is limited by ±15 V.

Fig. 2. QUBE-Servo 2 system and a metal disc load.

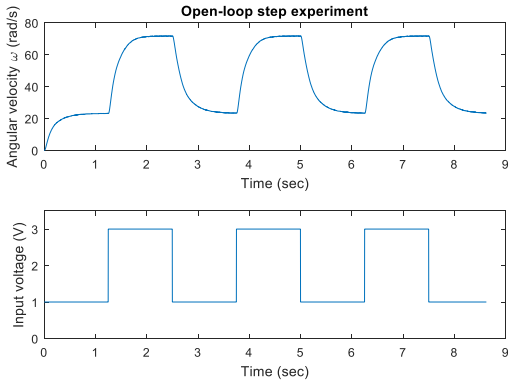


Fig. 3. Open-loop experiment for model fitting.

3.2 CNF control design

The CNF control system design begins from the results of Subsection 2.1. Here, controller tuning is chosen such that it yields good performance in real-time experiments. Because a second-order model captures the dominating dynamics of the DC motor, it is advisable to parameterize the gains of the linear feedback law in (3) as

$$K_L = \begin{bmatrix} K_1 & K_2 \end{bmatrix} = \frac{\tau}{K}\begin{bmatrix} \omega_0^2 & 2\zeta_0\omega_0 - 1/\tau \end{bmatrix} \qquad (25)$$

and

$$R_s = -[C_y(A - BK_L)^{-1}B]^{-1} = \omega_0^2\frac{\tau}{K} = K_1. \qquad (26)$$

The parameterized gains allow the designer to choose an appropriate initial damping ratio ζ_0 and an initial natural frequency ω_0. The parameters ζ_0 and ω_0 are chosen such that the step response of the resulting linear closed-loop system has short rise time and large overshoot, and that the control input u_L does not cause actuator saturation. Here, the initial poles of the closed-loop control system are placed at $s_{1,2} = -10 \pm j30$. The chosen poles yield $\zeta_0 \approx 0.3162$ and $\omega_0 \approx 31.6228$, which give $K_L \approx [7.1488 \;\; 0.1017]$ and $R_s \approx 7.1488$.

In what follows, the nonlinear feedback part is designed using the procedure explained in Subsection 2.2. The Lyapunov equation (8) is solved with Q = diag(17, 1) which gives

$$P \approx \begin{bmatrix} 25.5950 & 0.0085 \\ 0.0085 & 0.0252 \end{bmatrix},\; P = P^T > 0. \qquad (27)$$

The gain of the nonlinear part is then

$$K_N = B^T P \approx [1.1890 \;\; 3.5566]. \qquad (28)$$

Finally, the tuning parameters of the nonlinear function must be chosen such that the overshoot caused by the linear feedback part is automatically reduced by the nonlinear controller when $e \to 0$. However, the tuning parameters must be chosen with care in order to ensure that the actuator limits are not reached when the error becomes small. The following values have been assigned to the tuning parameters: α = 6.1 and β = 0.15.

Unfortunately, only the angular position θ is measured in real-time experiments, that is $m = \theta$. Therefore, the final control law is implemented using (18), which constructs the angular velocity estimate $\hat{\omega}$. The gain of the reduced-order observer has been set to L_R = 150, which completes the design of the CNF control law. The resulting CNF controller is implemented using the following equations

$$\dot{x}_v = -155.7803x_v + 139.8844\,\mathrm{sat}(v_m) - 23367\theta$$

$$v_m = -[7.1488 \;\; 0.1017]\left[\begin{pmatrix} \theta \\ x_v + 150\theta \end{pmatrix} - x_d\right] + 7.1488\theta_r + \rho(e)[1.1890 \;\; 3.5566]\left[\begin{pmatrix} \theta \\ x_v + 150\theta \end{pmatrix} - x_d\right], \qquad (29)$$

$$x_d = [1 \;\; 0]^T\theta_r,\; x_v(0) = 0,\; \theta(0) = m(0) = 0,\; x_v + 150\theta = \hat{\omega}$$

with

$$\rho(e) = -0.15\exp(-6.1\alpha_0|e|),\; e = \theta_r - \theta, \qquad (30)$$

where

$$\alpha_0 = \begin{cases} 1/|\theta_r - \theta(0)|, & \theta(0) \neq \theta_r, \\ 1, & \theta(0) = \theta_r \end{cases} \begin{cases} \Delta\theta_r \neq 0 \Rightarrow \theta(0) = \theta(t_n) \\ \theta(0) = \theta(t_{n-1}),\ \text{otherwise} \end{cases}. \qquad (31)$$

All initial conditions of (29)–(31) have been set to zero, because the shaft encoder provides relative angle measurements from the actual device i.e. the measured angle will always start from zero despite the absolute initial position of the shaft and disc load.

3.3 Experimental results

In this subsection, the CNF control law in (29)–(31) is tested with the DC-motor application. The tracking performance of the refined control law is compared with the original CNF law in steady-state and during transients. In the experiment, a step sequence that traverses in between 0 deg and 200 deg is used as a reference input. An individual step command is changed both in magnitude and direction at more or less random time instances. The percent overshoot/ undershoot, settling time within ±2% margins, and steady-state error are used as the main criteria for performance evaluation.

The results of the experiments are depicted in Fig. 4 and Fig. 5, respectively. The settling time and the steady-state error using the revised CNF are 54.3 milliseconds and 0.69 degrees, respectively. Hence, the response of the refined CNF control law is fast and accurate. The main reason for good performance is the profile of the revised nonlinear function that automatically resets correct value for $y(0)$, and hence, updates α_0, which appropriately scales the nonlinear function during each step change. In contrast, the output response of the original CNF starts to experience unwanted transients and steady-state errors from the second step onwards. The maximum undershoot > 30%, which occurs just after the downward input step at 4 seconds. The maximum steady-state error for the original CNF is 3.5 degrees, and hence, the response of the closed-loop system does not always even settle within ±2% margins. Clearly, the tracking performance is unacceptable, and it is caused by unsuitable scaling. Note that the actuator limits are exceeded several times using the original CNF, but the revised CNF keeps the control under the given limits at all times.

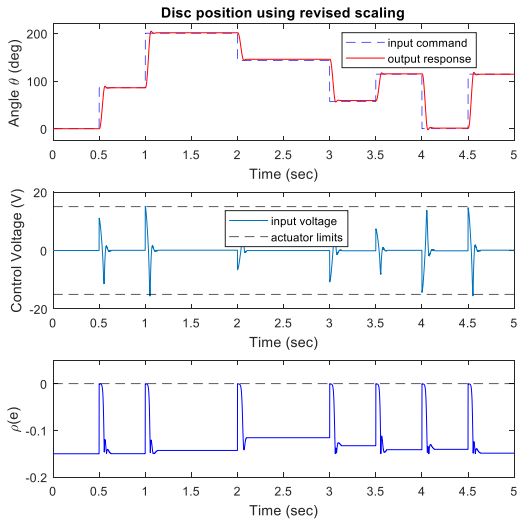


Fig. 4. Tracking performance, control input and profile of revised ρ.

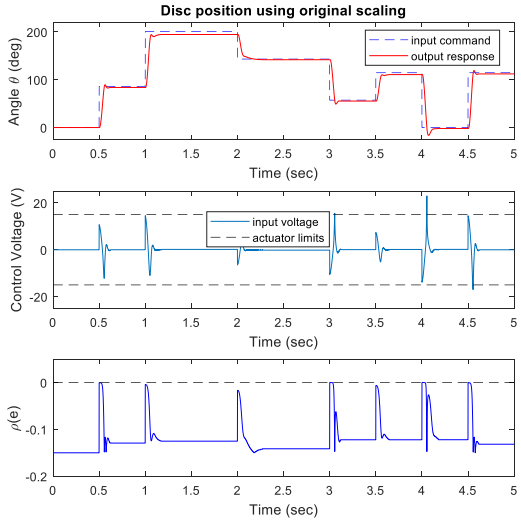


Fig. 5. Tracking performance, control input and profile of original ρ.

4 Concluding remarks

In this paper, new reset and hold feature was introduced for the scaling parameter of the nonlinear functions of composite nonlinear feedback controllers. To be more specific, the initial condition of the controlled output within the scaling parameter is now correctly reset, when step sequences are used as reference inputs. This helps closed-loop control systems to maintain good transient performance despite of variations in input magnitudes, while it also keeps control actions within the designed limits. The performance improvement obtained by the new feature was demonstrated using simulations and real-time experiments.

References

[1] Lin Z., Pachter M., Banda S., Toward improvement of tracking performance – nonlinear feedback for linear systems, Int. J. Contr., 1998, 70, 1–11

[2] Chen B.M., Lee T.H., Peng K., Venkataramanan V., Composite nonlinear feedback control for linear systems with input saturation: theory and an application, IEEE Trans. Autom. Contr., 2003, 48, 427–439

[3] Turner M.C., Postletwaite I., Walker D.J., Nonlinear tracking control for multivariable constrained input nonlinear systems, Int. J. Contr., 2000, 73, 1160–1172

[4] He Y., Chen B.M., Wu C., Composite nonlinear control with state and measurement feedback for general multivariable systems with input saturation, Syst. Contr. Lett., 2005, 54, 455–469

[5] Lan W., Chen B.M., He Y., On improvement of transient performance in tracking control for a class of nonlinear systems with input saturation, Syst. Contr. Lett., 2006, 55, 132–138

[6] Cheng G., Peng K., Chen B.M., Lee T.H., Improving transient performance in tracking general references using composite nonlinear feedback control and its application to high-speed XY-table positioning mechanism, IEEE Trans. Ind. Electron., 2007, 54, 1039–1051

[7] Pyrhonen V.-P., Koivisto H. J., On improvement of transient stage of composite nonlinear feedback control using arbitrary order set point filters, Proceedings of the 4th Annual IEEE International Conference on Control Systems, Computing and Engineering (28 November – 30 November 2014, Penang, Malaysia), IEEE, Penang, Malaysia, 2014, 147–152

[8] Hou Z., Fantoni I., Interactive leader-follower consensus of multiple quadrotors based on composite nonlinear feedback control, IEEE Trans. Contr. Syst. Technol., 2017, 26, 1732–1743

[9] Pyrhonen V.-P, Koivisto H.J., Vilkko M.K., A reduced-order two-degree-of-freedom composite nonlinear feedback control for a rotary DC servo motor, Proceedings of the 56th Annual IEEE Conference on Decision and Control (12 December – 15 December 2017, Melbourne, Australia), IEEE, Melbourne, Australia, 2017, 2065–2071

[10] Hu C., Wang R., Yan F., Chen N., Robust composite nonlinear feedback path-following control for underactuated surface vessels with desired-heading amendment, IEEE Trans. Ind. Electron., 2016, 63, 6386–6394

[11] Lan W., Thum C.K., Chen B.M., A hard disk drive servo system design using composite nonlinear feedback control with optimal nonlinear gain tuning methods, IEEE Trans. on Ind. Electron., 2010, 57, 1735–1745

[12] Cheng G., Peng K., Robust composite nonlinear feedback control with application to a servo positioning system, IEEE Trans. Ind. Electron., 2007, 54, 1132–1140
[13] Wendong P., Jianbo S., Tracking controller for robot manipulators via composite nonlinear feedback law, IEEE J. Syst. Eng. Electron., 2012, 20, 129–135
[14] Cai G., Chen B.M., Peng K., Dong M., Lee T.H., Modeling and control of the yaw channel of a uav helicopter, IEEE Trans. Ind. Electron., 2008, 55, 3426–3434
[15] Cheng G., Hu J–G., An observer-based mode switching control scheme for improved position regulation in servomotors, IEEE Trans. Contr. Syst. Technol., 2013, 22, 1883–1891
[16] Lan W., Chen B.M., On selection of nonlinear gain in composite nonlinear feedback control for a class of linear systems, Proceedings of the 46th Annual IEEE Conference on Decision and Control, (12 December – 14 December, 2007, New Orleans, LA, USA), IEEE, New Orleans, LA, USA, 2007, 1198–1203
[17] Apkarian J., Levis M., Martin P., Qube-servo 2 experiment for Matlab/Simulink users – instructor workbook, Quanser Inc. 2016, Ch. PD Control.

Tero Hietanen, Timo Heikkinen, Manne Tervaskanto ja Satu Vähänikkilä OAMK

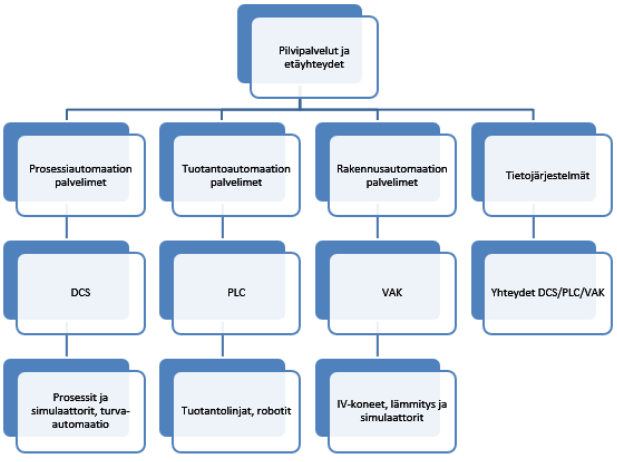
Verkottunut yhteistyö automaatiokoulutuksessa

Oululaiset automaatiotekniikan koulutusorganisaatiot kehittävät alueelle uutta keskitettyä digitaalista oppimisympäristöä (DigiAuto-hanke) ja uudenlaista toimintamallia automaatiokoulutukseen (EduAuto-hanke). Mukana ovat kaikki alueen ammatillisen koulutuksen kouluttajat (Oulun ammattikorkeakoulu Oamk, Oulun yliopisto OY, Oulun seudun ammattiopisto OSAO) ja tämä muodostaa uudenlaisen eri kouluttajien välisen toimintaympäristön. Tässä artikkelissa kuvataan opetusympäristöjen kehitystyötä sekä pilotoituja toteutuksia. Pilotoidut toteutukset ovat vähintään kahden eri oppilaitoksen välisiä yhteisprojekteja. Lisäksi tuodaan esille laitteistojen ylläpidon toteutusta sekä yhteisen varausjärjestelmän kehittämistä.

Asiasanat: IoT, koulutus, digitalisaatio

1 Johdanto

Hankkeissa on investoitu uuteen teknologiaan ja uusien yhteistyömuotojen kehittämiseen. Hankittua teknologiaa käytetään prosessiautomaation, tuotantoautomaation, rakennusautomaation sekä teollisuuden tietojärjestelmien kouluttamiseen, kuvan 1 mukaisesti. Yhteisinä tekijöinä ovat IoT, digitalisaatio sekä etäkäytettävyys.



Kuva 1. DigiAuto-hankkeen hankintalinjat.

Prosessiautomaation koulutukseen on hankittu mm. virtualisoidut automaatiojärjestelmien suunnittelu ja simulointityökalut Valmetilta. Tämä pitää sisällään mm. historiatiedon tallennusjärjestelmän moderneine datan analysointi- ja visualisointityökaluineen. Lisäksi järjestelmään kuuluu teollisen mittakaavan mallinprediktiivinen säätöjärjestelmä, jolla ohjataan mm. Oamkin pilot-prosessia. Säätötekniikan koulutukseen on hankittu kannettavat miniprosessit, joihin on liitetty prosessiasema sekä tarvittava kenttä-I/O. Oamkin pilot-prosessin kenttälaitteet on modernisoitu sekä liitetty HART- ja Profinet-väyliin.

Tuotantoautomaation tärkeimmät hankinnat ovat Mitsubishin robottijärjestelmä sekä Feston MPS-tuotantolinjasto. Feston järjestelmässä tuotantotieto kulkee RFID-tekniikalla erilliseltä MES-järjestelmältä ohjaavalle logiikalle. Lisäksi laitteiston käynnissäpidossa hyödynnetään AR-teknologiaa (Augmented Reality). Keskeisimpänä rakennusautomaation hankintana on laivakonttiin rakennettu energia- ja LVI-tekniikkaa sisältävä hybridijärjestelmä. Laitteisto sisältää mm. aurinkopaneelit ja -keräimet, tuuliturbiinin, ilmalämpöpumpun, akuston sekä lämminvesivaraajan. Laitteistoa ohjataan Fidelixin automaatiojärjestelmällä. Laitteiston etäohjauksessa hyödynnetään Tosibox-tekniikkaa.

Yhteisten resurssien ylläpidon ja varausten sekä koulutusmateriaalin yhteiskäytön mahdollistavan toiminnanohjausjärjestelmän on toteuttanut ALMA Consulting Oy. Hankkeessa toteutettava oppilaitosten yhteinen automaatioympäristö laajentaa jo olemassa olevaa koulutusyhteistyötä sekä tarjoaa monipuolisen ja laaja-alaisen ympäristön erilaisille automaatioalan tutkimus- ja yrityshankkeille.

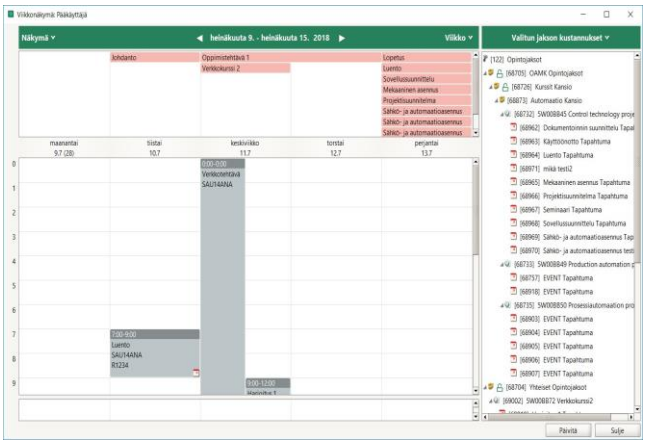
2 Toiminnanohjausjärjestelmän kehitys

Uusien hajautettujen koulutusresurssien sekä yhteisen koulutuksen toteuttaminen vaatii myös yhteisen toiminnanohjauksen kehittämistä. Hankkeessa automaatioteknologia sekä tilojen ja oppilasryhmien hallinta on viety ALMA-toiminnanohjausjärjestelmään.

ALMA® on ollut markkinoilla vuodesta 1986 ja tuotemerkki on rekisteröity 13 maassa. Sen takana on ALMA Consulting Oy, joka kehittää, markkinoi, toimittaa ja ylläpitää ALMA® suunnittelu-, teknisen tiedon ja tapahtumien sekä kunnossapidon hallintajärjestelmää sekä siihen liittyviä palveluita. /1/

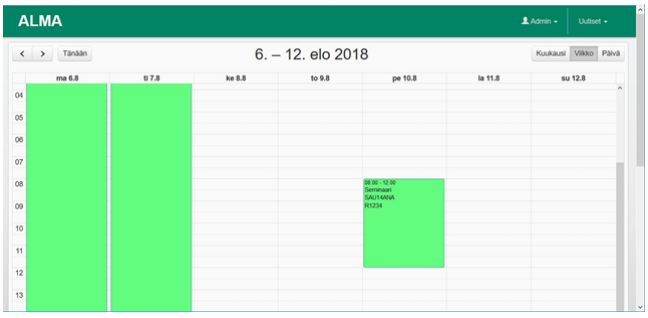
EduALMA-koulutuksen toiminnanohjausjärjestelmää voidaan hyödyntää mm. opetuksen suunnittelussa, oppimateriaalin hallinnassa, oppimisympäristöjen ja tilojen varauksissa, oppilasryhmien hallinnassa, automaatiolaitteistojen ylläpidossa sekä opintojaksojen toteutuksessa.

Kuvassa 2 on esimerkki opiskelijan viikko-kalenterinäkymästä. Viikoittainen opetus on koottu opintojaksojen tapahtumista.

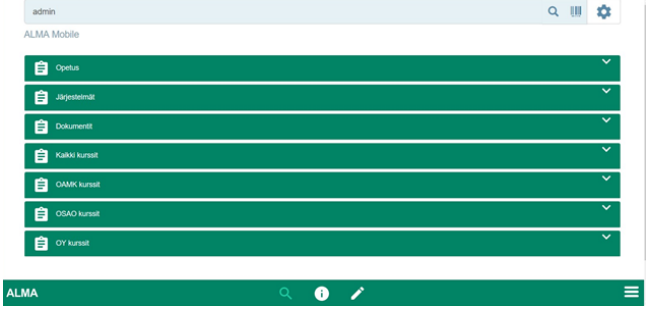


Kuva 2. ”Lukujärjestys” 7-päivää kalenterissa.

EduALMAn näkymät web- ja mobiilikäyttöliittymissä on esitetty kuvissa 3 ja 4. Opiskelijalla ja muilla käyttäjillä on erinomainen pääsy esim. oppimateriaaleihin, lukujärjestyksiin tai oppimisympäristöihin, kuten esim. automaatiosuunnittelun ohjelmistopalvelimelle kirjautumiseen.



Kuva 3. Web-ALMA kalenterinäkymän periaatekuva.

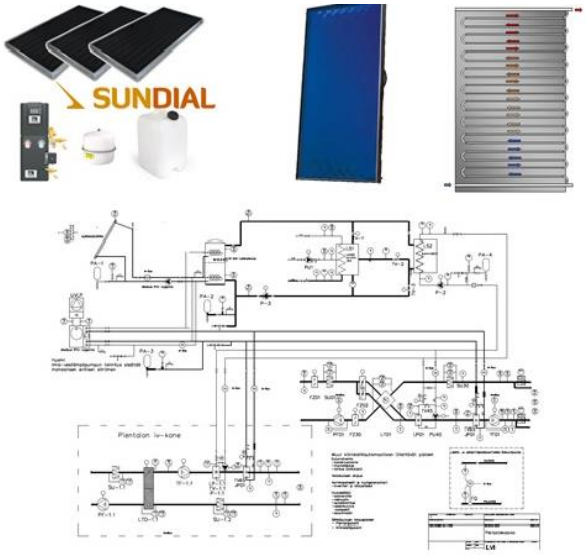


Kuva 4. Mobile-ALMA päänäkymä.

ALMAa käytetään lisäksi automaatio- ja sähköteknisessä dokumentoinnin hallinnassa sekä kunnossapitojärjestelmien (CMMS) opetuksessa.

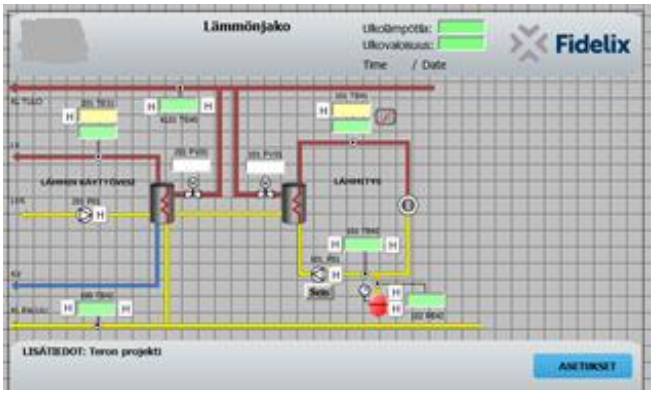
3 Rakennusautomaation oppimisympäristöt

Kuvassa 5 on esitetty hybriditekniikkaan perustuva kiinteistöautomaatiojärjestelmä. Lämpö tuotetaan aurinkokeräimillä, ilma-vesilämpöpumpulla sekä sähkövastuksella. Aurinkokeräimillä ja ilmavesilämpöpumpulla tuotettu lämpö ohjataan varaajaan. Varaajassa on lisäksi sähkövastus, joka tarvittaessa tuottaa lisälämpöä varaajaan. Varaajalta lämpö siirretään lämmönjakopaketille, joka lämmittää lämpimän käyttöveden, sekä lämmitykseen tarvittavan veden.



Kuva 5. Rakennusautomaation oppimisympäristön teknologiaa.

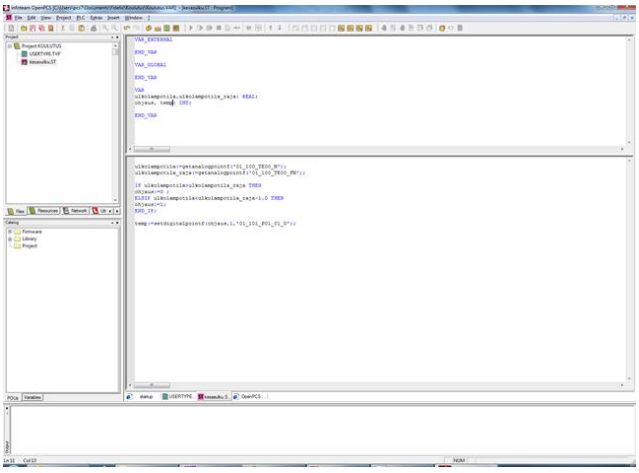
Rakennusautomaation oppimisympäristöä hallitaan Fidelixin-automaatiojärjestelmällä. Lisäksi Fidelix on toimittanut taloteknisten järjestelmien simulointiympäristön. Laitteistoon liittyen opiskelijat ovat tehneet mm. rakennusautomaatioprojekteja, joissa automaatiosuunnittelua on harjoiteltu mm. FX-editor työkalulla, kuva 6.



Kuva 6. Grafiikka suunnittelunäkymä FX-editor työkalussa.

Vaativampia automaatiotoimintoja on ohjelmoitu OpenPCS-ohjelmistolla. OpenPCS on IEC 61131-3-standardin pohjalta

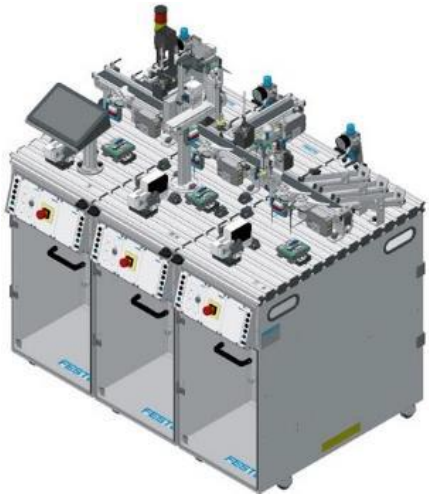
luotu ohjelmointityökalu. IEC 61131-3- standardin mukainen tekstipohjainen ohjelmointikieli Structured Text on hyvin samankaltainen kuin yleisimmät tekstipohjaiset ohjelmointikielet, kuten C++. Kuvassa 76 on esitetty esimerkkiohjelma lämmitysverkoston kesäsulusta. /2/



Kuva 7. Lämmitysverkoston kesäsulku OpenPCS-ohjelmistossa.

4 Tuotanto-automaation oppimisympäristöt

Tuotantoautomaation liitettyjä laitteistoja on käytetty yhteistyössä OSAO ammattiopiston ja OAMK välillä. Tähän opetuskokonaisuuteen sisältyy mm. Feston MPS tuotantolinja, jonka ohjauslogiikkaa voidaan konfiguroida Siemens S7-1500 logiikalla, kuva 8.

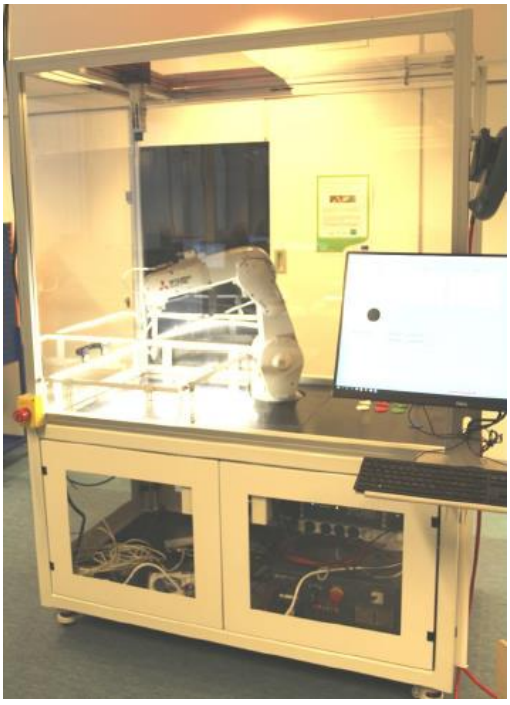


Kuva 8. Festo MPS tuotantolinjasto.

Linjasto kattaa myös erillisen MES-palvelimen, jolla voidaan toteuttaa tuotannonohjausjärjestelmä linjastolle RFID-tekniikan avulla. Mielenkiintoinen lisäpiirre on lisätyn todellisuuden (AR – Augmented Reality) hyödyntäminen linjaston käyttöönotossa ja operoinnissa (kuva 9).

Kuva 9. Festo MPS laitteiston AR –sovellus.

Syksyllä 2018 aloitettu ja kevään 2019 jatkuva yhteistyöprojekti tuotantoautomaatioon ja robotiikkaan liittyen OSAO:n ja OAMK:n välillä. Robotiikan opiskelu on aloitettu opettelemalla ohjelmointia RT ToolBox -ohjelmiston avulla ja tekemällä tällä ohjelmalla simulointeja. Tästä on siirrytty ohjelmien testaamiseen robottien avulla. Kevään aikana tarkoituksena on tehdä OSAO:n ja OAMK:n opiskelijoiden yhteistyöprojekteja robotiikkaan ja tuotantoautomaation liittyen. Kuvassa 10 DigiAuto-projektin puitteissa hankittu robotti.



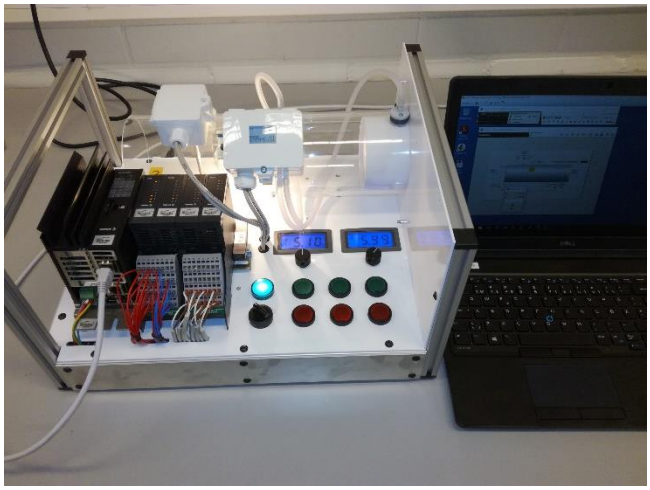
Kuva 10. DigiAuto-hankkeen robottimoduuli.

5 Prosessiautomaation oppimisympäristöt

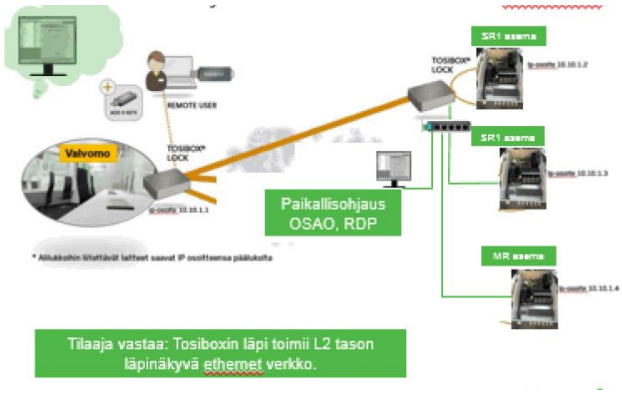
Uudistetut laitteisto- ja ohjelmistoratkaisut ovat tarkoittaneet myös projektityyppisten kurssien sisältöjen uudistamista ja kehittämistä. Projekteista on osa toteutettu jo kuluneen lukuvuoden aikana oppilaitosten välisinä yhteistyöprojekteina.

Näihin voidaan lukea mm. Automaatiotekniikan erikoistyö – kurssi, jossa OSAO:lla veden kierrätys- ja lämmitysprosessiin on liitetty uusi Valmet SR1-prosessiasema. Asemalle on etäyhteyden päässä olevalla EAS suunnitteluasemalla tehty sovellus- ja käyttöliittymäsuunnittelu sekä sovellustestaus (kuva YY). Ammattikorkeakoulun opiskelijat ovat tehneet prosessiin piirikaaviot Loop Circuit CAD:lla, joiden mukaisesti ammattiopiston opiskelijat ovat kytkeneet mittaus- ja toimilaitteet kiinni prosessiin sekä testanneet kytkentöjen toimivuuden. Tämän jälkeen ammattikorkeakoulun opiskelijoiden tehtävänä on ollut sovellussuunnittelun teko ja järjestelmän käyttöönotto sekä toiminnan raportointi. Projekti on vaatinut eri opiskelijaryhmien tiivistä yhteistyötä.

Yhteiskäyttöön on lisäksi hankittu Valmet SR1-prosessiasemia, joihin on kiinteästi integroitu kenttälaitemittauksia ja toimilaite. Kannettava järjestelmä sisältää 4 IO-korttia (DI, DO, AI ja AO), joilla on myös Hart-tuki (kuva 11). Sovellussuunnittelu- ja operointiasemat ovat erillisellä kannettavalla tietokoneella. Laitteistolla voidaan harjoitella ulkoisia kenttälaitekytkentöjä ja sovellus- ja näyttösuunnittelua sekä mm. PID-säätöjen viritystä. Kannettavuus mahdollistaa laitteiston helpon yhteiskäyttöisyyden eri oppilaitosten kesken.



Kuva 11. Kannettava Valmet SR1-logiikka ja sovellussuunnitteluun ja operointiin tarkoitettu työasema.



Kuva 12. Etäyhteys kahden oppilaitoksen välillä.

Teollisen mallinprediktiivisen säätimen (MPC) testausta ja simulointia on tehty yhdessä sekä ammattikorkeakoulun että yliopiston opiskelijoiden toimesta. AMK:lla on keskitytty historiatiedonkeruun konfigurointiin ja datan luotettavaan tiedonvälitykseen MPC-palvelimen ja automaatiojärjestelmän välillä. Yliopistolla suurempi painoarvo on ollut MPC:n konfiguroinnissa, monimuuttujasäätöjen simuloinnissa ja järjestelmätestauksessa. MPC-palvelimelle on pääsy etäyhteyden kautta 20 käyttäjälle molemmista oppilaitoksista, joten myös yhtäaikainen työskentely on ollut mahdollista.

6 Yhteenveto

Laitteistohankinnat mahdollistanut DigiAuto-hanke päättyi vuoden 2018 loppuun. Rakennusautomaation laitteistokokonaisuuden rakentamisessa, käyttöönotossa ja koulutuksissa oli pientä viivettä. Hankintakokonaisuudessa oman pääosin oppilasvoimin tehdyn suunnittelu- ja asennustyön osuus oli suurin. Kehitystyötä jatketaan vielä mm. etäyhteyksien käyttöönoton ja koulutusten osalta. Vastaavasti prosessiautomaation hankintalinjan osalta on eniten yhteisiä kokemuksia projekteista. Opiskelijat ovat ottaneet yhteistyön vastaan myönteisesti ja osallistuminen on ollut aktiivista. Yhteishankkeet antavat hyvän valmennuksen eri tehtäviin liittyvistä työelämärooleista.

Robotiikan osalta yhteisprojektit ovat myös olleet suosittuja. Feston laitteiston laajamittainen hyödyntäminen edellyttää vielä lisäkoulutusta ja ohjelmistopäivityksiä. Olemme saaneet ensikokemusta MES-järjestelmstä ja kunnossapidon AR-sovelluksesta. ALMA-järjestelmän osalta on hankittu ensi kokemuksia toiminnanohjauksesta ja resurssien yhteisistä kuvauksista. EduAuto-hanke päättyy vuoden 2019 loppuun mennessä, joten hankkeen puitteissa järjestelmää voidaan pilotoida syksyn 2019 opetuksessa.

Oppilaitosten välisessä projektitoteutuksissa onkin tärkeää, että tietyt tekniset ja työtapoihin asiat hallitaan jo etukäteen eikä aivan kaikkea asennuksesta sovellussuunnitteluun tarvitse opetella projektin aikana. Muussa tapauksessa on riskinä, että projekti venyy eikä yhteistyö tarjoa sitä mitä siltä on parhaimmillaan saavutettavissa. Toteutus vaatii myös jokaiselta organisaatiolta sitoutumista hankkeeseen ja huolellista etukäteissuunnittelua.

Laitteistojen, ohjelmistojen ja yhteistoiminnan osalta on vielä paljon opeteltavaa. Uuden teknologian käyttöönotto tuo uusia haasteita esim. palvelinten ja päätelaitteiden päivitysten suhteen. Etäyhteyksien ylläpito ja laitteistojen etäohjaus vaativat vielä kehittämistä mm. turvallisuus- ja valvontateknologian osalta. Kaikissa oppilaitoksissa on toteutettu hankkeen aikana säästöjä ja organisaatiomuutoksia, näistä huolimatta hanke on mahdollistanut uuden toimintakulttuurin kehittämisen.

Lähdeluettelo

1. https://www.alma.fi/ 4.2.2019
2. https://www.theseus.fi/bitstream/handle/10024/68484/Korhonen_Vili.pdf?sequence=1 3.2.2019

Tomi Räsänen* and Veli-Pekka Pyrhönen

State feedback control of a rotary inverted pendulum

Abstract: In this paper, we design a state feedback controller for a rotary inverted pendulum, which is mounted to a Quanser QUBE-Servo 2 unit. To be more specific, we use linear quadratic regulator to find suitable controller gains for QUBE-Servo 2 system. The essential characteristics of the QUBE-Servo 2 unit are presented and the performances of the closed-loop systems are evaluated based on rise time, settling time and overshoot of the rotary arm's step response. The design is validated using simulations and real-time experiments. The resulting controller stabilizes the rotary pendulum to upright position and is able to move the pendulum to desired angular position while keeping the balance under control.

Keywords: LQR-control, pole-placement, inverted pendulum, balance control, reference tracking

***Corresponding Author: Tomi Räsänen:** Tampere University (Bachelor's student), E-mail: tomi.rasanen@tuni.fi
Veli-Pekka Pyrhönen: Tampere University, E-mail: veli-pekka.pyrhonen@tuni.fi

1 Introduction

Regulating a link to its upright position is an application of a mechanical balancing problem. Combined with a reference tracking, it becomes an exciting problem for testing various control designs. State feedback is a widely used method for designing feedback controllers for linear-time-invariant (LTI) systems. LTI-design methods take advantage of the well-known theory of linear algebra to form simple controllers. A classical method is pole placement which, in theory, results in closed-loop system with arbitrary dynamics [1]. In practice, there are always e.g. physical restrictions that limit the achievable performance of the closed-loop control system such as parasitic effects and actuator saturation. Optimization can also be used in line with LTI-design algorithms to yield an optimal solution to the problem with given parameters. A well-known optimal control method is the linear quadratic regulator (LQR) which is one of the most important results in modern control theory [2].

Inverted pendulum systems have been studied extensively in recent years. For example, stabilization of a real inverted pendulum was studied in [3], controlling a wheeled inverted pendulum in [4] and characteristics of the control of a flying inverted pendulum in [5]. Related topics to the inverted pendulums were also studied in [6]–[8]. In this paper, we design state feedback controllers for a rotary inverted pendulum using the pole-placement and LQR design methods. The rotary inverted pendulum is attached to a Quanser QUBE-Servo 2 unit, which is a small-scale design platform for a variety of control methods. The characteristics of the system is captured by single-input–multiple-output (SIMO) model. The feedback controllers are designed to regulate the pendulum link to the upright position with carefully coordinated movements of the DC motor. Feedback control also enables the rotary arm to turn from one angle to another. The performance of the designed controllers are evaluated using the well-known classical measures such as rise time, settling time and overshoot.

The material of this paper is organized in the following order. Section 2 presents the pole-placement method and LQR-control, whereas Section 3 examines the Smith-McMillan form of the system. In Section 4, we introduce characteristics of the QUBE-Servo 2 unit and the model of the inverted pendulum. In Section 5, feedback controllers are designed and the resulting closed-loop control systems are tested using simulations as well as real-time experiments with QUBE-Servo 2 system. Finally, some concluding remarks are summarized into Section 6.

2 State feedback

In this section, a short introduction to state-space representation and state feedback law are discussed. In addition, theory of the pole-placement method and LQR-control are presented.

2.1 State-space representation

The system considered in this paper is SIMO of the following form

$$\begin{cases} \dot{\mathbf{x}}(t) &= A\mathbf{x}(t) + \mathbf{b}\mathrm{u}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t), \end{cases}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{1}$$

where $A \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ is the system matrix, $\mathbf{x} \in \mathbb{R}^{\mathrm{n}}$ is the state vector, $\mathrm{u} \in \mathbb{R}$ is control input, and vector $\mathbf{b} \in \mathbb{R}^{\mathrm{n}\times 1}$. Output matrix $C \in \mathbb{R}^{\mathrm{l}\times\mathrm{n}}$ describes the relationship between state vector and output vector $\mathbf{y} \in \mathbb{R}^{\mathrm{l}\times 1}$, which includes the angle measurements from the system. The system (1) is said to be realization $(A, \mathbf{b}, C)$.

In this paper, the system (1) is assumed to be controllable and observable. To be more specific, the rank of the controllability matrix

$$\mathrm{rank}C_k = \mathrm{rank}[\mathbf{b} \;\; A\mathbf{b} \;\; A^2\mathbf{b} \;\; \ldots \;\; A^{n-1}\mathbf{b}] = n \tag{2}$$

and the rank of the observability matrix

$$\mathrm{rank}O = \mathrm{rank}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = n, \tag{3}$$

where n is the number of states of the system. When examining the SIMO-models, realization which is both controllable and observable is said to be minimal.

2.2 State feedback law

In this paper, state feedback control laws are written by [1]:

$$\mathrm{u}(t) = K\mathbf{x}(t), \tag{4}$$

in which $\mathrm{u}(t)$ and $\mathbf{x}(t)$ are respectively the same input and state vector as introduced in (1) and $K \in \mathbb{R}^{1\times\mathrm{n}}$ is called a full-state feedback gain. The control law in (4) assumes that the full state vector is known without any errors. If there are unknown states, it is mandatory to use a state estimator to calculate the missing values. When missing measurements represents rates of changes of state variables, then simple high-pass filters can be used to supplement the state such that state feedback control can be applied. In this paper, the high-pass filters are of the following form

$$\frac{\omega_f s}{s + \omega_f}, \tag{5}$$

where ω_f is the cut-off frequency of the filter and s is the Laplace variable. The filter (5) produces an estimate for the rate of change of its input signal.

The control law defined by the equation (4) is valid only with a regulation task. If the control scenario also includes a reference state vector $\mathbf{x}_r$, the closed-loop system consisting of (1) and (4) with the coordinate transformation $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_r$ yields [1]:

$$\dot{\tilde{\mathbf{x}}} = \dot{\mathbf{x}} = (A - \mathbf{b}K)\tilde{\mathbf{x}} + A\mathbf{x}_r. \tag{6}$$

Therefore, there will always be steady-state error, unless $A\mathbf{x}_r = 0$. The condition $A\mathbf{x}_r = 0$ is met when the system matrix A is singular and the vector $\mathbf{x}_r$ belongs into the null space of this matrix, which is the situation in this paper.

2.3 Pole-placement method

Pole-placement is popular way to design control for closed-loop LTI-systems. Specifying desired locations of poles, a designer can ensure stability of the closed-loop and meet assigned control specifications. In theory, the closed-loop poles can be placed to arbitrary locations on the left-half complex plane if the system is controllable [2, 9]. Desirable places of the poles are achieved by choosing the correct coefficients to the gain matrix K.

There are a couple of different ways to calculate the correct value of the state gain matrix based on the given locations of poles. In this paper, feedback gain is calculated according to Ackermann's formula. The formula states that the state gain matrix K can be obtained by solving the equation [9, 10]

$$K = [0\ 0\ \ldots\ 0\ 1]C_k^{-1}\alpha(A) \tag{7}$$

in which C_k is the controllability matrix from (2) and $\alpha(A)$ is defined by

$$\alpha(A) = A^n + \alpha_1 A^{n-1} + \ldots + \alpha_{n-1}A + \alpha_n I, \tag{8}$$

where $I^{(n\times n)}$ corresponds to $(n \times n)$ identity matrix and coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ can be obtained from the desired characteristic polynomial

$$s^n + \alpha_1 s^{n-1} + \ldots + \alpha_{n-1}s + \alpha_n. \tag{9}$$

Ackermann's formula is feasible, because the system in this paper is well-conditioned and is of low-order.

2.4 Linear Quadratic Regulator

Linear quadratic regulator uses quadratic cost function to calculate the values of the state gain matrix K,

when the system is controllable [2]. In order to compare the results of optimizations, one has to declare a cost function J_k. An infinite horizon quadratic cost function, which is suitable for SIMO systems is

$$J_k = \int_0^\infty \mathbf{x}^\intercal Q\mathbf{x} + \mathrm{u}^2 R \; dt. \tag{10}$$

The vector $\mathbf{x}$ and the scalar u are from the state equation (1), whereas $Q \geq 0 \in \mathbb{R}^{n\times n}$ and $R > 0 \in \mathbb{R}$ are called weighting matrices [2].

An optimal state feedback gain, which minimizes the cost function (10) is [1]:

$$K = R^{-1}\mathbf{b}^\intercal P, \tag{11}$$

where $P > 0$ satisfies the algebraic Riccati equation [1]:

$$A^\intercal P + PA - P\mathbf{b}R^{-1}\mathbf{b}^\intercal P + Q = 0. \tag{12}$$

The Ricatti equation (12) is non-linear and can be solved e.g., using eigenvalues of the Hamiltonian matrix [1].

One can seek suitable weights by examining the physical characteristics of the system. Another approach for selecting the weights is trial and error, where initial selection (guess) is based on prior knowledge of the similar problems. The above-mentioned method of trial and error is often used at least in the final selection of weighting matrices. [1]

3 Transfer function matrix and Smith-McMillan form

A system's poles and, in the case of minimal realization of SIMO-system, transmission zeros can be found using Rosenbrock system matrix. Despite the usefulness of the Rosenbrock matrix, it could leave unrevealed information about the zeros of the system's transfer functions, which could cause nonminimal time domain behaviour. For this reason, a transfer function matrix is introduced. From the transfer function matrix, zeros and poles of the system is possible to determine using a system's Smith-McMillan form.

The transfer function matrix $\mathrm{G}(s) \in \mathbb{C}^{l\times 1}$ of a SIMO-system is [11]

$$\mathrm{G}(s) = C(sI - A)^{-1}\mathbf{b} \tag{13}$$

where matrices A, C and vector $\mathbf{b}$ are from equation (1). Poles and zeros of the transfer function from the input u to l^{th} output can be found in the nominator and in the denominator of the corresponding element in G(s).

The theorem of the Smith-McMillan form states that a rational matrix G(s) of normal rank n could be transformed into a pseudo-diagonal rational matrix [12]

$$\mathrm{M(s)} = diag\Big\{\frac{\varepsilon_1(s)}{\psi_1(s)}, \frac{\varepsilon_2(s)}{\psi_2(s)} ... \frac{\varepsilon_r(s)}{\psi_r(s),} 0, ...0\Big\} \tag{14}$$

in which polynomials $\frac{\varepsilon_i(s)}{\psi_i(s)}$ $(i = 1, 2, ..., r-1)$ have no common factors and

$$\begin{cases} \varepsilon_i(s)/\varepsilon_{i+1}(s) \\ \psi_{i+1}(s)/\psi_i(s) \end{cases} \tag{15}$$

are satisfied without remainder. The diagonal matrix M(s) is called the Smith-McMillan form of the transfer function matrix G(s), and in the case of SIMO-system, it reduces to

$$\mathrm{M(s)} = diag\Big\{\frac{\varepsilon_1(s)}{\psi_1(s)}\Big\} = \begin{bmatrix} \frac{\varepsilon_1(s)}{\psi_1(s)} \\ 0 \end{bmatrix}. \tag{16}$$

3.1 System's poles and zeros from the Smith-McMillan form

Poles of the system can be found from the roots of the pole polynomial of the constructed Smith-McMillan form. The pole polynomial is defined by denominators of M(s) as [12]

$$p(s) = \psi_1(s)...\psi_r(s), \tag{17}$$

which in SIMO-case only contains the first element $\psi_1(s)$. If the realization is minimal, the pole polynomial is given by $\det(sI - A)$ and the poles are the eigenvalues of the system matrix A.

In general, when the system is not in the form of SISO, there is defined different kind of zeros, called invariant zeros. In the case of minimal realization, the invariant zeros are the same as transmission zeros. Transmission zeros are found from the roots of the zero polynomial of Smith-McMillan form, where zero polynomial is [12]

$$z(s) = \varepsilon_1(s)...\varepsilon_r(s). \tag{18}$$

In SIMO-case, the zero polynomial is reduced to contain only the numerator $\varepsilon_1(s)$ of the fist element of M(s).

4 Quanser QUBE-Servo 2 system

Quanser QUBE-Servo 2 system is a rotary DC motor application, which can be used as a testbed for

reference tracking and regulation tasks. It consist of an Allied Motion CL40 Series 18V brushed DC motor (model 16705) and a PWM (Pulse-Width Modulation) voltage-controlled power amplifier, which is used to power the motor. The PWM accepts commands from Data Acquisition (DAQ) device. The DAQ is linked to a PC via USB connection. Quanser also provides Simulink/Matlab add on (QUARC), which includes necessary blocks to use the unit with Simulink software.

The rotary inverted pendulum is attached to the DC motor via magnets. Angular positions of the DC motor and the pendulum are measured with a single-ended optical shaft encoder (US Digital E8P-512-118) with the resolution of 2048 counts per revolution, which transforms to 0.176 degree of accuracy. The sampling rate of the angle measurements is 0.001 second. Rotary pendulum attached to Quanser QUBE-Servo 2 is depicted in Fig. 1. A more detailed description of the specifications of the unit can be found from the datasheet of the manufacturer.

Fig. 1. Quanser QUBE-Servo 2 system with the rotary pendulum

The equations of motion (EOM) of the rotary pendulum system is obtained using Euler-Lagrange method and defined by two non-linear differential equations:

$$(m_pL_r^2 + \frac{1}{4}m_pL_p^2 - \frac{1}{4}m_pL_p^2cos(\alpha)^2 + J_r)\ddot{\theta} - (\frac{1}{2}m_pL_pL_rcos(\alpha))\ddot{\alpha} + (\frac{1}{2}m_pL_p^2sin(\alpha)cos(\alpha))\dot{\theta}\dot{\alpha} + (\frac{1}{2}m_pL_pL_rsin(\alpha))\alpha^2 = \tau - D_r\dot{\theta} \quad (19)$$

$$\frac{1}{2}m_pL_pL_rcos(\alpha)\ddot{\theta} + (J_p + \frac{1}{4}m_pL_p^2)\ddot{\alpha} - \frac{1}{4}m_pL_p^2cos(\alpha)sin(\alpha)\dot{\theta}^2 + \frac{1}{2}m_pL_pg\sin(\alpha) = -D_p\dot{\alpha}, \quad (20)$$

and

$$\tau = \frac{k_m(V_m - k_m\dot{\theta})}{R_m}, \quad (21)$$

where θ is a rotary arm angle, $\dot{\theta}$ is the angular speed of the rotary arm angle, $\ddot{\theta}$ is the angular acceleration of the rotary arm, α is the inverted pendulum angle, $\dot{\alpha}$ is the angular velocity of the inverted pendulum angle and $\ddot{\alpha}$ is angular acceleration. The input voltage to the system is V_m. The remaining constants are the arms mass (m_r, m_p), the arms length (L_r, L_p), the equivalent viscous damping coefficients (D_r, D_p), the moments of inertia about pivot (J_r, J_p), the terminal resistance of the DC motor R_m, torque constant k_t and back-emf constant k_m. The subscripts 'r' and 'p' refer to the rotary arm and the inverted pendulum, respectively. The value of the inverted pendulum angle is zero when the pendulum is in the downward position. Numerical values of the parameters are listed in Table 1.

Table 1. Quanser QUBE-Servo 2 parameters

Symbol	Description	Value
	Rotary arm	
m_r	Rotary arm mass	0.095 kg
L_r	Rotary arm length	0.085 m
D_r	Equivalent Viscous Damping Coefficient	0.0015 $\frac{\text{Nms}}{\text{rad}}$
J_r	Moment of inertia about the center of mass	5.7 $\times 10^{-5}$ kgm^2
	Pendulum link	
m_p	Pendulum link mass	0.024 kg
L_p	Pendulum link lenght	0.13 m
D_p	Equivalent Viscous Damping Coefficient	0.0005 $\frac{\text{Nms}}{\text{rad}}$
J_p	Moment of inertia about the center of mass	3.4 $\times 10^{-5}\text{kgm}^2$
	DC Motor	
R_m	Terminal resistance	8.4 Ω
k_t	Torque constant	0.042 $\frac{\text{Nm}}{\text{A}}$
k_m	Back-emf constant	0.042 $\frac{\text{Vs}}{\text{rad}}$

The moments of inertia J_p and J_r in Table 1 are calculated by

$$J_p = \frac{1}{12}m_pL_p^2 \quad \text{and} \quad J_r = \frac{1}{12}m_rL_r^2. \quad (22)$$

The kinematics of the inverted pendulum system is sketched in Fig. 2. In Fig. 2, Cartesian axes are labeled

as x_0, y_0 and z_0, θ is the rotary arm angle, α is the pendulum link angle, L_r is the rotary arm length and L_p is the pendulum link length.

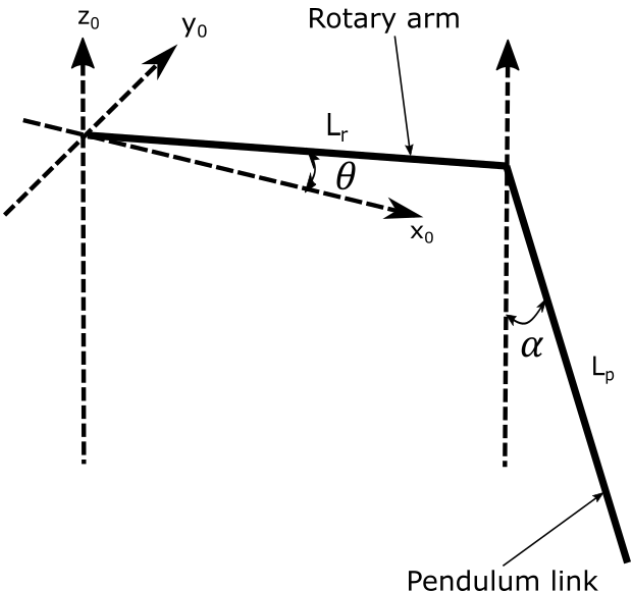


Fig. 2. Kinematics of the inverted pendulum

Equations (19) and (20) must be linearized for LTI-control algorithms introduced previously in this paper. Linearization is performed at the point where the pendulum link angle $\alpha = 180^\circ = \pi$ rad, and the rotary arm angle $\theta = 0^\circ$. The linearized EOM are

$$\begin{cases} (m_p L_r^2 + J_r)\ddot{\theta} + (\frac{1}{2} m_p L_p L_r)\ddot{\alpha} & = \tau - D_r\dot{\theta} \\ (-\frac{1}{2} m_p L_p L_r)\ddot{\theta} + (J_p + \frac{1}{4} m_p L_p^2)\ddot{\alpha} & \\ +(-\frac{1}{2} m_p L_p g)\alpha & = -D_p\dot{\alpha}. \end{cases} \tag{23}$$

Using the values from Table 1 and substituting equation (21) to equation (23), the state-space model of the QUBE-Servo 2 system can be written as

$$\begin{bmatrix} \dot{\theta} \\ \dot{\alpha} \\ \ddot{\theta} \\ \ddot{\alpha} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -41.6 & -4.16 & 1.37 \\ 0 & 72.4 & -4.11 & -2.40 \end{bmatrix} \begin{bmatrix} \theta \\ \alpha \\ \dot{\theta} \\ \dot{\alpha} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 13.9 \\ 13.7 \end{bmatrix} V_m$$
$$\begin{bmatrix} \theta \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \alpha \\ \dot{\theta} \\ \dot{\alpha}. \end{bmatrix} \tag{24}$$

and $\mathbf{x}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^\intercal$. The controllability and observability matrices are both full rank, and hence, the state-space model (24) is the minimal realization of the inverted pendulum system. Thus, the invariant zeros are equivalent to the transmission zeros. The transfer function matrix of the QUBE-Servo 2 system from the state-space model (24) is represented as

$$G(s) = \begin{bmatrix} \frac{13.9s^2+52.1s-1582}{s^4+7.14s^3-55.1s^2-541s} \\ \frac{13.7s-1.79\times10^{-6}}{s^3+7.14s^2-55.1s-541} \end{bmatrix}. \tag{25}$$

Using the algorithm in [12], the Smith-McMillan form of the system is:

$$M(s) = \begin{bmatrix} \frac{1}{s^4+7.14s^3-55.13s^3-541s} \\ 0 \end{bmatrix}, \tag{26}$$

in which the zero polynomial $z(s) = 1$ and the pole polynomial $p(s) = s^4 + 7.14s^3 - 55.1s^2 - 541s$. The above indicates that the rotary pendulum does not have any transmission zeros and the pole locations on the complex plain are $s = 0,\ s = 8.05$ and $s = -7.60 \pm 3.08i$, so the pendulum must be stabilized using feedback control.

5 Feedback controllers

In this paper, the designed controllers must satisfy three requirements:

- The control input V_m has to be between $\pm 15V$.
- The deviation of the pendulum link angle α must be kept with ± 20 degrees from the upright position.
- The rotary arm and pendulum should be moved from the given initial position to final position swiftly without causing overshoot or oscillation.

The first restriction is due to limited performance of the DC motor. The second is involved because of so-called swing-up control that Quanser has implemented for raising the pendulum link to the upright position.

As mentioned in Section 1, the performances of the closed-loop systems are evaluated using rise time (t_r), settling time (t_s) and maximum overshoot (M_p) of the response of the rotary arm. The rise time is defined to be the time in which the response rises from 10% to 90% of its final value. The settling time is the time which the system's response takes to settle 2% from its steady-state value, and the maximum overshoot is the percentage of the maximum value of the response compared to its final value.

Measurements from both angles θ and α are directly provided by QUBE-Servo 2 hardware, whereas the rate of changes of the θ and α are provided by the following high-pass filters

$$\frac{50s}{s+50}. \tag{27}$$

The filters transfer function (27) is provided by Quanser.

First, pole-placement method is used to compose the full-state gain matrix K. The locations of selected poles are

$$s = -5, \quad s = -8.8 \pm 5i \quad \text{and} \quad s = -10. \tag{28}$$

The characteristic polynomial (9) constituted from poles (28) is

$$s^4 + 32.6s^3 + 416.4s^2 + 2416.6s + 5122. \tag{29}$$

Using the system's controllability matrix and the coefficients of the characteristic polynomial (29), Ackermann's formula (7) yields

$$K = \begin{bmatrix} -3.2488 & 45.2021 & -1.9767 & 3.8578 \end{bmatrix}. \tag{30}$$

Competing LQR-design is achieved by selecting the weighting matrices for the cost function (10), and minimizing it by solving the algebraic Ricatti equation (12). After comparing the formed possible gain matrices, the one designed with LQR method performed in the best way i.e. the response yielded the smallest numbers for the chosen performance indicators. The chosen weights were

$$Q = \begin{bmatrix} 15 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix} \quad \text{and} \quad R = 1. \tag{31}$$

The weights in (31) results in the state feedback gain given by

$$K = \begin{bmatrix} -3.8730 & 51.2299 & -2.2650 & 4.3458 \end{bmatrix}. \tag{32}$$

The optimal state gain matrix (32) assigns the system's poles to the locations $s = -5.10, s = -9.88 \pm 4.19i$ and -10.4.

5.1 Simulation results

The simulation model is constructed using the linearized state-space representation (24). Connecting the state gain matrix (32) with the filters (27) to control the inverted pendulum yields the step responses of the pendulum link angle α and the rotary arm angle θ, which are represented in Fig. 3. Corresponding values of the input voltage V_m is in Fig. 4. The size of the step of the rotary arm angle in reference state $\mathbf{x}_r$ is $\frac{\pi}{2}$ rad (90 degrees). There are also Gaussian measurement noise included with mean = 0, variance = 1, sample time = 0.01, and which is multiplied by 0.001.

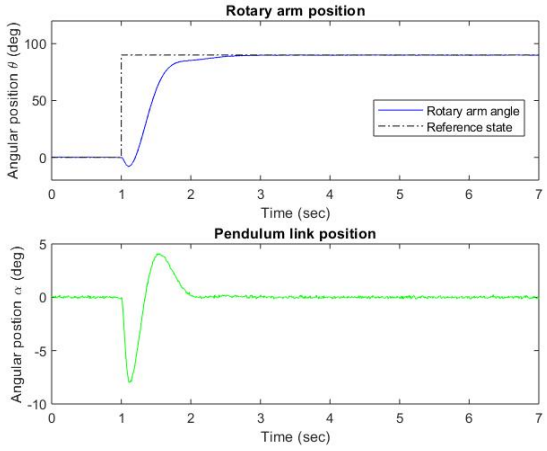


Fig. 3. Reference tracking of the simulated closed-loop system

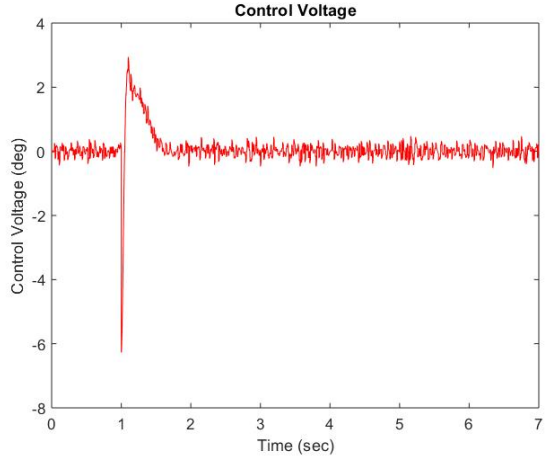


Fig. 4. Control voltage of the simulated LQR controller

According to Fig. 3 and Fig. 4, the control requirements are satisfied. The step response of the rotary arm angle has no overshoot and the settling time is 1.37 seconds. The input voltage (±6.5 V) is well below the given limits, and the maximum deviation of the pendulum angle is approximately 8 degree from the upright position. Therefore, it is possible to use even larger step change of the reference state. As expected from (6), the state error is driven to zero, because $A\mathbf{x_r} = 0$.

The influence of the zeros of the inverted pendulum can be seen from the angle responses. To analyze the effect of zeros, we look back at the system (25). Both transfer function elements in the matrix have nonminimal zero(s), which causes the inverted behavior at the beginning of the step responses. When the DC motor starts turning, the upright positioned pendulum link deviates to the wrong direction due to the moving

rotary arm and the gravitation. To be able to track the given reference state, without tipping over the link, DC motor has to perform the corrective move to be able to maintain its balancing property.

5.2 Implementation results

The Simulink model of the physical QUBE-Servo 2 unit differs from that of the one used in Section 5.1 and is found from the documentation of the pendulum link system by Quanser. Difference between the model used in this paper and the one by Quanser is caused by compensation term due to relative measurement of the angles. In practice, the compensation term causes the activation of the step signal to take place at the same moment as the reference state changes. That produces an offset to the response of the rotary arm angle before the step change. With the compensation term, responses of the angles θ and α are in Fig. 5 and values of the input voltage in Fig. 6. The step change of the reference state of the rotary arm angle is the same magnitude ($\frac{\pi}{2}$) as in the simulations. The zero value of the pendulum link angle is defined to be in the lower position, so values of the y-axis in Fig. 5 are +180 degrees compared to simulation results.

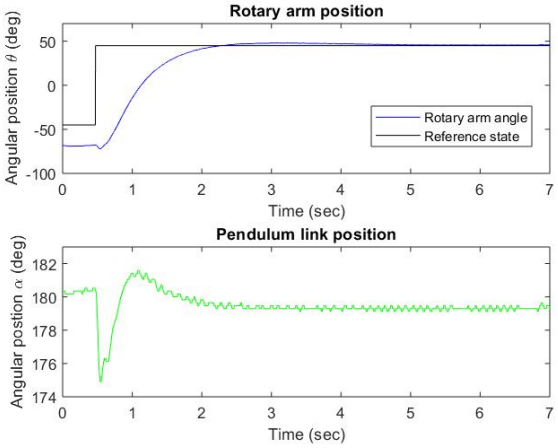


Fig. 5. Experimental result of the angular position of closed-loop system

The performance of the system is similar compared to simulation results in Section 5.1. Both of the given restriction is met with secure margins, and there is practically no steady-state error as discussed in Subsection 2.2. The model of the system does not describe the pendulum system accurately, so the

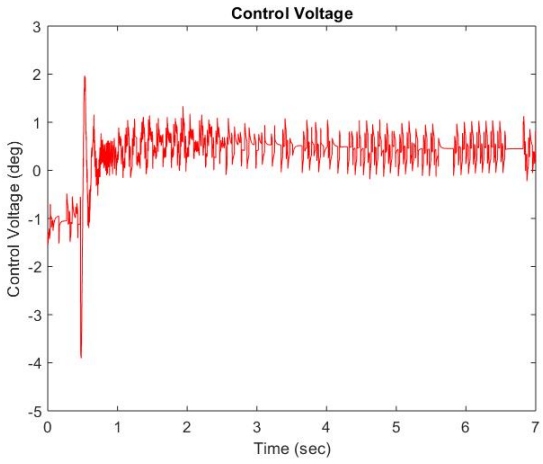


Fig. 6. Experimental result of the control voltage of chosen LQR controller

experimental response of the rotary arm poses a slight overshoot. The resolution of the pendulum link measurement is seen from the constant vibration of the pendulum link position. The vibration is also due to balancing movements of the DC motor, which stabilize the pendulum link to the upright position. The transient response characteristics of the system are

$$M_p = 2.48 \quad t_s = 3.90 \quad \text{and} \quad t_r = 0.966. \qquad (33)$$

The performance could slightly be improved with a parallel-connected integrating controller. The disadvantage of that approach is the selection of the integration gain which could lead, in the worst scenario, to more oscillating responses and even instability of the system. The effect of the minimal-zeros of the system is in line with the simulation result. As mentioned in Section 5, the stability of the system is sensitive to small changes in the values in state gain matrix.

6 Conclusion

In this paper, we have designed an LQR state feedback controller for Quanser QUBE-Servo 2 inverted pendulum system. The LQR controller was not only able to stabilize the system, but it also satisfied all control requirements and design constraints. The closed-loop system yielded fast and accurate set-point tracking of the rotary arm angle despite of nonminimum phase system characteristics. The controllers presented in this paper could also be potentially used in other applications e.g., in Segway kind of conveyors.

References

[1] Burl J. B., Linear Optimal Control, H_2 and H_{inf} Methods, Addison Wesley Longman Inc., Menlo Park, Calif., 1999

[2] Belanger P. R., Control Engineering: A Modern Approach, Saunders College Publ., Fort Worth : New York (NY), 1995

[3] Muskinja N., Tovornik B. A., Swinging up and stabilization of a real inverted pendulum, IEEE Transactions on Industrial Electronics, 2006, 53, 631-639, DOI: 10.1109/TIE.2006.870667

[4] Pathak K., Franch K., Agrawal S. K., Velocity and position control of a wheeled inverted pendulum by partial feedback linearization, IEEE Transactions on Robotics, 2005, 21, 505-513, DOI: 10.1109/TRO.2004.840905

[5] Hehn M., D'Andrea R., A flying inverted pendulum, 2011 IEEE International Conference on Robotics and Automation (9 May - 13 May 2011, Shanghai, China), IEEE, Shanghai, 2011, 763-770

[6] Matsuoka K., Williams V., Learning to balance the inverted pendulum using neural networks, Proceedings of the 1991 IEEE International Joint Conference on Neural Networks (18 November - 21 November 1991, Singapore, Singapore), IEEE, Singapore, 1991, 214-219

[7] Takei T., Imamura R., Yuta S., Baggage Transportation and Navigation by a Wheeled Inverted Pendulum Mobile Robot, IEEE Transactions on Industrial Electronics, 2009, 56, 3985-3994, DOI: 10.1109/TIE.2009.2027252

[8] Tang Z., Joo Er M., Humanoid 3D Gait Generation Based on Inverted Pendulum Model, 2007 IEEE 22nd International Symposium on Intelligent Control (1 October - 3 October 2007, Singapore, Singapore), IEEE, Singapore, 2007, 339–344

[9] Williams II R. L., Lawrence D. A., Linear State-Space Control Systems, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007

[10] Ogata K., Modern Control Engineering, 5th ed., Pearson Education Inc., Prentice Hall, 2010

[11] Glyn J., Advanced Modern Engineering Mathematics, 3rd ed., Pearson Education Limited, Harlow, 2011

[12] Maciejowski J.M., Multivariable feedback design, Addison-Wesley Publishers Ltd., Wokingham, 1989

Mats Friman*

Fault-Tolerant Valve Control

Abstract: Valves have a key position in safe, reliable, and economical operation of process industry plants. Valves consist of moving parts subject to wear and dirt. During a fault, it is important that the valve controller eagerly tries to keep the valve under control, despite of changes in the operating environment. In this paper we discuss valve faults on a general level, and we present a solution for keeping valve under control despite of a missing valve position measurement. Our solution utilizes valve/actuator models and real-time simulations to generate a virtual valve position sensor. When valve position measurement is lost, the controller will continue as before, but the real position measurement is replaced with a soft sensor value.

Keywords: industry, valve, valve controller, fault-tolerant control

***Corresponding Author: Mats Friman:** Metso Flow Control, E-mail: firstname.lastname@metso.com

1 Introduction

Process industry plants, such as oil refineries, pulp mills and chemical production plants, are highly automatized. The automatic operation is based on a network of sensors, controllers, and actuators.

The most common actuators are the valves, which are used to control liquid and gas flows in pipelines. An industrial valve consists of three parts: 1) a valve body, attached to the pipeline, 2) a pneumatically powered actuator, and 3) a valve controller, which controls the valve position according to the setpoint obtained from the process automation system [1]. Today's valve controllers are intelligent digital devices with various features that supports plant operation during its entire life time.

The valves have a key position with respect to safe and reliable operation of the plant. It is important that all devices contribute to safe and controlled responses during a failure. A wide range of options must be supported: from continued operations (with some reduced performance) to a safe and controlled shut down.

Some valves may have a critical position regarding plant operation, i.e. if a valve fails, a part of, or the entire plant must be shut down. Therefore, a valve controller should eagerly fight against faults and try to keep valve under control despite of minor faults. A difficult case is loss of valve position measurement. In this paper we discuss some safety and fault-tolerant features of the Neles NDX valve controller, and we present a method for keeping control over the valve in case of a faulty valve position sensor.

2 Valve controller basics

A valve controller has two main tasks. First it receives a valve position setpoint, typically from a process automation system, and secondly, it controls the valve position according to the setpoint.

A typical valve controller is illustrated in Figure 1. The main components include a PCB with integrated sensors, a Prestage unit (an I/P converter), and an Output Stage (a pneumatic relay).

A local user interface enables easy commissioning and a possibility for manual operation and parameter changes. The milliampere signal powers the device and provides analog setpoint and standardized HART digital communication. A set of sensors provide necessary measurements needed for valve control, and a position transmitter enables an analog valve opening signal.

The microprocessor compares valve position measurement to its setpoint and generates an electrical signal to the Prestage. The Prestage pressure actuates the Output stage, which controls air flow into or out of the actuator. The valve controller keeps adjusting the Prestage signal until the valve reaches its desired position [2].

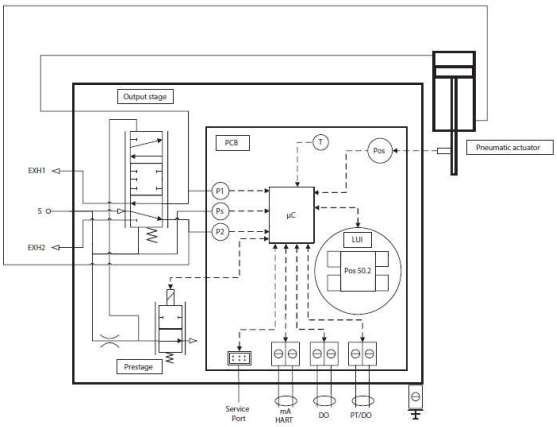


Figure 1. Operating principle of a valve controller.

Valve controllers have traditionally employed some feedback control algorithm for controlling the valve position. If actuator pressure measurements are available, some cascade control structure can be used to speed up valve control. An example cascade control structure is illustrated in Figure 2.

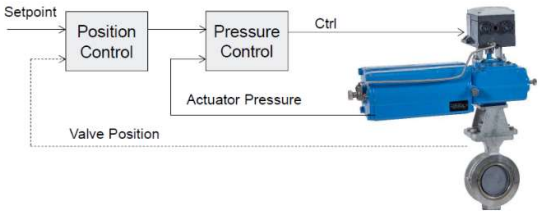


Figure 2. A typical valve control structure based on valve position and actuator pressure measurements.

3 Valve Faults

Various faults may occur during valve operation. We can divide faults into two categories: faults that imply loss of controllability and faults with retained controllability. If we lose valve controllability, e.g. during loss of pneumatic pressure or electrical power, then valve moves to a pre-determined fail-safe position.

If controllability of valve is retained, we can further divide the faults into two categories: 1) mechanical faults and 2) sensor faults. Typically, mechanical faults affect control performance but does not necessarily imply a need for urgent replacement of faulty valve package part. However, depending on the fault and its severity, we typically get some increased position control error or valve hunting as a result.

The most critical sensor for valve control is the valve position sensor. For any other sensor fault, we can switch to simple position feedback control during a fault.

The Neles NDX valve controller position measurement is based on a magnet sensor. This solution has many advantages as it enables accurate position sensing without mechanical links in harsh conditions. However, if the magnet is significantly moved away from its original position or if the positioner or its bracket is moved, tilted, or rotated, it may affect sensor reliability. Therefore, situations may occur where NDX cannot access the position measurement. To prevent an unplanned shutdown because of missing valve position readings, we have developed a method for valve position control without position measurement.

4 Fault-tolerant control

With fault tolerant control we mean the ability of a controller to retain controllability despite of faults in the control network. Usually we tolerate some reduction in the control performance, but the main goal is to continue running the process despite of the fault.

For valves, fault-tolerant control due to mechanical faults and sensor faults are discussed next.

Mechanical faults

Typical problems in valve control include air leakage in pneumatic actuator, increased friction in valve body, freezing, and valve controller faults (e.g. defectives in pneumatic components). These faults may affect valve control accuracy, but normally we can continue operation without a need for replacement of devices or spare parts.

For mechanical faults, fault-tolerant control typically means detuning of controller to avoid valve hunting. Large control errors may also need special attention.

Mechanical faults are recognized by valve controller diagnostics, and the faults are communicated to the maintenance organization [3].

Auxiliary sensor faults

In addition to valve position, which is the controlled value, valve controllers typically have auxiliary sensors, which speeds up and improves position control. Auxiliary sensors are e.g. supply and actuator pressures, and temperature. In case of a failure in any of the auxiliary sensors, the controller switches to a position-feedback mode where control actions are based on valve position only.

Faults in position sensor

On a general level, a feedback control loop needs both a setpoint and a measurement. A fault in the controlled

value (i.e. measurement fault) prevents us from utilizing feedback control. Instead, feedforward control must be used to ensure that the (unmeasured) controlled value responds to changes in the setpoint.

For a valve controller, if there is a fault in the position sensor, the only option is *feedforward* control of valve position.

An intuitive solution for the cascade control setup in Figure 2, where the dotted line indicates a faulty or missing position measurement, would be replace the "Position Control" block, with e.g. a look-up table that picks a setpoint for actuator pressure based on given position setpoint. In this case, the look-up table would act as a feedforward controller, and replaces the feedback controller, which cannot operate because of the missing measurement needed for feedback control.

Next, we will present an alternative solution. We will replace the missing position measurement with a soft-sensor and continue with the same feedback controller as before [4]. The advantage of this solution is that there is no need for a separate feedforward controller. Instead, the same feedback controller can be employed both for ordinary feedback control and for control during a fault in position measurement. All we need is a model and a valve position simulation engine, which generates a virtual valve position value during a valve position failure. A simple model, which is easy to simulate is utilized [5,6,7].

Valve and actuator model

Consider a single acting, spring/piston actuator connected to a valve (Figure 3). The actuator consists of a spring pushing in one direction, and air pushing in the opposite direction. When air flows into/out of actuator, actuator pressure changes, and valve moves.

Compressed air in the actuator initiates a force that is proportional to air pressure. According to Hook's Law, the spring force is proportional to spring contraction [8] and considering pneumatic and spring forces (before considering friction) we notice that actuator travel (and valve opening) is proportional to actuator pressure. Introducing Coulomb friction, the net spring and pneumatic force must exceed the Coulomb friction threshold to ensure that the valve is moving.

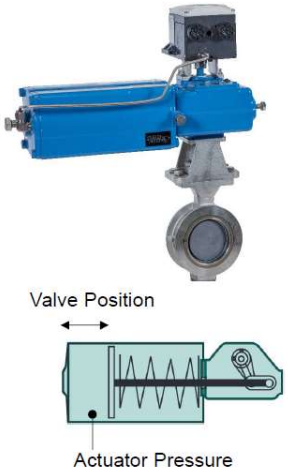


Figure 3. A valve package (above) and a detailed view of the single acting spring-return actuator (below).

A typical response of actuator movement to pressure changes is shown below (Figure 4) where we have plotted valve position vs. actuator pressure for an example actuator. Different colors indicate different movement directions. From this figure it is clearly seen that valve position is linear with respect to actuator pressure for each movement directions. However, because of friction forces, there is a clear gap between movement up and down curves (i.e. the Coulomb friction).

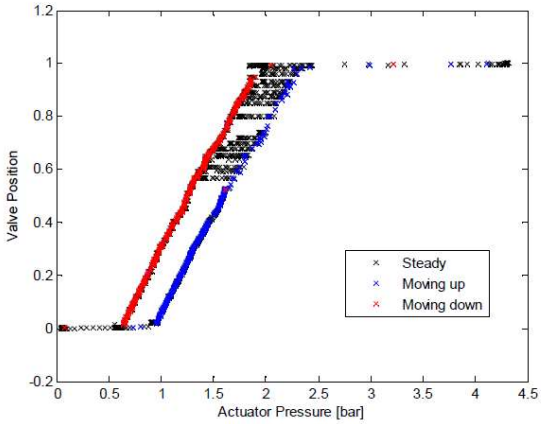


Figure 4. Valve position vs. actuator pressure for an example spring-return actuator.

Soft sensor

Above we observed a linear relationship between valve position and actuator pressure, when moving in one direction (up/down cases indicated by red/blue colors). Based on this finding we used the following equality for estimating new valve position h_e actuator pressure p_a

$$h_e = \begin{cases} \min\left(1, max\left(0, \frac{p-p0-pf}{k_s}\right)\right) & if p > h_o k_s + p_0 + p_f \\ \min\left(1, max\left(0, \frac{p-p0+pf}{k_s}\right)\right) & if p < h_o k_s + p_0 - p_f \\ h_o & \text{otherwise} \end{cases} \quad (1)$$

where the parameters p_0 is the actuator pressure which equals spring pretension, p_f is net coulomb friction in pressure units, and k_s is spring constant (in pressure units / full stroke). The simulated position h_o is the only state variable needed for the simulation.

The parameters of Eq. 1 are identified during device calibration. Device calibration includes an automatic tuning sequence. This tuning sequence moves valve in both directions, which enables identification of the three parameters p_0, p_f, and k_s of Eq. 1. Note that for valve opening values other than extreme values (fully open/close), Eq. 1 is linear in the parameters.

The valve controller can switch to missing-valve-position-measurement mode automatically if it recognizes problems in position measurement. Alternatively, we can manually switch to fault-tolerant mode.

5 Results

We tested the suggested fault-control strategy by running a control valve in the laboratory. During the test we used a manual mode selector to switch between normal control and fault-tolerant mode with real position measurement replaced by soft-sensor value.

To demonstrate the robustness of the suggested method, we selected a high-friction valve for testing. For the test valve, the pressure change needed for valve reversal is 0.6 bar (i.e. pressure to compensate for friction), which can be compared to pressure change of 1.0 bar needed to compensate for spring forces during entire moving range (close to open). With such a large friction values, it is difficult to position the valve, especially when running in fault-tolerant mode.

An example test run is shown below in Figure 5 where trends for ordinary control which uses valve position, and fault-tolerant mode are shown. We used a setpoint sequence consisting of a ramp, and some step changes. The colors indicate the two different experiments: blue lines for ordinary control (which utilizes position measurement) and green lines for fault-tolerant control mode (when position measurement was neglected by the controller but recorded for trend plotting purposes).

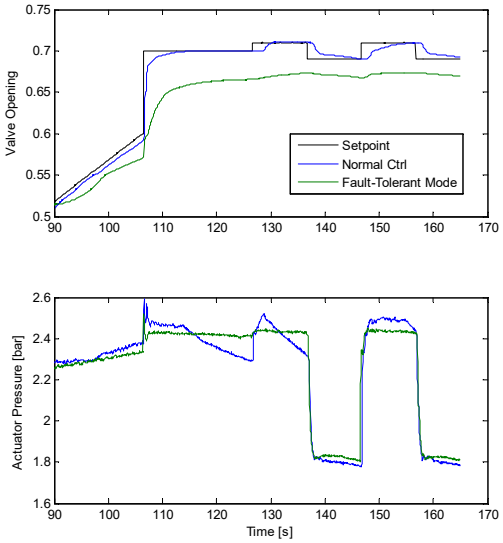


Figure 5. Comparison of control performance with normal control (blue) and with fault-tolerant control (green), which does not utilize position measurement. Above valve opening and setpoint, below actuator pressure.

6 Summary

We have developed a method for keeping a valve under control despite of loss of valve position measurement. Our solution is to replace the missing position measurement with a virtual measurement obtained from real-time simulations of valve. The same controller is used in both modes: closed-loop control (with position measurement from real sensor), and fault-tolerant control mode (with virtual measurement used for control).

Our test results from running a high-friction valve in laboratory suggest that valve control based on a virtual measurement works very well. The results demonstrate that the control accuracy suffers a little bit, as the control error increases with a few percentage points when operating the valve in position sensor-fault mode.

Because of the missing position sensor, it is impossible to for the valve the valve to follow its setpoint exactly. This is not a serious problem for valves operated by a PID control loop, because valve position errors are compensated by the PID controller. Control loops operated in manual mode, on the other hand, are expected to have a steady-state deviation in valve position.

The advantages with the suggested feature is that we can avoid an unplanned shut-downs of plant. This is expected to provide cost savings, added flexibility and more options for maintenance planning.

References

[1] Kirmanen J., Niemelä I., Pyötsiä J., Simula M., Hauhia M., Riihilahti J. The Flow Control Manual,.4th ed. Metso Automation. 1997

[2] Metso Flow Control: NELES® INTELLIGENT VALVE CONTROLLER, SERIES NDX. Technical Bulletin. 2016.

[3] Manninen T. Fault Simulator and Detection for a Process Control Valve. PhD Thesis, Aalto University, Espoo, Finland 2012.

[4] Friman M., Heikkinen P. Method and Controller for Actuator, Patent Application, WO2018/055229Al, 2018

[5] Hietanen V., Friman M., Pyötsiä J., Manninen T. Laatua järjestelmien simuloinnista. Automaatio XIX seminar. Finnish Society of Automation. Helsinki, Finland. 2011.

[6] Friman M. Model-Based Design: Experiences from Valve Controller Development. Automaatio XXII seminar. Finnish Society of Automation. Vaasa, Finland. 2017.

[7] Pyötsiä J. A Mathematical model of a Control Valve. PhD Thesis. Helsinki University of Technology, Espoo, Finland. 1991.

[8] Wikipedia. https://en.wikipedia.org/wiki/Hooke%27s_law Accessed 10.3.2019

Jussi Sihvo*, Joona Leinonen, Tomi Roinila ja Tuomas Messo

Jatkuva-aikaiset impedanssimittaukset osana älykkäitä akkujärjestelmiä

Tiivistelmä:

Tutkimukset ovat osoittaneet, että Li-ion akun sisäistä impedanssia voidaan tehokkaasti hyödyntää akun varaustilan ja elinkaaren analyysiin, jolloin myös akkujärjestelmän luotettavuutta ja turvallisuutta voidaan merkittävästi parantaa. Tässä työssä esitellään kehittyneitä signaalinkäsittelyn menetelmiä Li-ion akun sisäisen impedanssin nopeaan ja luotettavaan mittaukseen. Menetelmässä hyödynnetään laajakaistaista, pseudo-satunnaista binääristä herätesignaalia akun impedanssin nopeaan mittaamiseen. Menetelmässä akkua herätetään (ladataan/puretaan) pieniamplitudisella herätteellä, ja tämän herätteen aiheuttama vaste (virta/jännite) mitataan. Tämän perusteella voidaan laskea suoraviivaisesti akun impedanssi. Menetelmä mahdollistaa impedanssin laskemisen murto-osassa verrattuna aikaan, joka kuluu perinteisellä impedanssispektroskopialla. Lisäksi, menetelmä on mahdollista toteuttaa erittäin edullisesti, mikä antaa mahdollisuuden implementoida teknologia useiden eri akkusovellusten yhteyteen.

Avainsanat: Litium-ioni akku, Akun impedanssi, jatkuva-aikaiset mittaukset, älykkäät akkujärjestelmät, kuntotila, varaustila

***Vastaava kirjoittaja: Jussi Sihvo:** Tampere University of Technology, E-mail: jussi.sihvo@tuni.fi

Joona Leinonen: Tampere University of Technology, E-mail: joona.leinonen@tuni.fi

Tomi Roinila: Tampere University of Technology, E-mail: tomi.roinila@tuni.fi

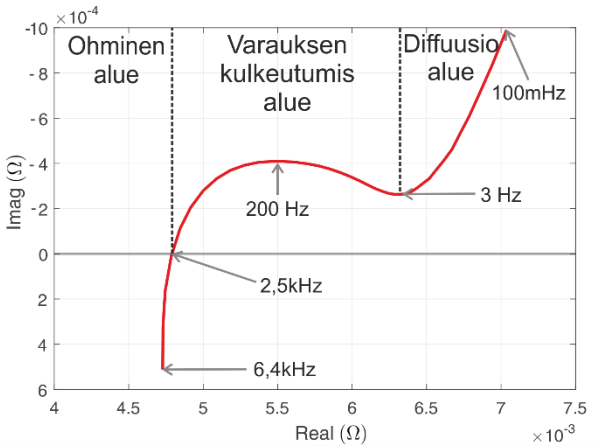
Tuomas Messo: Tampere University of Technology, E-mail: tuomas.messo@tuni.fi

1 Johdanto

Li-ion akkuja hyödyntävien sovellusten, kuten esimerkiksi sähköisen liikenteen ja uusiutuvan energian, määrä on merkittävästi kasvanut viimeisten vuosien aikana. On arvioitu, että akkujen globaalit markkinat ylittävät 100 miljardia euroa vuoteen 2030 mennessä [1]. Tämä on asettanut haasteen kiertotaloudelle, sillä noin 95% akuista päätyy jätteeksi kierrättämisen sijaan. Tutkimukset ovat kuitenkin osoittaneet, että jopa 95% li-ion akuista olisi mahdollisuus uusiokäyttää osittain tai kokonaan [2].

Li-ion akut tarvitsevat toimiakseen akun valvonta-järjestelmän (battery-management-system (BMS)), joka mahdollistaa akun turvallisen, sekä optimaalisen käytön. Akun tilaa indikoidaan akun valvonta-järjestelmässä akun tilaparametrien, kuten varaustilan (state-of-charge (SOC)) ja kuntotilan (state-of-health (SOH)) avulla. Varaustila kertoo, kuinka paljon akussa on varausta jäljellä ja kuinka paljon akku pystyy toimittamaan energiaa [3]. Akun kuntotila taas indikoi akun jäljellä olevaa käyttökapasiteettia, sekä akun kykyä toimia sen nykyisessä sovelluksessa [4]. Parametrien tarkka määrittäminen on kuitenkin haasteellista, sillä SOC ja SOH riippuvat myös akussa vallitsevista olosuhteista, kuten lämpötilasta ja akun sisällä tapahtuvista kemiallisista reaktioista [5]. Lisäksi kyseiset parametrit täytyy määrittää epäsuorasti akun jännite-, virta- ja lämpötilamittauksin [6]. Erityisesti akun kuntotilan monitorointi on haasteellista, sillä akun kunto riippuu myös nykyisen kapasiteetin lisäksi hyvin paljon siitä, kuinka ja missä olosuhteissa akkua on aiemmin käytetty ja kuinka kapasiteetti on laskenut. Luotettavaan kuntotilan monitorointiin tarvitaankin läpi akun eliniän kestävää monitorointia, jotta vikaantuneet kennot voidaan havaita ennen kuin niiden kunto on laskenut liikaa [7].

Akun kapasiteetti voidaan mitata purkamalla akku täydestä tyhjäksi ja laskea akusta purettu varauksen määrä. Kyseinen menetelmä kuitenkin vie hyvin paljon aikaa ja kuluttaa akkua turhaan eikä näin ollen sovellu akun kuntotilan mittauksiin. Tutkimukset ovat kuitenkin osoittaneet, että Li-ion akun sisäinen impedanssi vaihtelee paitsi akun kapasiteetin ja kuntotilan, myös varaustilan funktiona [3-8]. Akun impedanssin mittaus tarjoaa siis vaihtoehtoisen tavan määrittää akun kunto- ja varaustila. Akun impedanssi voidaan mitata elektrokemiallisen impedanssi-spektroskopian (EIS) avulla. EIS-menetelmässä akkua puretaan/ladataan sinimuotoisella virta- herätteellä, jonka tuottama jännitevaste mitataan



Kuva 1. Akun impedanssi ja sen eri alueet kuvattuna kompleksi-tasossa

jännitesensoreilla. Menetelmällä voidaan impedanssi mitata tarkasti ja luotettavasti, mutta tekniikka on huonosti sovellettavissa käytännön sovelluksiin sen hitauden ja kompleksisuuden takia.

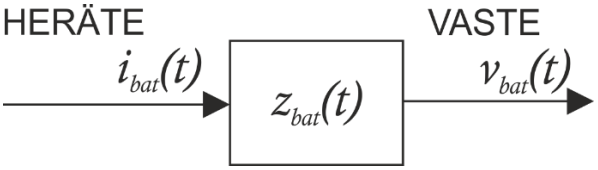
Tässä työssä hyödynnetään laajakaistaista, pseudo-satunnaista binääristä herätesignaalia (PRBS) sekä Fourier-tekniikoita akun impedanssin mittauksiin [11]. Menetelmässä akkua herätetään (ladataan/puretaan) pieniamplitudisella, kaksitasoisella herätesyklillä, jonka avulla akun impedanssi voidaan mitata useilla eri taajuuksilla samanaikaisesti. Näin mittaukset voidaan suorittaa vain murto-osalla ajasta joka kuluisi perinteisiin EIS-menetelmän mittauksiin. Nopeutensa ansiosta PRBS-menetelmää voi käyttää jatkuva-aikaisiin sovelluksiin, joita ovat esimerkiksi akun impedanssin muutosten hyödyntäminen akun varaustilan ja kuntotilan estimointiin. Lisäksi, menetelmä vaatii ainoastaan kaksi eri signaalitasoa, mikä mahdollistaa menetelmän yksinkertaisen ja edullisen toteutuksen esimerkiksi jo akkujärjestelmässä olemassa olevan akun balansointipiirin yhteyteen. Tässä artikkelissa menetelmästä julkaistut tulokset pohjautuu julkaisuissa [9-10] esiteltyihin tuloksiin ja teoriaan.

2 Impedanssin vaikutus akun tilaan

Akun impedanssille voidaan taajuustasossa kirjoittaa

$$Z_{\mathrm{bat}}(j\omega) = \frac{V_{\mathrm{bat}}(j\omega)}{I_{\mathrm{bat}}(j\omega)} \qquad (1)$$

missä V_{bat} on akun napajännite, I_{bat} akun läpi menevä virta ja ω kulmataajuus. Tyypillisesti kirjallisuudessa akun impedanssi esitetään kompleksitasossa, ja se voidaan jakaa eri alueisiin, jotka muodostuvat tiettyjen kemiallisten reaktioiden tuloksena. LiFEPO4-akkukennon impedanssi kompleksitasossa, sekä sen eri alueet on esitetty kuvassa 1. Matalat taajuudet



Kuva. 2. Impedanssin yksinkertainen lohkokaavio

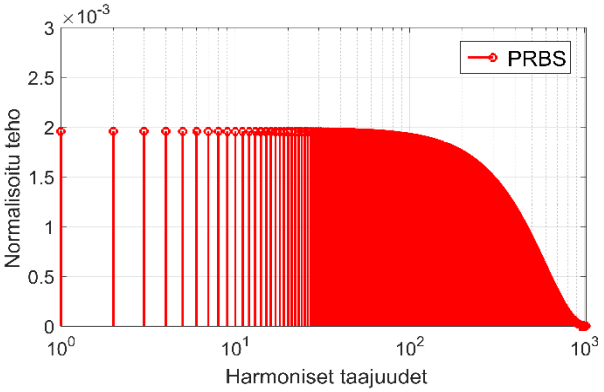
impedanssista muodostaa ns. diffuusioalueen (kuvassa 1 ”diffusion”), joka muodostuu litium-ioneiden kulkeutumisesta akun napojen elektrodeissa. Keskitaajuuksilla impedanssia kutsutaan varauksen-kulkeutumisalueeksi (kuvassa 1 ”charge-transfer”), joka muodostuu varauksen-kuljettajien kulkeutumisesta elektrolyyttimateriaalin ja elektrodien rajapinnan läpi. Suurilla taajuuksilla impedanssi taas koostuu lähinnä akun varauksenkuljettajien ja elektrodien resistiivisyydestä (Ohmic).

On huomion arvoista, että akun impedanssin riippuvuus varaus- tai kuntotilasta on eri suuruinen riippuen impedanssin alueesta. Esimerkiksi litium-ioni akkujen varaustilan muutokset näkyvät parhaiten impedanssin diffuusioalueella pienillä taajuuksilla [5-6][8-9], kun taas akun kuntotila voidaan havaita ohmisella alueella suurilla taajuuksilla [4-5]. Impedanssin muuttumisen suuruutta voidaan havainnoida mm. impedanssin sähköisen vastinpiirin parametrien muuttumisen kautta [6].

3 Impedanssin mittaus ulkoisella herätteellä

Akun impedanssi voidaan ymmärtää siirtofunktiona akun virran (sisäänmeno) ja jännitteen välillä (ulostulo), jolloin akun virta toimii herätteenä akulle tuottaen jännitevasteen. Tilannetta kuvaa kuvan 2 yksinkertainen lohkokaavio. Herätesignaalilla ja sen ominaisuuksilla on merkittävä vaikutus myös vasteeseen, minkä takia on suotavaa käyttää herätettä, jonka taajuussisältö on tunnettu. Yleensä akun impedanssimittaukseen herätesignaalina on käytetty sinimuotoista herätettä, jolla on tietty amplitudi ja taajuus. Kyseistä menetelmää kutsutaan myös elektrokemialliseksi impedanssispektroskopiaksi (EIS) ja on kirjallisuudessa laajalti käytetty menetelmä [4-6][8]. Sen avulla saadaan hyvinkin tarkkoja tuloksia, mutta menetelmän käytettävyys muissa, kuin laboratorio-olosuhteissa on rajoittunutta pitkälti tekniikan hitauden takia. EIS-menetelmässä joudutaan jokainen taajuuspiste mittaamaan erikseen, mikä vie aikaa varsinkin, jos mitataan todella matalia taajuuksia. Lisäksi sinimuotoisen virtaherätteen toteutus vaatii kompleksisen ja tilaa vievän toteutuksen, mikä lisää sovelluksen kustannuksia.

Mikäli impedanssin mittaukset halutaan yhdistää osaksi



Kuva 3. PRBS:n taajuustason energiaspektri

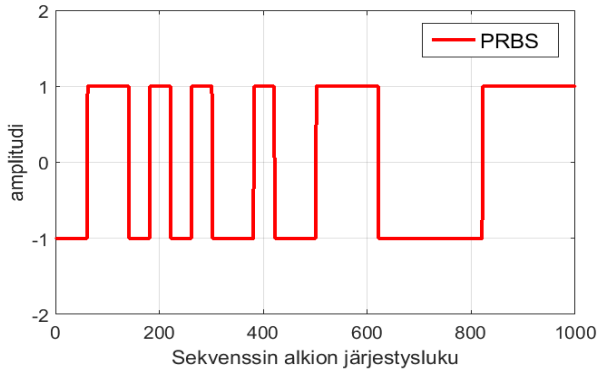
akkujärjestelmää, täytyy mittausmenetelmän olla paitsi nopea ja luotettava, myös digitaalisesti ja fyysisesti kevyt toteuttaa. Niin sanottujen laajakaista-menetelmien avulla impedanssinmittauksia voidaan merkittävästi nopeuttaa. Laajakaista-menetelmissä herätesignaali sisältää energiaa useammalla kuin yhdellä taajuudella, jolloin kyseiset taajuudet voidaan mitata samanaikaisesti. Tällainen signaali voidaan muodostaa esimerkiksi eri taajuuksisten siniaaltojen summana [10], mutta menetelmä on siniaaltojen käytöstä johtuen edelleen digitaalisesti varsin raskas toteuttaa. Systeemin identifiointimenetelmistä on kuitenkin historiassa tutkittu myös niin sanottuja pseudo-satunnaisia-binäärisiä-signaaleja (pseudo-random-binary-sequence (PRBS)) [11]. Kyseiset signaalit omaavat laajakaistamenetelmien nopeuden, mutta ne ovat myös yksinkertaista toteuttaa, sillä heräte voidaan karakterisoida vain kahdella eri signaalitasolla. Kyseisiä menetelmiä on myös akun impedanssin mittaukseen tutkittu [12-13]. Tutkimuksissa on kuitenkin raportoitu menetelmän epätarkkuutta johtuen, sekä akun epälineaarisuuksista, sekä akun toimintapisteen muuttumisesta kesken mittausten. Kuitenkin oikein toteutettuna ja suunniteltuna on osoitettu, että PRBS-menetelmällä voidaan saada varsin tarkkoja tuloksia [9-10].

3.1 PRBS –heräte

PRBS –heräte on tarkkaan suunniteltu binäärinen sekvenssi, jolla on energiaa useilla eri taajuus-komponenteilla. Lisäksi, ideaalitapauksessa, taajuus-komponenttien energiat ovat yhtä suuret mikä lisää mitatun taajuusvasteen mittausvarmuutta. PRBS signaalin generointialgoritmi voidaan esittää muodossa

$$s_{PRBS}(n+i) = \sum_{r=1}^{n} C_r s_{PRBS}(i-r) \quad mod\ 2, \quad (2)$$

missä s_{PRBS} on itse signaali ja parametri C sisältää n asteisen primitiivisen polynomin kertoimet. Signaalin seuraava alkio määritetään aina käyttäen signaalin n-



Kuva 4. PRBS:n aikatason esitys

edellistä alkiota, mistä johtuen sekvenssin n-ensimmäistä alkiota täytyy alustaa alkuarvoihin. Alku-arvoiksi käy kaikki muut numeroiden 0 ja 1 kom-binaatiot paitsi puhdas nollavektori. Koska muodostuva PRBS on binäärinen, täytyy aritmetiikka kaavassa (2) toteuttaa moduulissa 2. Huomionarvoista on, että C voi olla minkä tahansa polynomin kertoimet, joka on astetta n ja primitiivinen moduulissa 2. [11]

3.1.1 Herätesignaalin suunnittelu

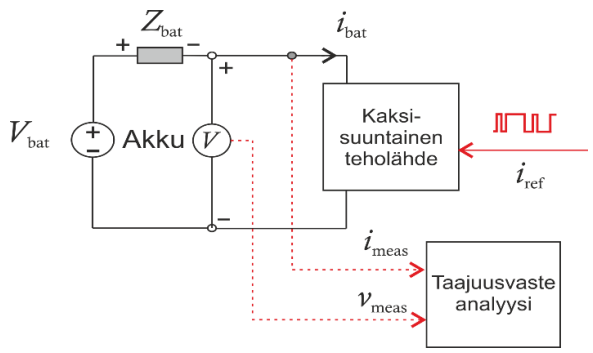
PRBS sekvenssin ominaisuuksia aika- ja taajuustasossa on esitetty kuvissa 3 ja 4. Tärkeimmät parametrit PRBS taajuussisällön kannalta ovat PRBS:n pituus ja gene-rointitaajuus. PRBS:n pituus N muodostuu kaavasta

$$N = 2^n - 1, \quad (3)$$

missä n on kaavassa (2) käytetty primitiivisen polynomin asteluku. Tällöin puhutaan myös n-bittisestä sekvenssistä. Generointitaajuus f_{gen} on PRBS:n peräkkäisten alkioiden näytteistystaajuus. Generointitaajuus määrittää myös suurimman taajuuskomponentin, jolla PRBS:llä on energiaa ja joka herätteellä voidaan mitata. Käytännön mittauksissa PRBS:n jokaista alkiota tulee myös erikseen näytteistää. Tämän näytteistystaajuuden on Nyquist-teoreeman mukaan oltava vähintään kaksinkertainen generointitaajuuteen verrattuna, jotta PRBS:n taajuussisältö pysyy mahdollisimman muuttumattoma-na. Tästä tosin aiheutuu PRBS:n taajuuskomponenttien energian pieneneminen suurilla taajuuksilla, kuten kuvassa 3 on esitetty. Tästä johtuen luotettava taajuussisältö rajoittuukin todellisuudessa noin puoleen generointitaajuudesta. Yhdessä N ja f_{gen} muodostavat sekvenssin taajuusresoluution kaavalla

$$f_{res} = \frac{f_{gen}}{N}, \quad (3)$$

joka määrittää myös pienimmän taajuuden, joka voidaan mitata. Siitä saadaan myös teoreettinen



Kuva 5. Mittausten periaatekuva

mittauksiin tarvittava vähimmäisaika, joka on $t_{meas} = \frac{1}{f_{res}}$. Taajuusresoluutiolla on siis varsinkin akun impedanssin mittauksiin merkittävä vaikutus, sillä impedanssi sisältää tietoa myös hyvin pienillä taajuuksilla varsinkin akun varaustilasta. Toisaalta lyhyttä mittausaikaa voidaan vaatia sovelluksen puolesta varsinkin reaali-aikasovelluksissa. Lisäksi, hyvin pienillä taajuuksilla akun diffuusioalueella myös akun epälineaarisuudet nousevat dominoivaksi, jolloin mittaustuloksiin tulee epätarkkuutta. Kompromissi sallitun mittausajan ja mitattavan taajuuskaistan välillä voidaankin siis määrittää taajuusresoluution avulla.

Sovelluskohtaisesti tärkeä parametri on myös sekvenssin amplitudi, joka akkumittauksissa käytännössä määrittää akun virran. Suurempi amplitudi parantaa mittausten luotettavuutta, mutta aiheuttaa myös häviöitä ja voi aiheuttaa myös epälineaarisuuksia mitattavaan vasteeseen tehden siitä epäluotettavan. Verrattuna muihin sovelluksiin, akkusovelluksissa amplitudin täytyy olla varsin suuri, jotta mitään vaihtelua napajännitteessä voidaan havaita. Lisäksi sekvenssin DC-taso tärkeää asettaa nollaksi kuten kuvassa 4, jotta akun toimintapiste voidaan pitää mittausten ajan mahdollisimman muuttumattomana.

4 Mittaukset

Käytännön mittaukset toteutettiin laboratorio-olosuhteissa, missä akun impedanssin taajuusvaste mitattiin litium-rauta-fosfaatti kennosta (LiFEPO4). Akun kapasiteetti oli 2.3Ah ja nimellisjännite 3.3V. Mittausten lohkokaavio on esitettynä kuvassa 5, missä PRBS -heräte syötetään järjestelmään kaksisuuntaisen teholähteen virtaohjeena, joka tällöin vastaa myös akun virtaa. Järjestelmän jännite- ja virtamittaukset suoritettiin samalla näytteistyksellä ja saman pituisina kuin PRBS heräte. Akun impedanssi mitattiin akkukennosta eri varaustiloissa 10% resoluutiolla välillä 20% - 100%. Akkua purettiin mittausten välillä 1C:n suuruisella virralla, joka vastaa akun kapasiteetin suuruista virtaa. Varaustilaa monitoroitiin yksinkertaisella varauksen summausmenetelmällä [6].

Taulukko 1. mittausmenetelmien parametrit

	EIS	PRBS
Amplitudi	0.7A – 2A	1.15A
Taajuusalue	0.2Hz – 3kHz	0.2Hz – 3kHz
Generointitaajuus	-	3kHz
Sekvenssin pituus		32767
Näytteistystaajuus	-	6kHz

PRBS –herätteen suunnittelu on täysin riippuva akkukennon ominaisuuksista. Taajuusalue, jolla impedanssin muutokset ja trendit ovat suurimpia, riippuu paljon akkukennon kemiasta. Taajuusalue riippuu myös siitä, halutaanko akun impedanssista mitata kaikki kuvassa 1 esiintyvät alueet vai ainoastaan jokin tai jotkut niistä. Tässä artikkelissa mittausalue valittiin ulottumaan diffuusioalueelta ohmiselle alueelle, mikä kyseiselle LiFEPO4-kennolle vastaa taajuusaluetta 100mHz – 3kHz varmuusrajoineen. Vastaavat mittaukset suoritettiin myös EIS-menetelmällä käyttäen samaa taajuusaluetta. Suunnitteluparametrit sekä PRBS-, että EIS-menetelmille mittauksia varten on esitetty taulukossa 1.

4.1 Mittausdatan jatkokäsittely

Jännite- ja virtamittaukset muunnetaan taajuustasoon diskreetin Fourier-muunnoksen (DFT) avulla, joka voidaan matemaattisesti ilmaista

$$F_i = \frac{1}{N}\sum_{k=0}^{N-1} f_k e^{-\frac{j2\pi i}{N}k}, \quad (4)$$

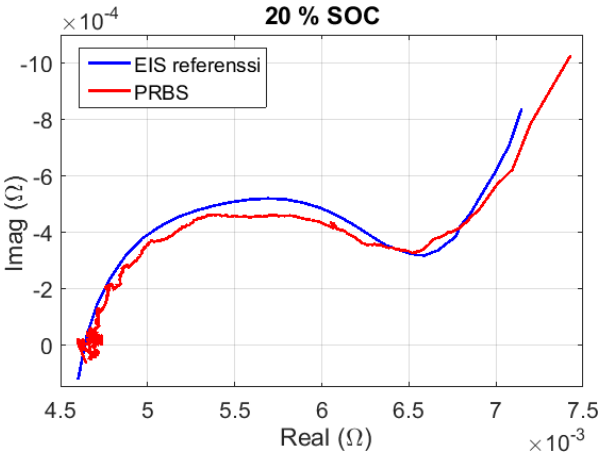
missä F on muunnettu taajuustason signaali, f muunnettava aikatason ja N muunnettavan signaalin pituus. Lisäksi mittaustuloksiin lisääntyvän mittaus-kohinan poistoon käytettiin liukuvan-keskiarvon-suodatinta (moving-average-filter (MAF)) [10][14]. Kohinan poistoon voi myös käyttää peräkkäisten herätteitten keskiarvotusta, mutta tällöin mittausten kokonaisaika kasvaa ja menetelmän käytettävyys jatkuva-aikaisissa sovelluksissa heikkenee [9].

4.2 Mittaustulokset ja analysointi

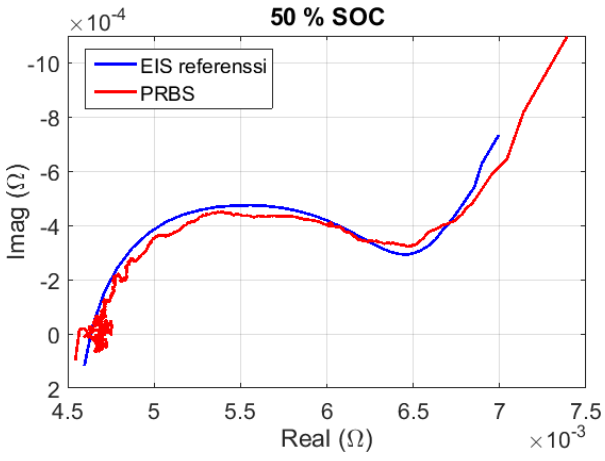
PRBS-menetelmällä saadut mittaustulokset rinnastettuna EIS-referenssimittauksiin on esitetty kuvissa 6-8, joista menetelmien vastaavuus voidaan hyvin todentaa. Ainoastaan diffuusioalueen vasteissa erot ovat

Taulukko 2. Mittauksiin kulunut aika eri herätesignaaleilla

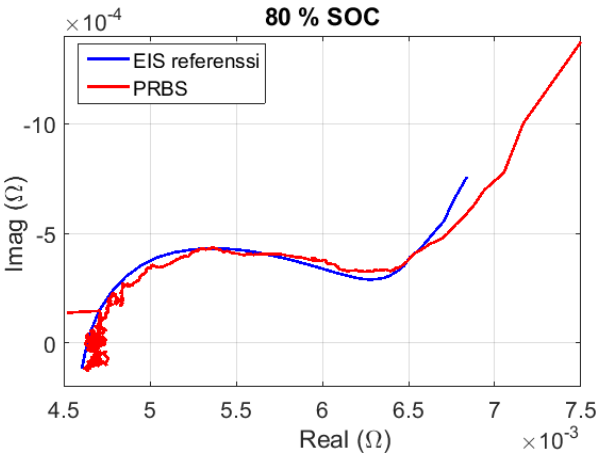
	EIS	PRBS
Mittausaika	60s	4.7s



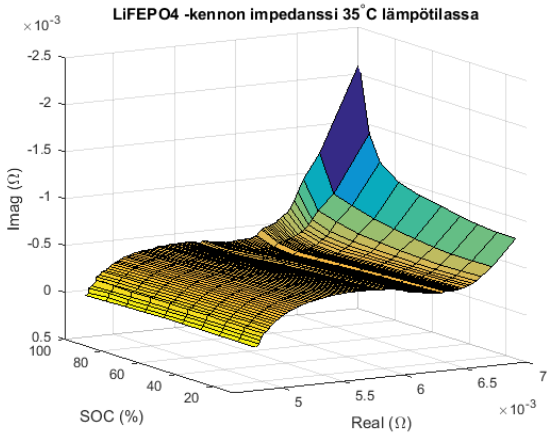
Kuva 6. Mitatut impedanssit 20% varaustilassa 35-asteen lämpötilassa



Kuva 7. Mitatut impedanssit 50% varaustilassa 35-asteen lämpötilassa



Kuva 8. Mitatut impedanssit 80% varaustilassa 35-asteen lämpötilassa



Kuva 9. PRBS-menetelmällä mitattu impedanssi varaustilan funktiona 35-asteen lämpötilassa

hieman suurempia, mikä johtuu osittain diffuusio-alueen karakterisoivista matalataajuuskomponenteista. Mittausten aikana akun toimintapiste ei pysy täysin vakiona, mikä kertautuu erityisesti pienillä taajuuksilla, jolloin mitattava aikavakio on pidempi. Kuitenkin PRBS-mittaukset todentavat akun impedanssin muuttumisen varaustilan funktiona erityisesti pienillä taajuuksilla, kuten kuvasta 9 voidaan todeta (taajuus kasvaa reaaliakselia oikealta vasemmalle mentäessä). Jo aikaisemmin mainittu etu on PRBS:n mittausajan merkittävä lyhentyminen verrattuna EIS-menetelmään. Taulukosta 2 voidaan huomata, että mittausaika on EIS-menetelmällä 12-kertainen verrattuna PRBS-menetelmään. Lisäksi PRBS:n kaksi signaalitasoa voidaan akkujärjestelmässä helposti luoda minimissään käyttäen yhtä kytkintä, jolloin menetelmän integrointikustannukset sovellukseen pysyvät hyvin pieninä.

Merkittävin käyttösovellus PRBS-menetelmällä on akkujärjestelmässä akun kennojen kuntotilan estimoinnissa, sillä impedanssilla on yhteys akun kapasiteettiin, kuten jo aiemmin todettiin. Näin voidaan järjestelmän älykkäällä monitoroinnilla havaita mahdolliset vaihdettavat kennot ja vaihtaa ne. Lisäksi menetelmää voi myös käyttää akun varaustilan määrittämiseen, vaikka varaustilan estimointi onkin nykypäivän litium-ioni akkujärjestelmissä jo toteutettu. Varaustilan estimoinnin tarkkuus kuitenkin korostuu, kun puhutaan teholtaan ja energialtaan yhä suuremmista sovelluksista ja tällöin jatkuva-aikaiset impedanssi-mittaukset voivat parantaa estimoinnin tarkkuutta.

5 Yhteenveto

Tässä artikkelissa käsiteltiin PRBS-menetelmän käyttöä akun jatkuva-aikaisiin impedanssimittauksiin. Menetelmällä suoritettiin käytännön mittaukset litium-rauta-fosfaatti akkukennosta ja tuloksia verrattiin perinteisellä EIS-menetelmällä mitattuihin referenssituloksiin. Mittauksissa todettiin menetelmän antavan luotettavan vasteen mitatulle impedanssille.

Luotettavuuden lisäksi menetelmä on myös nopea, sillä sen avulla voidaan mitata akusta impedanssi vain murto-osassa perinteisellä EIS-menetelmällä kuluvaan mittausaikaan. Mittausaika oli PRBS-menetelmällä vain 4.7s, jolloin voidaan jo tietyissä tilanteissa puhua jatkuva-aikaisesta käytöstä. Menetelmä on myös digitaalisesti kevyt toteuttaa käytännön sovellukseen, sillä se vaatii ainoastaan kaksi signaalitasoa. Mitattua impedanssia voidaan käyttää mm. akkujärjestelmän kennojen kuntotilan, sekä varaustilan estimointiin. Yhdistettäessä impedanssinmittaukset ja analysointi osaksi akun valvontajärjestelmän algoritmeja, voidaan akkujärjestelmän turvallisuutta ja luotettavuutta merkittävästi parantaa.

Lähteet

[1] J. S. John, “Global energy storage to double 6 times by 2030, matching solars spectacular rise,” Bloomberg New Energy Finance (report), 2017.

[2] S. King, N. Boxall, and A. Bhatt, “Lithium battery recycling in Australia - Current status and opportunities for developing a new industry,” CSIRO, 2018.

[3] J. Rivera-Barrera, N. Munoz-Galeano, and H.Sarmiento-Maldonado, “SoC Estimation for Lithium-ion Batteries: Review and Future Challenges,” Electronics, vol. 6, no. 4, p. 102, 2017.

[4] D. I. Stroe, M. Swierczynski, S. K. Kær, and R. Teodorescu, “Degradation Behavior of Lithium-Ion Batteries During Calendar Ageing - The Case of the Internal Resistance Increase,” IEEE Trans. Ind. Appl., vol. 54, no. 1, pp. 517–525, 2018.

[5] J. Vetter, P. Nov, M. R. Wagner, and C. Veit, “Ageing mechanisms in lithium-ion batteries,” Journal of Power Sources, vol. 147, pp. 269–281, 2005.

[6] P. Weicker, A system Approach to Lithium-ion Battery management. Artech House, 2013, no. 2013.

[7] M. Berecibar, I. Gandiaga, I. Villarreal, N. Omar, J. Van Mierlo, and P. Van den Bossche, “Critical review of state of health estimation methods of Li-ion batteries for real applications,” Renewable and Sustainable Energy Reviews, vol. 56, pp. 572–587, apr 2016.

[8] A. Zenati, P. Desprez, and H. Razik, “Estimation of the SOC and the SOH of Li-ion batteries, by combining impedance measurements with the fuzzy logic inference,” IECON Proceedings Industrial Electronics Conference), pp. 1773–1778, 2010.

[9] J. Sihvo, T. Messo, T. Roinila, and R. Luhtala, “Online internal impedance measurements of li-ion battery using prbs broadband excitation and fourier techniques: Methods and injection design,” in 2018 International Power Electronics Conference (IPEC-Niigata 2018 -ECCE Asia), May 2018, pp. 2470–2475.

[10] J. Sihvo, T. Messo, T. Roinila, and D. I. Stroe, “Online identification of internal impedance of Li-ion battery cell using ternary-sequence injection,”in Energy Conversion Congress and Exposition (ECCE), 2018 IEEE, 2018, pp. 1–7.

[11] K. Godfrey, Perturbation Signals for System Identification. Prentice Hall, 1994.

[12] R. Al Nazer, V. Cattin, P. Granjon, M. Montaru, M. Ranieri, and V. Heiries, “Classical EIS and square pattern signals comparison based on a well-known reference impedance,” World Electric Vehicle Journal, vol. 6, no. 3, pp. 800–806, 2013.

[13] A. J. Fairweather, M. P. Foster, and D. A. Stone, “Battery parameter identification with Pseudo Random Binary Sequence excitation (PRBS),” Journal of Power Sources, vol. 196, no. 22, pp. 9398–9406, 2011.

[14] P. Manganiello, G. Petrone, M. Giannattasio, E. Monmasson, and G. Spagnuolo, “FPGA implementation of the EIS technique for the on-line diagnosis of fuel-cell systems,” IEEE International Symposium on Industrial Electronics, 2017.

Jukka Koskinen*, Timo Salmi, Pekka Kilpeläinen ja Pertti Lahdenperä

Robotiikan mahdollisuudet rakentamisessa

Tiivistelmä: Tässä paperissa esitetään robotiikan mahdollisuuksia suomalaisessa talonrakentamisessa ja taloelementtiteollisuudessa. Talonrakentamisen mahdollisuuksia selvitettiin analysoimalla rakentamisen työlajeja käymällä läpi rakentamisen työvaiheita Ratu-kortiston avulla. Elementtiteollisuudessa potentiaalisimpia automatisointikohteita etsittiin haastattelemalla ja vierailemalla yrityksissä. Muutamalle potentiaalisimmalle sovelluskohteelle tehtiin analyysi, jonka avulla selvitettiin taloudellisia edellytyksiä automatisoinnille.

Avainsanat: robotiikka, automatisointi, rakentaminen

*Yhteyshenkilö: Teknologian Tutkimuskeskus VTT Oy, Email: Jukka.Koskinen@vtt.fi

1 Johdanto

Rakentamisessa tuottavuuden kasvu on ollut heikompaa kuin esimerkiksi tehdasteollisuudessa, jossa on tapahtunut huomattavasti isompi tuottavuuden kasvu tuotantoprosessien digitalisoinnilla, robotisoinnilla ja kehittyneillä työstökoneilla [1]. Varsinainen rakentamistyö tehdään yhä pääosin perinteisin menetelmin, tosin digitalisoituja suunnitelmia hyödynnetään. BIM-tietomalleja (Building Information Model) rakennuksen tietomalli) hyödynnetään laajasti rakennusten suunnittelussa ja toteutuksessa. Mallit sisältävät mm. rakennuksen geometriaa koskevat tiedot digitaalisina, esimerkiksi rakennuksen tila- ja rakennemalleja

Automatisoituja koneita ja laitteita rakennustyömailla ei ole juurikaan voitu hyödyntää samalla tavalla kuin esimerkiksi konepajateollisuudessa. Automatisoinnin esteinä ovat mm. rakennusten yksilöllisyys, lyhyet työvaiheet, isot materiaalivirrat, matala käyttöaste yksittäisellä laitteella, epätarkat suunnitelmat automatisoinnin kannalta ja ahtaat tilat laitteiden siirtämiseen.

Rakentamisessa on kuitenkin paljon matalaa osaamista vaativia ja työolosuhteiltaan haastavia tehtäviä, kuten esim. poraus, hionta, vanhan materiaalin poisto, tavaroiden kuljetus. Automatisointi tulisi toteuttaa tehtävissä, joissa automatisoinnin edut olisivat sinänsä ilmeiset.

Rakentamisessa tarvittavien komponenttien, kuten esimerkiksi katto- ja seinäelementtien, valmistamisessa automaattisia laitteita käytetään jonkin verran ja myös robotteja hyödynnetään. Rakentamisen robotisoinnin kannalta haasteet ovat hyvinkin erilaiset eri rakentamisvaiheissa (perustus-, runko- ja ulkotyöt, sisärakenteet, LVIS-työt). Lisäksi erilaiset rakennustavat asettavat omat haasteensa (puu- ja betonielementtitalot, tiilitalot). Rakentamisen perinteet ja rakentamismääräykset ovat eri maissa hyvinkin paljon toisistaan poikkeavia, mikä on omalta osaltaan hankaloittanut rakentamisen menetelmien ja laitteiden kehittämistyötä. Monissa maissa edullinen työvoima on rakennuttajalle halvempi ja joustavampi vaihtoehto kuin automatisoidut laitteet.

Rakentamisen robotiikkaa on 1980-luvulta lähtien tutkittu ja erilaisia ratkaisuja on esitelty erityisesti Japanissa [2]. Elementtiteollisuutta lukuun ottamatta kaupallisiksi tuotteiksi asti kehitettyjä automaattisia laitteita tai järjestelmiä on markkinoilla vähän ja varsinaista läpilyöntiä eli laajamittaista käyttöönottoa ei ole minkään sovelluksen osalta vielä tapahtunut [3]. Haaste onkin löytää robotiikalle ne käyttökohteet, jotka mahdollistavat uuden liiketoiminnan luomisen ja kehittämisen. Rakentamisalalla on nähtävissä tilaa uusille robotiikkaa ja automaatiota hyödyntäville innovaatioille.

Automatisoinnin kannalta BIM-tietomalien hyödyntäminen on lähes välttämätöntä. Pääsy malleihin antaa mahdollisuuden viimeisimpiin muutoksiin suunnittelutiedoissa ja siten robotin tehtävien suunnittelussa [4]. Tosin tällä hetkellä BIM-malleja ei voida vielä kovin hyvin hyödyntää robotin tehtävien ohjaukseen. BIM-mallien sisältämä tieto ei ole välttämättä automatisoinnin kannalta riittävän tarkkaa tai se voi olla muutoin puutteellista.

Tämän paperin kappaleessa 2 on esitetty analyysi talonrakentamiseen liittyvien työtehtävien automatisoinnin mahdollisuuksista. Analyysissä hyödynnettiin Ratu-kortistoa [5], joka on rakennusalan tietopalvelu ja se sisältää mm. työmenetelmä- ja työohjekuvaukset rakennustyömaille sekä näihin menetelmiin liittyvää ositeltua työmenekkitietoa. Analysoimalla työohjeet tunnistettiin muutamia potentiaalisia robotisointikohteita, jotka otettiin tarkempaan tarkasteluun (Kappale 4). Valituille

työtehtäville määriteltiin karkeat ratkaisumallit ja laadittiin niitä vastaavat kustannuslaskelmat, joiden avulla arvioitiin robotisoinnin kannattavuutta kyseisissä työlajeissa (Kappale 5). Betonielementtitehtailla ja seinäelementtitehtailla (puiset ulkoseinät) potentiaalisimpia automatisoitavia tehtäviä selvitettiin tehtailla tehtyjen haastattelujen perusteella (Kappale 3). Näkökulma on suomalaisessa talonrakentamisessa ja taloelementtiteollisuudessa.

2 Potentiaalisia kohteita talonrakentamisessa

Tässä osiossa tarkastellaan työmaalla tapahtuvan rakennustyön automatisoinnin edellytyksiä. Kyseessä on pääpiirteinen, eräänlainen esikarsinta potentiaalisten automatisointikohteiden tunnistamiseksi. Työ tukeutuu Rakennustietosäätiön Ratu-kortistoon, sen Talo-Ratu -korttien läpikäyntiin. Korjaus-Ratu-kortistoa tai taloteknisiä töitä ei tarkastella. Työlajeja on pyritty arvioimaan seuraavien toteutettavuuskriteerien näkökulmasta:

- **Työvaltaisuus**. Kokonaisuudessaan suuri työmaatyön määrä ja työt liittyvät rakennusosiin, joiden kehittäminen esivalmistusta lisäämällä haastavaa.
- **Toistuvuus**. Yksittäinen työtehtävä, liikesarja tai liikesarjojen yhdistelmä toistuu työssä ja hankkeessa (ja markkinoilla) samanlaisena hyvin suurina määrinä.
- **Jatkuvuus**. Työtehtävät toistuvat jatkuvina samalla, yhtenäisellä työalueelle ilman oleellisia/vaikeita sijaintimuutoksia, monia keskeytyksiä tai välivaiheen tehtäviä.
- **Keveys**. Käsiteltävät aineet, tarvikkeet ja perinteiset työkalut ovat suhteellisen kevyitä (mahdollistamaan kevyet/edulliset ja helposti siirrettävät koneet).
- **Käsiteltävyys**. Käsiteltävät tarvikkeet ja tuotettava rakenne ovat koostumukseltaan ja mittasuhteiltaan säännönmukaisia ja helposti hallittavia (olosuhteet huomioon ottaen).
- **Työergonomia**. Haasteellinen ergonomia.
- **Laskennallisuus**. Muutoin suoraviivaisen toteutuksen haastava tuotesuunnitelma ja sen tulkinnan monimutkaisuus (mosaiikkikuviot, erityiset 3D-rakenteet).
- **Säännönmukaisuus**. Työ ei sisällä monia erilaisia hankalasti tulkittavia muuttujia

Eri kriteerien täyttyminen esitetään taulukossa 1 useimpien eri työlajien (pl. 1 Maarakennustyöt) osalta suuntaa-antavasti. Näin mm. siksi, että monissa tapauksissa kunkin työlajin sisällä tehdään hyvin erityyppisiä rakenteita, joiden osalta automatisoinnin edellytykset poikkeavat toisistaan jo lähtökohtaisesti. Näin ollen esitetyt arviot eivät voi olla yksikäsitteisiä, mutta niiden uskotaan auttavan huomion kohdistamisessa potentiaalisimpiin kohteisiin.

Analyysin perusteella, erityisesti muurauksessa, kivitöissä, raudoituksessa ja laatoituksessa on edellytyksiä täyttäviä ehtoja. Näistä muuraus ja laatoitus valittiin mukaan taloudellisuustarkasteluun. Maalauksen robotisointiin on tehty maailmalla paljonkin tutkimusta, joten myös se arvioitiin potentiaaliseksi ja otettiin mukaan tarkasteluun. Taloudellisuustarkastelu on esitetty kappaleessa 4.

Taulukko 1. Talonrakentamisen työmaatehtävien automatisoinnin mahdollisuuksien alustavaa arviointia. Taulukossa arvioidaan suuntaa-antavasti työlajien automatisoinnin edellytysten täyttymistä värikoodein (˜ (ei täytä ehtoa), › , › , › , ˜ (täyttää ehdon)).

Korttinumero Kortin nimi ja sisältö	Työvaltaisuus	Toistuvuus	Jatkuvuus	Keveys	Käsiteltävyys	Työergonomia	Laskennallisuus	Säännönmukai-
2 Betonirakennetyöt								
Ratu 0397 Lautamuottityö.	›	›	›	›	›	›		›
Ratu 0398 Levymuottityö.	˜	›	›	›	›	›		›
Ratu 0399 Kasetti- ja kupumuottityö.	˜	›	›	›	›	›		›
Ratu 0400 Pöytä- ja kulmamuottityö.	›	›	˜	˜	˜			›
Ratu 0401 Suur- ja erikoissuurmuottityö.	›	›	˜	˜	˜			›
Ratu 0402 Raudoitus.	›	›	›	›	›	›	›	›
Ratu 0403 Betonointi.		›	›	˜	›		˜	›
Ratu 0404 Pintabetonityöt.	›		›	›			›	
Ratu 0405 Lattiatasoitetyö.	›		›	›				
Ratu 0390 Kuorilaattaelementti- ja liittolevytyö.		›		˜	˜	›		
Ratu 0408 Betonipintojen etuoikaisu ja ruiskubetonointi.		›	›					›
Ratu 0406 Piikkaus, paikkaus, timanttiporaus ja -sahaus.		›	›			›		
3 Metallirakennetyöt								
31 Teräsrunkotyö		›		›	›			
Ratu 0410 Metalliovi- ja -ikkunatyö.		›	›	›	›			›
Ratu 0411 Metallielementtityö.	›	›		›	›			
Ratu 0412		›	›		˜	›		

Ohutlevytyö, kate.								
Ratu 0413 Ohutlevytyö, julkisivut ja täydentävät rakenteet.		›	›		~			›
Ratu 0414 Metallirakennetyö.		›	›	›	›			
4 Muuraus- ja kivityöt								
Ratu 41-0289 Tiilimuuraus.	›	~	~	›	›	›	›	›
Ratu 42-0290 Harkkomuuraus.		›	›	›	›	›		
Ratu 42-0291 Ohutsaumamuuraus.	›	›	~	›	›	›		›
Ratu 43-0292 Kivityö.	›	›	›	›	›	›	›	›
Ratu 44-0293 Ladonta. Tiilikate.	›	›	›	›	~	›		
5 Puutyöt								
Ratu 0416 Puurunkorakentaminen, paikalla rakennettu puurunko.				›	~			
Ratu 0417 Puurunkorakentaminen, platform-menetelmä.				›	›			
Ratu 0418 Puurunkorakentaminen, ulkopuolinen puuverhous.	›	›	›	›	›	›		›
Ratu 0424 Puuelementtirakentaminen, seinät.				~	~			
Ratu 0425 Puuelementtirakentaminen, tilaelementit.				~	~			
Ratu 0434 Saunan puutyöt.	›	›	›	›	›	›		›
Ratu 0435 Puuelementtirakentaminen, pilarit ja palkit.				~	~			
Ratu 0436 Puuelementtirakentaminen, väli- ja yläpohjaelementit.				~	~			
Ratu 0426 Levyrakentaminen, väliseinät.	›	›	›	›	›	›		›
Ratu 0420 Levyrakentaminen, levytys.	›	›	›	›	›	›		›
Ratu 0427 Puupintarakentaminen, sisäpuolinen puuverhous ja -päällystys.			›	›	›			›
Ratu 0428 Listoitus.			›	~	›			
Ratu 0429 Vakiovarustaminen, heloitus ja lukitus.			›	›	›			
Ratu 0422 Kalustepuusepäntyö.			›	›	›			
6 Eristys- ja saumaustyöt								
Ratu 0437 Lämmöneristys.				›	›			›
Ratu 0438 Ääneneristys.				›	›			›
Ratu 0430 Perustusten vedeneristys.		›		›	›			
Ratu 0431 Vesikaton vedeneristys.			›		~			
Ratu 0433 Sisäpuolinen vedeneristys.			›		~			›
Ratu 0432 Saumaus.					~			›
Ratu 0439			›		›			
Palosuojaus.								
7 Pintatyöt								
Ratu 71-0307 Rappaus.	›		›	›	›			
Ratu 72-0308 Tasoitetyö.	›		›		›			›
Ratu 73-0309 Sisämaalaus.		~	~	~	›			›
Ratu 73-0310 Tapetointi.					~			
Ratu 73-0311 Ulkomaalaus.					›			
Ratu 74-0312 Laatoitus.	›	~	›	~	›	~	›	
Ratu 75-0313 Mattotyö, kuivat tilat.					~			
Ratu 75-0314 Mattotyö, märkätilat.					~			
Ratu 76-0315 Massapäällystystyö.		›	›	›	›			›
Ratu 77-0316 Parketti- ja laminaattipäällystetyö.		›	›	›	›			›
Ratu 78-0317 Alakattotyö.		›	›		›	›		
Ratu 79-0318 Lasitus.		›	›		›			

3 Potentiaalisia kohteita puu- ja betonielementtitehtailla

Puuelementtitehtaat

Rakennustyömaihin verrattuna automatisoinnin lähtökohdat puutaloelementtien valmistuksessa tehtaassa ovat paljon helpompia. Samat tehtävät ovat kuitenkin kyseessä. Järjestely vain on toinen. Seinäelementtilinjoilla on automatisoinnin kannalta aikakin seuraavia etuja:

- Työ tapahtuu linjassa, elementti tulee robotin luo sen sijaan, että robotin pitäisi tulla työmaalle.
- Ulottuvuusvaateet ovat rajallisempia. Vaikka perinteisen teollisuusrobotin ulottuvuus ei sellaisenaan riitäkään, tehdasolosuhteissa ulottuvuutta on helpompi kasvattaa ulkoisilla akseleilla (portaalit ja lineaariakselit).
- Jatkuva materiaalivirta robotille järjestettävissä osittain linjan avulla. Laitetta ei tarvitse siirtää työmaalla eikä työmaalta toiselle. Käyttösuhde on helpompi saada kuntoon. Sitä tukee myös useammassa vuorossa työskentely.

Seinäelementtien valmistukseen kuuluu rungon kasausta, eristeiden laittoa, tuulensuojalevytystä, höyrynsulkumuovin asennusta, sisäpinnan levytystä, ulkolaudoitusta, ikkunoiden asennusta ja ikkunoiden ympärillä olevien vuorilautojen asennusta. Työ sisältää paikalleen asennusta, naulausta, niittausta ja sahausta. Li-

säksi tehtaissa tehdään erinäisiä aidakkeita yms. Periaatteessa suuri osa tehtävistä on automatisoitavissa. Toki se vaatii tutkimus- ja kehitystyötä. Jotkut kohteet enemmän ja toiset vähemmän. Suomalaisissa tehtaissa ei ole yleensä niin isoa volyymia, että robotti voisi olla vain yhteen tehtävään dedikoitunut. Myös joustavuusvaateet ovat kovat. Seinäelementtivalmistuksessa on olemassa kaksi merkittävää kansainvälistä laitetoimittajaa, saksalainen Weinnman ja ruotsalainen Randek, joiden ratkaisut ovat portaaleihin perustuvia. Suomalainen Kivioja Engineering on myös tullut mukaan alueelle. Vakiintuneita ratkaisuja on suomalaisten kokemusten mukaan pidetty turhan jäykkinä. Laitteiden hyödyntämisessä on erilaisia käyttötapoja täysautomaatiota ja erilaisia automaation ja manuaalisen työn sekoituksia. Haastavaa on löytää sopiva automaatiotaso, joustavuus ja kustannustehokkuus kohdilleen.

Julkisivujen laudoitus on automatisoinnin kannalta vaativimpia tehtäviä. Varsinainen laudoitus on kohtuullista adaptiivisuutta vaativaa yleensäkin, kun materiaali on joustavaa ja laudat pitäisi saada ponttiin luotettavasti. Lisäksi tulisi mukautua sekä mitoissa että muodoissa oleviin laatupoikkeamiin. Lisäksi pitäisi tehdä pinnan tarkastus. Manuaalinen laudoitus on pullonkaula puuelementtitehdaslinjalla. Laudoituksen nopeuttaminen automatisoimalla nostaisi merkittävästi tehtaan kapasiteettia ja siten tuottavuutta.

Myös eristeiden automaattinen asennus on haastavaa. Miten saada villan leikkaaminen, käsittely ja asennus toimimaan luotettavasti. Julkisivujen maalauksen automaatio on myös kiinnostavaa. Maalin levityksen voi ajatella tapahtuvan ruiskulla, mutta haluttu laatu vaatii siveltimen käyttöä. Tarvittava maalimäärä riippuu kuitenkin pinnan laadusta, joka vaihtelee. Haasteena on saada tämä prosessi hallintaan, jotta laatu on kohdallaan. Laadunvalvonnassa konenäöllä voisi olla oma rooli, mutta tilanne ei oletettavasti pelkästään sitä kautta ratkea.

Tilaelementtien valmistuksessa mielekäs lähtötilanne on aloittaa seinäelementeistä. Siitä mennään valmiusasteessa pidemmälle. Silloin tulee lisäksi vastaan muita kiinnostavia ratkaisuja, kuten laatoitus, jota käsitellään työmaatöiden yhteydessä. Mahdollisesti myös vedeneristyksen levitys ja sisämaalaukset ovat automatisoitavissa. Myös muille vaiheille automatisoinnin lähtökohta tilaelementtien valmistuksessa on työmaata helpompi. Vastaan tulee myös se kysymys, voisiko jotain LVIS-töihin liittyviä esiasennuksia tehdä automaattisesti.

Betonielementtitehtaat

Betonielementti tehtaiden osalta arvioidaan sitä, mitä mahdollisuuksia olisi automatisoida seinäelementtien (sandwich-tyyppinen) ja ontelolaattojen valmistusta esimerkiksi robottien avulla. Isoin hyöty olisi saatavissa automatisoimalla eristeiden asennus sandwich-elementtiin. Betonielementtitehtailla on jo portaalirobotteja esimerkiksi betonielementin muotien reunojen kasaamiseen sekä betoniraudoitteen kokoamiseen

Villaeristelevyjen asennus elementtiin tehdään manuaalisesti. Työntekijät asentavat eristelevyjä muottiin valetun betonikerroksen päälle eri kokoisina levyinä ja tarvittaessa leikkaavat levyt pienemmiksi paloiksi. Vaihtoehtona on asentaa ne robotilla, joka poimisi eristeet ja asentaisi ne elementtiin. Haasteena robotiikan kannalta on eri kokoisten palojen leikkaaminen sekä eristeiden tarkka paikoitus oikeaan kohtaan. Paikoituksen voisi tehdä tietomallien (BIM) ja konenäön avulla. Tosin BIM-mallit eivät tällä hetkellä tue tätä, koska mallit eivät sisällä tietoa siitä miten eriste tulisi koota erikokoiset eristelevyistä; eristekerros on mallissa yhtenäisenä isona kappaleena. Haasteena on myös eristelevyjen asennus elementtiin riittävän tahtiajan puitteissa. Muottiin valettavan sandwich-elementin ylä- ja alaosat kiinnitetään toisiinsa ansaiten avulla, jotka työnnetään eristelevyjen asennuksen yhteydessä eristelevyn vierestä tai läpi siten, että ansain kiinnittyy elementin alaosaan ja ansaimen toinen pää jää eristelevyn yläpuolelle valettavaan betonikerrokseen. Tämän voi tehdä myös robotilla, toisaalta pistokastyyppisten ansaiten asennukseen on jo olemassa automaattisia ratkaisuja maailmalla.

4 Valittujen työlajien automatisoinnin taloudellisen toteutettavuuden arviointia rakennuksilla

Kappaleessa 2 esitettiin rakentamisen työlajien automatisointimahdollisuuksia yleisesti. Tässä osiossa käydään läpi valittuja työlajeja ja niiden automatisointiedellytyksiä tuotannollistaloudellisesta näkökulmasta. Kohteena ovat rakennustyömaalla tehtävät työlajit seuraavasti:

- Maalaus
- Tiilimuuraus
- Lasitus
- Laatoitus

Taloudellisuusarviot tehtiin kaikille neljälle valitulle me-

netelmälle ja niiden tuloksia käsitellään jäljempänä. Esimerkkinä tarkastelutavasta esitetään tarkemmin maalaus. Esimerkissä on tarkoitus tuoda keskusteltavaksi arvioinnissa käytettävät ratkaisuolettamat esityksen ollessa luonteeltaan alustava. Vasta kun laskelman rakenteesta ja muuttujien arvoista on olemassa hyvä näkemys, voidaan taloudellisuusarviota kehittää päätöksenteossa käytettäväksi.

Esimerkki esitetään vaiheittain rakennettuna taulukossa 2, joissa värillisellä kirjaimella viitataan yleensä aiempaan riviin, jossa kulloinkin käytetty lähtötieto esitetään perusteltuna/laskettuna. Vaihtoehtoisesti viitataan kyseisen rivin vasemmalla olevaan sarakkeeseen (”vas.”).

Yleiskuva maalauksesta:

- Automatisoinnin kohteena on laajojen ja tasaisten pintojen, kuten sisäkattojen ja -seinien maalaus rakennustyömaalla.
- Laitteisto perustuu telalla maalaukseen ja maalinsyöttö tapahtuu ruiskuttamalla maali maalattavalle pinnalle tai syöttämällä maalia telalle.
- Laitteen oletetaan toimivan luotettavasti ja itsenäisesti siinä rajatussa tehtävässä, mihin se on tarkoitettu.

Työn suoritus ja rajaus

- Maalauskohteen ympäristön ja siihen liittyvien rakenteiden suojaaminen jäävät työntekijöiden tehtäviksi.
- Laitteisto tekee vain varsinaista maalaustyötä, ei tasoitusta, hiontaa tai pölynpoistoa.
- Laitteiston valvonta ei edellytä erityistä panosta, vaan työntekijät hoitavat sen oman maalaustyön ohessa.
- Laitteen siirrot maalattavien alueiden välillä ja laitteen ohjelmoinnin tekevät maalaustyötä tekevät työntekijät.
- Kulmat, reuna-alueet ja muut maalauksen kannalta vaikeat alueet jäävät työntekijöiden tehtäviksi.

Taloudellisuus

- Laskelmat vihjaavat, että automaatiojärjestelmän voisi olla yrityksen näkökulmasta kannattavaa, jos laite maksaisi alle 120 k€ ja sitä käytettäisiin yhdessä työvuorossa hyvällä käyttöasteella viisi vuotta. Käyttö kahdessa vuorossa tekisi automatisoinnista luonnollisesti kannattavampaa, mutta sellainen ei rakennustyömaaolosuhteissa lienee kovin todennäköistä – tehdasolosuhteissa tämä lienee todennäköisempää.
- Maalauslaitteelle on periaatteessa varsin laajat markkinat, koska käyttötarkoitusta vastaavia töitä esiintyy kaikissa rakennuksissa, eikä rakennustyyppi tai siinä käytettävät rakenteet rajaa sovellusmahdollisuuksia kohtuuttomasti.

Taulukko 2. Sisämaalauksen automatisoinnin taloudellisuus.

Rivi	Laskelmaoletus	Laskelma
	Muutokset työsaavutuksissa	
A	Tarkasteluyksikkönä on oletettu työntekijän/maalarin tehollisen työajan työsaavutus aikayksikköä kohti.	”100 %” kuvaa yhden työntekijän tehollista työsaavutusta yhtä aikayksikköä kohti (teoreettinen). Kohtien B–E, H ja I arvot lasketaan suhteessa tähän.
B	Ihmisen toiminnassa käytetystä kokonaisajasta vain osa on tehollista työtä (tehollisen menetelmäajan ja ns. työvuoroajan suhde). Tässä osuus on **80 %**.	Työvuoroaikaa vastaavan työsaavutuksen osuus vastaavan aikajakson tehollisesta työsaavutuksesta: 80 % (vas.) * 100 % (A) = 80 %
C	Robotin tehollinen, menetelmäajasta laskettu työsaavutus on hieman työtekijän tehollista työsaavutusta korkeampi, esim. **1,2** kertainen.	Robotin tehollisen työn työsaavutus suhteessa työntekijän tehollisen työajan työsaavutukseen: 1,2 * 100 % (A) = 120 %
D	Robotin ohjelmointiin yms. käytetty aika vähentää (yhden) työntekijän panosta **30 %**.	Robottia käyttävän työntekijän todellinen maalaustyösaavutus suhteessa pelkkään maalaukseen keskittyvän työntekijän tehollisen ajan työsaavutukseen: (1 – 30 % (vas.)) * 80 % (B) = 56 %
E	Robotti on käytettävissä tehollisen työhön vain noin **70 %** kokonaisajasta.	Robotin työvuorotason työsaavutus suhteessa työntekijän tehollisen työajan työsaavutukseen: 70 % (D) * 120 % = 84 %
F	Robotti tekee maksimissaan tietyn osuuden varsinaisesta maalaustyön kokonaistyöpanoksesta.	Osuus on tässä 40 %.
G	Tuottavuusedun hyödyntäminen edellyttää robotin mahdollisimman keskeytymätöntä käyttöä. Robotin käyttö on mielekästä lähinnä osana kahden työntekijän ja robotin työryhmää.	Robotin suhteellinen työsaavutus työvuoroaikana on 84 % (E). Robottia ohjaavan työntekijän vastaava työsaavutus on 56 % (D). Tällöin robotin suhteellinen työsaavutus 84 % on yli 40 % (F) robotin ja yhden työntekijän suhteellisesta työsaavutuksesta: 84 % / (84 % + 56 %) = 60 % > 40 %

H	Kahden työntekijän vertailutyösaavutus.	2 tt * 80 % (B) = 160 %
I	Kahden työntekijän ja robotin työsaavutus	1 tt * 80 % (B) + 1 tt * 56 % (D) + 84 % (E) = 220 %
J	Työsaavutuksen paraneminen.	(220 % (I) / 160 % (H)) – 100% = 37,5 %
	Kustannussäästöt	
K	Laskennallinen vuotuinen kokonaistyöaika on 2 000 h, kun vuodessa on n. **250 työ/arkipäivää** (työvuoroa, tv) ja työvuoron pituus on **8 tuntia** (h).	250 tv (vas.) * 8 h/tv (vas.) = 2 000 h
L	Työntekijäkustannus tuntia kohti.	35 €/h
M	Laskennallisen kahden työntekijän työryhmän vuositason henkilökustannus.	2 tt * 2 000 h (K) * 35 €/tth (L) = 140 000 €
N	Ilman robottia toimivan kahden työtekijän työryhmän henkilökustannus sellaista vuosityösaavutusta kohti, jonka robotin kanssa toimiva kahden hengen työryhmä saa aikaan.	140 000 € (M) * (1 + 37,5 % (J)) = 192 500 €
O	Työntekijäkustannusten teoreettinen vuosisäästö	192 500 € (N) – 140 000 € (M) = 52 500 €
P	Robotin vuosittainen käyttöaste	80 %
Q	Laitteiston vuotuiset lisäkustannukset	10 000 €
R	Työ- ja käyttökustannusten vuosisäästö	80 % (P) * 52 500 € (O) – 10 000 € (Q) = 32 000 €
	Investoinnin kannattavuus (esimerkki)	
S	Investoinnin tuottovaatimus (laskentakorko)	10 %
T	Takaisinmaksuaika	5 v
U	Robotin hankintakustannuksen yläraja	32 000 €(R) * [1-1/(1+10 %(S))^5 (T)]/10 %(S) = 121 305 €

Edellä arvioitiin maalauksessa muodostuvia työ- ja käyttökustannusten vuositason säästöjä, mikäli kyseisissä töissä hyödynnetään automaatiota/ robotiikkaa. Samalla näistä säästöistä laskettiin esimerkinomaisesti yksi raja-arvo robotin hankinta-kustannukseksi, jolla investointi voisi vielä olla kannattava. Laskelmat perustuivat tiettyyn, yhteen laskentakorkoon/tuottovaatimukseen ja takaisin-maksuaikaan. Nämä arvot eivät ole ehdottomia vaan kyse on enemminkin päätöksentekijän valinnoista.

Oheisessa taulukossa 3 hankintakustannustarkastelua laajennetaan tehtäväksi eri takaisinmaksuajoilla ja tuottovaatimuksilla, kun tuottona on taulukossa aiemmin laskettu vuosittainen säästösumma (laskentalogiikan noudattaessa niin ikään edellä esitettyä). Taulukossa esitetään vastaavat tulokset myös muille kolmelle työssä tarkastelluille menetelmille, vaikka niiden tarkempi käsittely tässä muuten ohitetaan

Taulukossa 2 esitetiin laskentatapa, kun robottia käytetään yhdessä työvuorossa. Robotilla on mahdollista tehdä myös kahta työvuoroa. Tällöin vuosittainen tuotto/säästö olisi kaksinkertainen taulukon lukuihin nähden, sillä sekä työkustannuksissa saatavat säästöt, että laitteiston vuotuiset käyttökustannukset kaksinkertaistuvat. Tämä sallisi oleellisesti suuremmat laiteinvestoinnit.

Taulukko 3. Robotin hankintakustannuksen yläraja [€] kannattavalle laiteinvestoinnille.

Tuottovaatimus	5 %		10 %	
Takaisinmaksuaika	3 vuotta	5 vuotta	3 vuotta	5 vuotta
Maalaus	87000	139000	80 000	121000
Tiilimuuraus	116000	184000	106 000	161000
Lasitus	55000	8 000	50 000	76000
Laatoitus	179000	284000	163 000	249000

5 Yhteenveto ja pohdinta

Rakentamisessa teknisesti moni kohde on automatisoitavissa, mutta toimiva toteutus vaatii isoja kehityspanoksia. Käytännössä tulisi kehittää robotiikkaan perustuvia ratkaisuja, joita markkinoitaisiin kansainvälisesti. Se vaatisi riskisijoituksia. Laskelmiemme perusteella robotilla saavutettava kustannussäästö vaatii korkeaa käyttöastetta, joka voi olla vaikeasti toteutettavissa, mikäli robotti voi tehdä vain yhtä työtä/työvuoroa. Hyvä käyttösuhde on vaikea toteuttaa työmaaoloissa. Liikutteleminen ja työn aloittaminen aina uudessa tilanteessa vie resursseja. Robotin sovittaminen työmaiden aikatauluihin tuo myös omat haasteensa. Robotin monikäyttöisyys voisi parantaa tilannetta. Robotti vaatii kuitenkin rakentamisessa todennäköisesti aina ihmisen avustamaan työtä tai valvomaan robotin toimintaa, joka osaltaan vähentää kiinnostusta robottien hyödyntämiseen. Rakentamisessa onkin siirrytty ja tullaan myös jatkossa siirtämän töitä yhä enemmän esivalmistukseen, jossa automatisoinnilla on jo paremmat edellytykset.

Viitteet

[1] Lohilahti, O. Rakennusalalla työn tuottavuus ei ole

kasvanut 40 vuodessa – onko allianssista tai leanista apua?, Rakennuslehti, 4.7.2017.
[2] Council for Construction Robot, March 1999, Construction Robot System Catalog in Japan, Japan.
[3] Vähä, P., Heikkilä, T., Kilpeläinen, P., Järviluoma, M., & Heikkilä, R., Survey on Automation of the Building Construction and Building Products Industry, VTT technology 109, VTT, 2013.
[4] Vähä, P., Heikkilä, T., Kilpeläinen, P., Järviluoma, M., & Gambao, E., Extending automation of building construction: Survey on potential sensor technologies and robotic applications. Automation in Construction, 2013, 36 168-178.
[5] Rakennustietosäätiö, 2019, Ratu-kortisto (päivittyvä), Rakennustieto, Helsinki.

Igor Trotskii and Jukka Pulkkinen

Unsupervised machine learning model for heat flow monitoring in a geothermal energy storage in a near-zero-energy-building

Abstract: With a fast-paced development in IoT, information processing and process monitoring techniques, building automation systems become more and more complex and advanced. Evolution of these technologies allows to greatly improve buildings' energy performance and substantially decrease maintenance costs by utilizing such ideas as Condition Based Maintenance (CBM) and Machine Learning, which CBM heavily relies on.

As most of the building are still maintained through normal means, such as reactive and schedule-based maintenance, because of lack of property managers' interest in investing in more efficient approaches, this paper's aim is to cover this gap by providing short overview of most basic machine learning and data processing algorithms used in building maintenance domain and by providing case study results. The case study was conducted in a near-zero-energy building, Sheet Metal Center in Visamäki, Hämeenlinna by constructing Principal Component Analysis based solution for monitoring energy flows inside geothermal energy storage for further usage in CBM. The solution helps to quickly evaluate the state of the system and allows to simplify diagnosis and faults localization.

Keywords: Condition Based Maintenance, Principal Component Analysis, near-zero-energy building

Igor Trotskii: HAMK, E-mail: igor.trotskii@hamk.fi

Jukka Pulkkinen: HAMK, E-mail: jukka.pulkkinen@hamk.fi

1 Introduction

CBM has been an interesting topic since long ago as it allows to greatly reduce downtime of the process without the need for frequent equipment checks and replacements as scheduled maintenance dictates. [19] Nevertheless, today maintenance is still done mainly based on traditional scheduled maintenance. [18] The main limitation of CBM adoption is a high initial cost for modern equipment [17], which has embedded self-diagnostic capabilities, therefore could report about its own condition. On the other side, Internet of Things has been developing rapidly, bringing the ease of obtaining enormous amount of data from monitored process without any big investment costs. This makes it possible to utilize measured data in conjunction with machine learning to build inexpensive CBM systems, avoiding costly investments in a new equipment by moving part of expenses to the software by monitoring and analyzing parameters which have indirect impact on the monitored equipment. [8] As mentioned above, CBM heavily relies on machine learning in order to forecast building and equipment performance to be able to evaluate best possible repairment schedule. The reason for that is the complexity of the data from modern automation systems. Even the simplest system can have dozens of different metrics and measurements, which have to be analyzed in order to estimate devices and equipment status and if they should be replaced or fixed. One way to solve that problem is by applying statistical methods and machine learning.

The main purpose of this study is to determine possible data analytics techniques to utilize in building maintenance domain. Even though, many algorithms for fault and anomaly detection are already present, it is still hard to select one for simplifying CBM implementation.

This paper provides a brief overview of available techniques, which are widely used in building maintenance domains and their possible applications. The usefulness of the CBM and machine learning approach for building maintenance is further proved by conducting a case study in Sheet Metal Center in Visamäki, Hämeenlinna, which belongs to near-zero-energy buildings class.

2 Literature review

Numerous techniques are known, which can be applied for improving building maintenance efficiency by helping to adopt key principals of CBM. Most of these

techniques belong to anomaly detection domain as the information about faults, abnormalities or their possibilities is of utmost importance for performing repairment in economical way. The list of the available algorithms includes, but is not limited to, Principal Component Analysis, Support Vector Machine and Neural Networks.

The problem encountered by the authors is the lack of research in building maintenance and especially in maintenance and monitoring of near-zero-energy building domains. Even though, amount of developed techniques related to anomaly detection and Condition Based Maintenance is high, their possible applications and implementation for nZEB is still under study.

2.1 Principal Component Analysis

Principal component analysis is one of the most widely used multivariate statistical techniques used for dimensionality reduction and anomaly detection. PCA is based on an orthogonal decomposition of original correlated measurements into new space of lower dimensionality, which consists of new uncorrelated variables called principal components. New variables are selected in a way to maximize explained variance, what allows to significantly reduce amount of measurements by preserving as much information from original data as possible. [21]

PCA is a well-known technique in a process monitoring domain and even in its simplest form can be very useful for automatic fault detection. For example researchers have used this approach in asynchronous generators [14], wastewater treatment plant [7] and reciprocating compressors. [1]

Anomaly detection applications based on PCA often use such metrics as SPE and Hotelling's T^2 metrics (Mujica, Rodellar, Fernández, & Güemes, 2011), which are relatively simple to calculate and give accurate estimation if situation anomalous or not. However, it can be very difficult to localize the source of the fault. Fortunately, various techniques were developed to handle this problem, e.g. contribution plots. [16]

2.2 Other techniques

Principal component analysis is not the only popular tool in process monitoring and fault detection. Two other popular techniques in building maintenance domain is support vector machine (SVM) and Artificial Neural Networks (ANN).

Support vector machine is a supervised learning algorithm mainly used for classification, however, it is possible to perform regression tasks as well. The main idea behind SVMs is transforming a dataset to a higher dimensional space in order to make initial complex non-linear relations into simple linear separable clusters, but of higher dimensionality. [3]

SVMs are widely used for building energy efficiency evaluation [4], predicting electrical energy consumption [5] and forecasting of cooling/heating load for HVAC systems. [13] These methods can be useful for improving building management or tuning building automation system, however, they are not directly relevant to maintenance efficiency. More direct approach is applying SVM for fault detection. [2, 9]

Artificial Neural Networks have already revolutionized many industries and so it is only natural that they are a powerful tool in maintenance domain. The benefit of utilizing ANNs is their ability to solve both classification and regression problems. Neural Networks are widely used for forecasting building energy consumption and for CBM directly. [6, 12]

ANNs provide superior accuracy of prediction or fault detection, but they usually require big amount of recorded measurements in order to achieve high efficiency and so their usefulness can be severely limited especially in new buildings.

3 Case Study: Sheet Metal Center

3.1 Background of the study

The new testing hall for HAMK Sheet Metal Center (SMC) was built in 2015. The building is near-zero-energy building based on different technologies such as compact envelope, energy saving windows, effective heat recovery in air handling units, building automation and renewable energy sources. The renewable energy consists of the solar and geothermal energy units. The geothermal part is the main heat supplier and it includes energy piles and heat wells. The solar energy units are used to fill energy piles with thermal energy.

Fig 1. Sheet Metal Center testing hall

The behavior of the heating and cooling system depends primarily on the season, hence it is possible to define two main operational modes: heating season and cooling season. The operation of the system during one of the modes can be described by thermal energy flows between different parts shown in the Fig. 2. During heating season both energy piles and heat well must provide energy to the heat pump as long as their outlet liquid temperatures are not bellow certain threshold in order to prevent the formation of ice. There is no heating demand in the system during the cooling season, so heat pump doesn't operate, and energy piles are separated into a separate loop to fill them with heat from solar collectors and air handling unit exhaust air via heat exchanger; heat well is used as a heat sink for the building.

Although, the state of the system is very clear and easy to monitor with normal means during summer or winter it is not the case during autumn, spring or maintenance when it can oscillate between two states, what can lead to different anomalies and faults. The designed monitoring method addresses this problem.

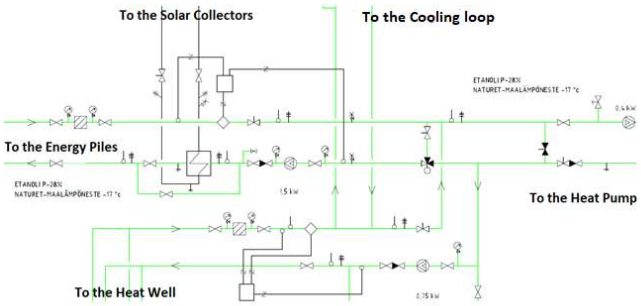


Fig. 2. Part of the heating system under the study

3.2 Data Selection

More than 130 process variables are monitored for the SMC building. Therefore, feature selection is required to select only relevant variable, otherwise the resulting model could be irrelevant to the area of interest, or worse, it is possible that no reasonable output could be derived. The feature selection for the designed model was based on physical locality principle, as a result, only measurements from the sensors close to the energy piles, heat well, heat pump and cooling loop connection were used. The final measurement list is shown in Table 1.

Measurement
Sun-heated glycol return temperature
Sun-heated glycol temperature at the top of the storage tank
Sun-heated glycol incoming temperature (meter data)
Sun-heated glycol return temperature (meter data)
Well pipes outcoming ethanol temperature
Well pipes incoming ethanol temperature
Well pipes ethanol flow rate
Well pipes incoming ethanol temperature (meter data)
Well pipes return ethanol temperature (meter data)
Well pipes outcoming ethanol temperature
Well pipes incoming ethanol temperature
Ethanol temperature after well pipes and roof radiant heaters
Ceiling heating outcoming ethanol temperature
Ceiling heating incoming ethanol temperature
Energy piles ethanol flow rate
Energy piles incoming ethanol temperature
Energy piles return ethanol temperature
Ethanol temperature before heat-exchanger with sun-heated glycol
Well pipes ethanol mixing valve

Table 1. List of the measurements used for analysis

3.3. Modelling technique selection

For this case study traditional principal component analysis was chosen instead of ANNs or SVM/SVR for following reasons:

1. Authors had access to limited amount of data: HAMK Sheet Metal Center is relatively new building and amount of available measurements is less than one year and a half, what limits the usefulness of ANNs.
2. There are gaps in available data due to various reasons, the ability to evaluate autocorrelation of the studied system. PCA considers every measurement point as a separate one, and thus its accuracy does not suffer because of the gaps.
3. SVM/SVR are very hard to use for process

monitoring and fault detection in case the faults themselves are not strictly defined as SVM/SVR are supervised learning algorithms.
4. PCA is well supported by such metrics as SPE and Hotelling's T2 statistics, which make fault detection a lot easier.

3.4. Modelling

The dimensionality of the selected dataset strongly suggests the use of some dimensionality reduction techniques such as principal component analysis (PCA). The main idea of PCA is transforming high dimensional datasets by projecting points onto a new lower dimensional space. The resulted uncorrelated variables – Principal components can be regarded as a linear combination of the original variables. PCA is very sensitive to the scaling of the variables as the method is based on calculation of the covariance matrix, hence it requires standardization of the data, so zero mean – unit variance scaling was applied. [10]

PCA is one of the best techniques for multivariate statistical analysis and it was chosen as it doesn't require much of prior knowledge about the process, which generated data. Another concern is inconsistency in data and missing values: traditional PCA does not consider autocorrelation, therefore it is not sensitive to gaps in training data unlike such methods as Dynamic PCA. [20]

3.5. Results

3.5.1. Interpretation and verification of principal components.

The resulting principal components were analyzed based on a correlation matrix for the principal components and original parameters. The results are shown in Table 2.

Principal Component	Explained variance	Meaning
1	64.14%	Heat flow into the energy piles and heat well
2	20.12%	Heat flow from the heat well into the heat pump and cooling system
3	6.26%	Random oscillations of the parameters
4	4.50%	

Table 2. Explained variance and meaning behind each principal component

The assumptions made in the Table 2 were further confirmed by applying k-means clustering to the data obtained from principal component analysis, the result of which is shown in the Fig 3. As it can be seen from the Table 2, most of the variance is explained by first two principal components (PCs) and the difference between them is essentially the heat flow direction, meaning that the simultaneous changes in both PCs can only be explained by changing the operational mode from heating to cooling or vice versa, hence the result of clustering should display the state of the heat pump: if it is on or off.

K-means clustering was performed on obtained four principal components, which cover three different operation periods:

1. From 23.10.2018 to 05.11.2018 — normal winter operation before heat pump maintenance.
2. From 05.11.2018 to 20.11.2018 — heat pump maintenance.
3. From 20.11.2018 to 01.02.2019 — normal winter operation after heat pump maintenance.

Labels obtained from clustering can be explained as follows:

- 0 – The heat pump is on during the third period.
- 1 – The heat pump is on during first period or off during the first period.
- 2 – The heat pump is under maintenance.

Class 1 brings some confusement as it can mean both on and off states of the heat pump depending on the period of time it belongs to. The reason for this is that obtained clusters are based on the heat flow between energy piles, heat well and heat pump and not heat pump production, and the flows changed a lot after the heat pump maintenance. However, heat pump production is strongly correlated to the heat flows in the system. In general the cluster can represent if energy flows into the ground, into the heat pump or if something anomalous, e.g. maintenance, happens. Oscillation of the clusters' labels are explained by the fact that the heat pump is currently used on/off controller and so it changes its state frequently. Accuracy metric was defined based on previously defined meaning of the labels and heat pump production level and essentially represent how accurately clustering is able to differentiate between on/off and maintenance states of the heat pump. The label was correct in 77.36% of cases.

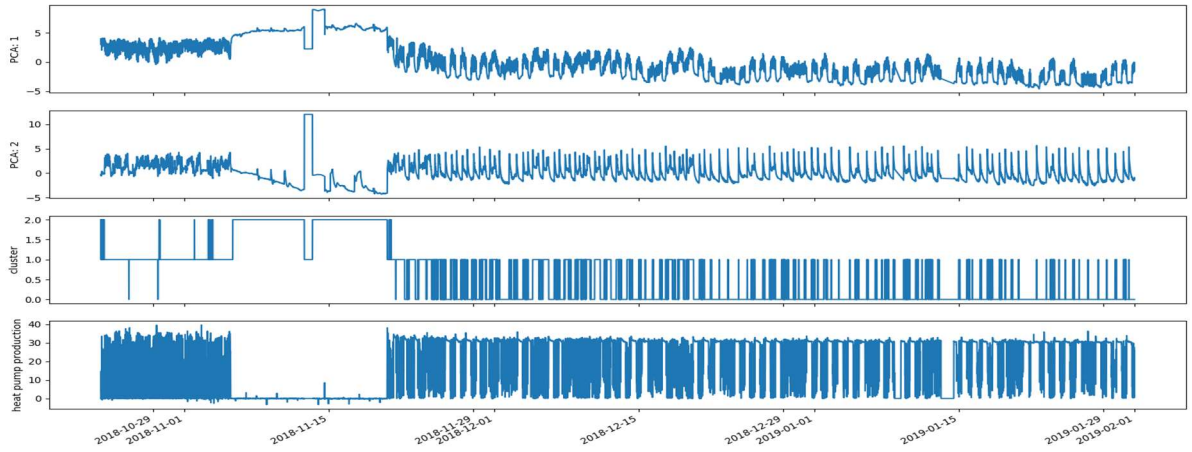


Fig 3. Clustering and comparison between clustering results and real measurements from the heat pump.

3.3.2. Fault detection with Hotelling's T^2 and SPE statistics.

Process monitoring and fault detection with principal component analysis is majorly based on T^2 and SPE metrics. Both these metrics can be used in order to find outliers in respect to original training set. Big SPE or T^2 value does not necessary mean that the system is in faulty state, however, if these values are abnormally high over extended periods of time then it might be worth some investigation.

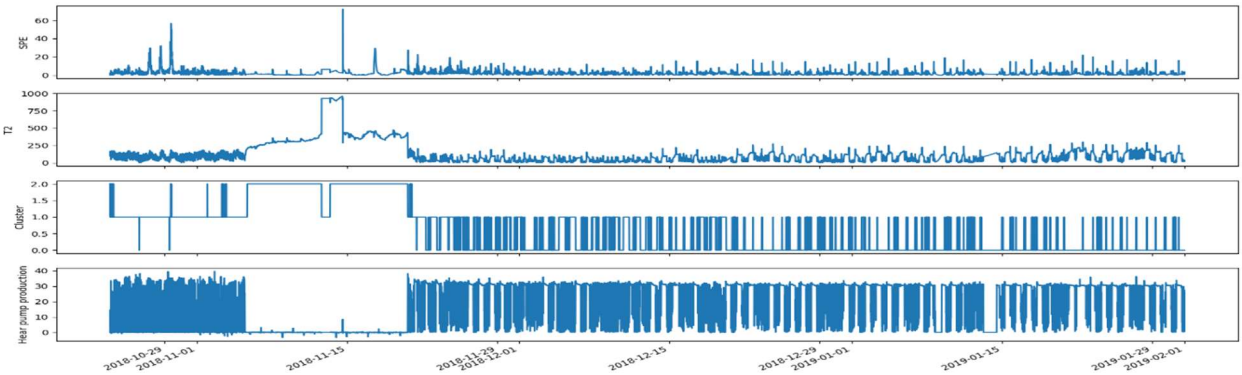


Fig 4. Hotelling's T^2 and SPE metrics.

Result of using T^2 and SPE metrics is shown on Fig 4. Most noticeable outliers happened during maintenance break. On 15.11.2018 heat pump was on for a very short period of time, when it was supposed to be off, what resulted in SPE spike. The second anomaly is covered by T^2: maintenance period itself can be considered as a deviation from the normal operation and thus it causes growth in the statistic. The squared spike is explained by missing measurements, which caused even further deviation from normal values.

4 Conclusion

In this paper a data analyzing method based on principal component analysis was developed for a geothermal energy storage, allowing to easily monitor

the current state of the geothermal energy storage in order to schedule maintenance efficiently by adapting the key principles of the Condition Based Maintenance without high investment costs. However, due to the indirect nature of the measurements used for analysis, the developed tool can only be used as a supplementary tool due to its inability to accurately show the exact faulty or degrading part as the model evaluates performance of the system as a whole. Nevertheless, the designed model produces reliable results which can be used to better pinpoint time of the fault occurring.

References

[1] Ahmed, M., Baqqar, M., Gu, F., & Ball, A. D. (2012). Fault detection and diagnosis using Principal Component Analysis of vibration data from a reciprocating compressor. In *Proceedings of 2012 UKACC International Conference on Control* (pp. 461–466). IEEE. https://doi.org/10.1109/CONTROL.2012.6334674

[2] Chitrakar, R., & Chuanhe, H. (2012). Anomaly detection using Support Vector Machine classification with k-Medoids clustering. In *2012 Third Asian Himalayas International Conference on Internet* (pp. 1–5). IEEE. https://doi.org/10.1109/AHICI.2012.6408446

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

[4] de Wilde, P., Martinez-Ortiz, C., Pearson, D., Beynon, I., Beck, M., & Barlow, N. (2013). Building simulation approaches for the training of automated data analysis tools in building energy management. *Advanced Engineering Informatics*, *27*(4), 457–465. https://doi.org/10.1016/J.AEI.2013.05.001

[5] Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, *37*(5), 545–553. https://doi.org/10.1016/J.ENBUILD.2004.09.009

[6] Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software*, *40*(5), 356–362. https://doi.org/10.1016/J.ADVENGSOFT.2008.05.003

[7] Garcia-Alvarez, D. (2009). FAULT DETECTION USING PRINCIPAL COMPONENT ANALYSIS (PCA) IN A WASTEWATER TREATMENT PLANT (WWTP).

[8] Ghasemi, A., Yacout, S., & Ouali, M.-S. (2010). Parameter Estimation Methods for Condition-Based Maintenance With Indirect Observations. *IEEE Transactions on Reliability*, *59*(2), 426–439. https://doi.org/10.1109/TR.2010.2048736

[9] Heller, K., Svore, K., Keromytis, A. D., & Stolfo, S. (2003). One Class Support Vector Machines for Detecting Anomalous Windows Registry Accesses, 2–9. https://doi.org/10.7916/D84B39Q0

[10] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

[11] Krenek, J., Kuca, K., Blazek, P., Krejcar, O., & Jun, D. (2016). Application of Artificial Neural Networks in Condition Based Predictive Maintenance (pp. 75–86). https://doi.org/10.1007/978-3-319-31277-4_7

[12] Kumar, R., Aggarwal, R. K., & Sharma, J. D. (2013). Energy analysis of a building using artificial neural network: A review. *Energy and Buildings*, *65*, 352–358. https://doi.org/10.1016/J.ENBUILD.2013.06.007

[13] Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. *Energy Conversion and Management*, *50*(1), 90–96. https://doi.org/10.1016/J.ENCONMAN.2008.08.033

[14] Martón, I., Sánchez, A., Carlos, S., & Martorell, S. (2013). Application of Data Driven Methods for Condition Monitoring Maintenance. In *Chemical Engineering Transactions* (Vol. 33, pp. 301–306). https://doi.org/10.3303/CET1333051

[15] Mnassri, B., Adel, E. M. El, Ananou, B., & Ouladsine, M. (2009). Fault Detection and Diagnosis Based on PCA and a New Contribution Plot. *IFAC Proceedings Volumes*, *42*(8), 834–839. https://doi.org/10.3182/20090630-4-ES-2003.00137

[16] Mujica, L., Rodellar, J., Fernández, A., & Güemes, A. (2011). Q-statistic and T2-statistic PCA-based measures for damage assessment in structures. *Structural Health Monitoring: An International Journal*, *10*(5), 539–553. https://doi.org/10.1177/1475921710388972

[17] Rastegari, A., & Bengtsson, M. (2015). Cost effectiveness of condition based maintenance in manufacturing. In *2015 Annual Reliability and Maintainability Symposium (RAMS)* (pp. 1–6). IEEE. https://doi.org/10.1109/RAMS.2015.7105079

[18] Srivastava, N., & Mondal, S. (2013). *Maintenance Practices in Indian Manufacturing Industries*. *International Journal of Computer Science & Management Studies* (Vol. 13).

[19] Stenholm, A., & Andersson, D. (2014). On

condition-based maintenance for machine components. Retrieved from https://lup.lub.lu.se/student-papers/search/publication/8593562

[20] Vanhatalo, E., Kulahci, M., & Bergquist, B. (2017). On the structure of dynamic principal component analysis used in statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, *167*, 1–11. https://doi.org/10.1016/J.CHEMOLAB.2017.05.016

[21] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Petri Hietaharju* and Mika Ruusunen

Forecasting and optimization of the heat demand at city level

Abstract: Computational methods have been developed for the predictive optimization of the heat demand to increase energy efficiency in heating by taking into account the point of view of both the energy producers and consumers. Research methods included the modelling of the individual buildings indoor temperature and heat demand, which can then be expanded to a larger scale to optimize the heat demand at the city level. The developed models are accurate and easily adaptable enabling the city level predictive optimization of the heat demand. This makes it possible to better adapt to and prepare for future changes in the outdoor temperature while at the same time ensuring the normal living conditions and optimized energy efficiency, also enabling the demand side management in the heating network. However, the full realization of the concept requires proper real-time and two-way information flow through the whole energy chain.

Keywords: district heating, modelling, prediction, demand side management, optimization

***Corresponding Author:** Control Engineering, Environmental and Chemical Engineering, University of Oulu, E-mail: petri.hietaharju@oulu.fi

Second Author: Control Engineering, Environmental and Chemical Engineering, University of Oulu, E-mail: mika.ruusunen@oulu.fi

1 Introduction

Energy Efficiency Directive (EED) [1] sets binding measures for EU countries to improve energy efficiency by 20% at EU level by 2020. In 2016, an update to EED was proposed setting a new 30% energy efficiency target for 2030 [2]. At the same time, buildings represent 20–40% of the total energy consumption and half of this energy is used for heating, ventilation and air conditioning (HVAC) [3]. Furthermore, for 15 of the 28 EU countries the annual heat demand in buildings is larger than electricity and cooling demands [4]. Also, fossil fuels are still used to produce most of the heat [5]. For the aforementioned reasons, the implementation of new energy efficiency measures for the heating and building sectors is of utmost importance.

As the requirements for energy efficiency are becoming stricter, it is no longer sufficient to consider buildings as isolated elements in energy systems [6]. They have to be treated as active participants having storage capabilities and even their own energy production. Therefore, buildings have to be taken into consideration when developing new control and optimization schemes for district heating systems. A concept for optimizing the heat demand in district heating systems has been proposed by the authors [7] and is presented in Fig. 1. The concept approaches the subject by predicting the heat demand and then optimizing the heat production utilizing demand side management (DSM). DSM refers to the change in energy consumption by the end user in response to the changes in the price or the production of the energy [8]. However, city level consumption forecasts can be extremely time-consuming if the simulations are done on a single building level, due to data gathering, simulation and monitoring efforts and the estimation of uncertainties [9]. Consequently, forecast models are widely used for individual buildings, but their application at the large scale is lacking [9–11]. It has been even stated that it is impossible to model every building separately, one of the main reasons being the lack of real measurement data [12]. However, today many buildings are equipped with smart meters that record heat consumption in intervals of an hour or less. Furthermore, model predictive control (MPC) has been one of the most studied control strategies for buildings during the last decade, offering an efficient way to perform demand response actions in buildings, but the amount of modelling work required makes the implementation expensive [13–15]. Ease of modelling would make the forecasting of heat demand and the implementation of predictive control strategies at the building and city level more cost-effective. In this regard, the applied models have to be easily reproducible for multiple buildings. This sets requirements for the simplicity and ease of parametrization of the models. The straightforward implementation in real applications should also be kept in mind. In modern automation, the cost of implementation work plays a key role while the cost of the hardware is decreasing.

In this work, the developed modelling approaches are presented to realize the predictive optimization

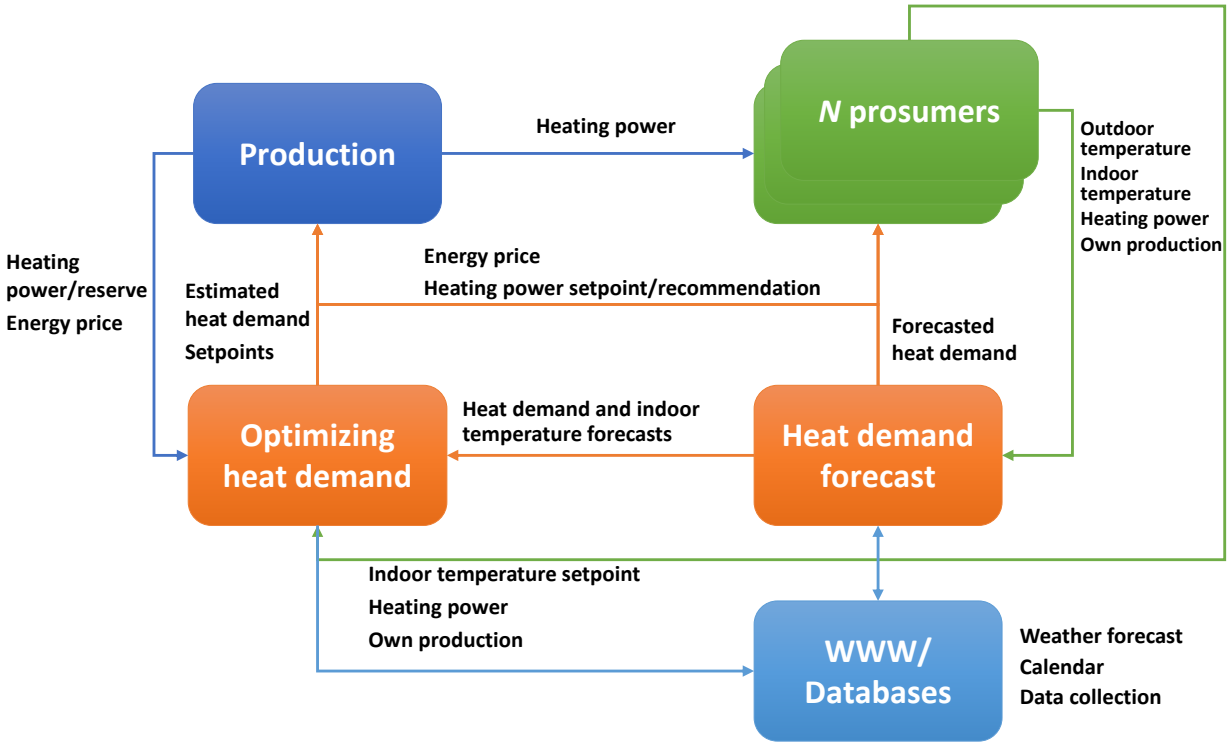


Fig. 1. Concept for the predictive optimization of the heat demand.

concept illustrated in Fig. 1. Then, the application of the developed modelling methods to optimize the heat demand at city level are discussed.

2 Modelling

Models are the basis for any MPC. Straightforward modelling methods would enable MPC to be implemented in buildings at city level. The concept presented in Fig. 1 builds on the forecast of the heat demand of individual buildings thus enabling DSM. Many of the previous works have considered only the total heat demand forecast of a district heating system. These approaches would not enable DSM actions as forecasts for the heat demands of individual buildings are not included. Furthermore, when optimizing the heat demand utilizing DSM, maintaining the indoor temperature at an acceptable level in buildings is important as the control actions should ensure the quality of the living conditions for the residents. Therefore, a mathematical model for the indoor temperature of a building is critical for enabling control and optimization strategies aiming at higher energy efficiency [16].

2.1 Forecasting the indoor temperature

For wide use of any indoor temperature model, it should be applicable to different types of buildings with minimum extra implementation work. However, most of the research have focused on a single building for the development and testing of the models. Furthermore, many of the models found in the literature need detailed information about the building properties and large amount of representative measurements together with many parameters. All of this would increase the complexity and the implementation work of the models thus limiting their application to different buildings in real environments. If the models are only tested in one building, it is very much possible that they cannot be transferred directly to another building. Then it comes to the amount of work needed to transfer these models into the different buildings. If the model is to be used only to control an individual building, the implementation time will not necessarily be an issue. However, as the intention of the authors is to perform the optimization of the heat demand at district and city level, the short implementation time for the model is crucial. When the model is implemented in hundreds or thousands of buildings, days of modelling work on one building is not acceptable.

To overcome these modelling issues, a new dynamic modelling approach was developed to predict and optimize the indoor temperature in large buildings [17]. To ensure the model generalizability to the whole building stock with reasonable prediction accuracy, the modelling approach combines easily available, existing measurements, building information and tabular values while minimizing the number of model parameters and inputs. A low number of parameters, easily available

measurements and generalizable model structure make the parameter identification of the model easy in comparison to present modelling methods. The average relative modelling error of the developed model was below 5%. The results confirmed that the model can be used to predict and optimize the indoor temperature in large buildings. A low number of needed measurements and generalizable model structure would allow the implementation and adaptation of the model to a wide variety of different buildings as a part of city level energy optimization concepts.

2.2 Forecasting the heat demand

Most of the studies that have considered heat demand forecast in individual buildings have had only one building for the model development and testing. Application of the models to a larger building stock using the same model structure would not necessarily result in the same accuracy. Some studies have also utilized data from simulated buildings. Then the model performance in real buildings may remain questionable. Although, there are studies that have considered more than one real building, there appears to be no study where hourly heat demand for a large district heating system has been forecasted utilizing models for real individual buildings.

Considering the above, two different straightforward modelling approaches were developed to forecast the hourly heat demand at city level considering more than 4000 individual buildings [18]. The proposed modelling approaches forecast the heat demand for individual buildings and at city level, enabling DSM. The results showed that the relative error was 4% for the city level heat demand forecast. Low amount of estimated parameters reduced the calculation time and easily attainable measurement data facilitates the implementation of the models for thousands of buildings.

3 Optimization of the heat demand

Today the heat demand for district heating is forecasted based on the outdoor temperature. This forecast is for the production of the heat and does not take into account the real heat demand of the buildings that the heat is being provided to. This result in non-optimal heat production. Furthermore, the lack of information from the consumption side prevents any DSM actions that could provide flexibility for the heat production. Heat demand forecast based on the forecasted heat demand of individual buildings together with the information on the indoor temperature of the buildings would enable different DSM action as presented in Fig. 2. These include peak load cutting, the minimization of the heat demand and timing of the energy production.

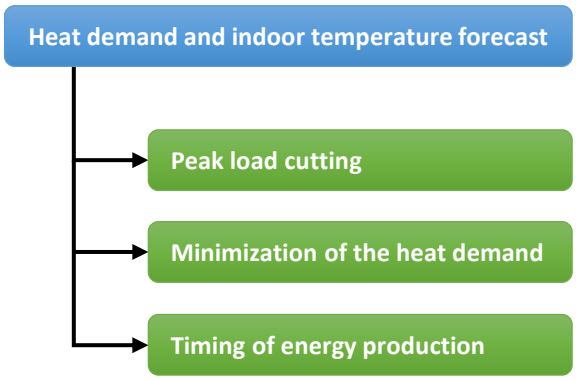


Fig. 2. Different optimization strategies for heat demand enabled by the demand side forecast.

3.1 Peak load cutting

Peak loads refer to times of high energy consumption that exceeds the production capacity of the power plant. The heat demand forecast would be used to identify these peak loads before they happen and the thermal mass of the buildings could be used to cut them. Fig. 3 shows an example of the peak load cutting by utilizing the thermal mass of buildings by preheating them thus lowering the heat demand during the forecasted peak load. It should be noted that the total heat demand remains similar in both cases. So there are not necessarily any direct benefits to building owners, rather the benefits are for the heat producer for not needing to start auxiliary power plants which would increase the production costs. However, it should be noted that this is highly case dependent and there could be energy savings when applying peak cutting. This could happen if buildings are already overheated or the outdoor temperature profile is favorable. This is also highly dependable on the allowed indoor temperature limits.

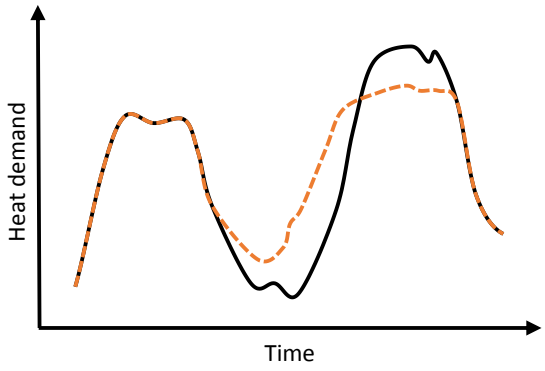


Fig. 3. An example of the peak load cutting. The black line is the heat demand without peak load cutting and the red dashed line is the heat demand with peak load cutting.

Simulations of different peak load cutting scenarios have been performed in two apartment buildings by utilizing the developed indoor temperature model [19]. One building was built in 1972 and the other in 2011. The results showed that the studied buildings had very different heat storage capacities. In the newer building, even 70% peak load cuts were possible without compromising the indoor temperature. However, in the older building 30% peak load cuts decreased the indoor temperature below the desired level. The results confirmed that the system level effect of peak load cutting cannot be concluded based on the results of a single building. Only by investigating systems with multiple buildings, the city level peak load cut capacity utilizing heat storage in buildings can be reliably evaluated.

3.2 Minimization of the heat demand

Optimization strategy that would have direct benefit for the building owners would be the minimization of the heat demand. This requires knowledge about the indoor temperature inside the building and its future projections. The minimization of the heat demand would be enabled by more stable indoor temperature control taking better into account the future outdoor temperature for example avoiding overheating when the outdoor temperature is rising. Fig. 4 illustrates what the result of minimization of the heat demand could be at city level. Again, to have an effect on the city level, the method would need to be implemented in multiple buildings.

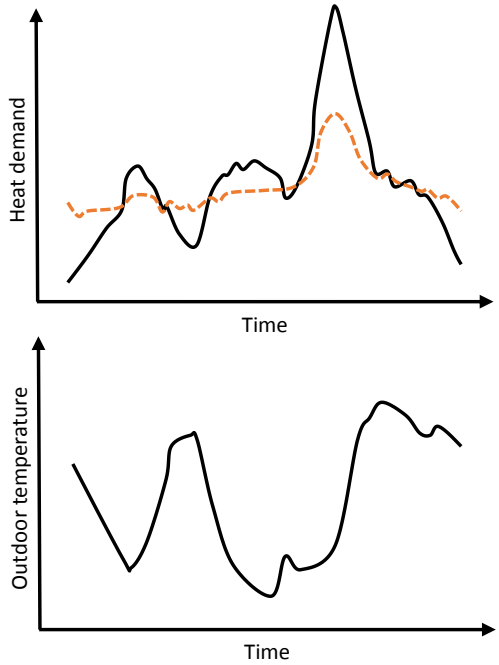


Fig. 4. The minimization of the heat demand. The black line is the heat demand without minimization and the red dashed line is the heat demand with minimization.

Preliminary results from a field test, where the optimization of the heat demand was performed in a school building, showed that significant savings in heat consumption and reduction in peak loads are possible [7]. Compared with the reference day, 14% energy savings were achieved in one day by optimizing the heat demand. It meant 1 MWh reduction in the heat consumption and additionally an average of 25% cut in peak loads. This demonstrates that there is a huge energy saving potential in the heat demand of buildings.

3.3 Timing of energy production

The timing of energy production refers to the timing of electricity production in combined heat and power (CHP) plants. At favorable times, electricity production could be increased and the extra heat could be stored in the buildings. As the trading in the Scandinavian electricity market is performed one day in advance, the predictive information on the heat demand and indoor temperature of the buildings is crucial.

3.4 Realization of the concept

In the context of the concept in Fig. 1, all the aforementioned predictive optimization strategies would utilize buildings as short term heat storages which is an effective and efficient way to store heat [20, 21]. It is well known that the peak loads can be cut, the indoor temperature swings can be reduced and the time of the heat demand can be shifted by utilizing buildings as short term heat storage [22–26]. As the heat storage capacity of buildings is already existing, only proper ways to utilize it are needed. The easily adaptable models discussed in Section 2 [17, 18] would enable the application of the predictive optimization methods to the whole building stock providing predictive information on the heat demand and indoor temperature in buildings. The optimization could be implemented as a continuous process where the buildings minimize their own heat consumption while maintaining the living comfort. On the other hand, the heat demand forecast model could also be used to provide predictive information on the future heat demand and DSM actions could then be executed when needed. This could be related to peak load cutting or timing of electricity production in case of CHP plants. Of course these two approaches could be combined. In any case, the full realization of the concept would require proper real-time and two-way information flow through the whole energy chain.

4 Conclusions

Computational methods for the predictive optimization

of the heat demand to increase energy efficiency in heating have been developed. Models for the indoor temperature and heat demand have been developed. The developed indoor temperature model can predict the indoor temperature in buildings with under 5% relative error. The average relative error of the total heat demand forecast was 4%. The utilization of buildings as short term heat storages to optimize the heat demand have been discussed. Simulations and the performed field test have shown that the buildings can be used for short term heat storage to achieve significant reduction in peak loads and an increase in energy efficiency by applying the developed modelling methods.

In conclusion, the presented modelling approaches enable the city level optimization of the heat consumption due to their reproducibility and accuracy. However, the realization of the concept requires proper real-time and two-way information flow through the whole energy chain. In addition, although the simulations with real data give valuable information on the feasibility of the developed methods, more actual testing in the real environment would be crucial to commercialize the results.

Acknowledgements

This research was funded by TEKES through the project KLEI (40267/13) and the Academy of Finland through the project SEN2050 (287748).

References

[1] Directive 2012/27/EU of the European Parliament and of the Council, http://data.europa.eu/eli/dir/2012/27/oj

[2] Proposal for a Directive of the European Parliament and of the Council amending Directive 2012/27/EU. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0761

[3] Pérez-Lombard L., Ortiz J., Pout C., A review on buildings energy consumption information, Energy Build., 2008, 40, 394–398, DOI: 10.1016/j.enbuild.2007.03.007

[4] Connolly D., Heat Roadmap Europe: Quantitative comparison between the electricity, heating, and cooling sectors for different European countries, Energy, 2017, 139, 580–593, DOI: 10.1016/j.energy.2017.07.037

[5] Eisentraut A., Brown A., Heating without Global Warming - Market Developments and Policy Considerations for Renewable Heat, IEA, Paris, 2014

[6] Allegrini J., Orehounig K., Mavromatidis G., Ruesch F., Dorer V., Evins R., A review of modelling approaches and tools for the simulation of district-scale energy systems, Renew. Sustain. Energy Rev., 2015, 52, 1391–1404, DOI: 10.1016/j.rser.2015.07.123

[7] Hietaharju P., Ruusunen M., A concept for cutting peak loads in district heating, Proceedings of the Automaatio XXI, Helsinki, Finland, 17–18 March 2015

[8] Albadi M.H., El-Saadany E.F., A summary of demand response in electricity markets, Electr. Power Syst. Res., 2008, 78, 1989–1996, DOI: 10.1016/j.epsr.2008.04.002

[9] Tardioli G., Kerrigan R., Oates M., O'Donnell J., Finn D., Data Driven Approaches for Prediction of Building Energy Consumption at Urban Level, Energy Procedia, 2015, 78, 3378–3383, DOI: 10.1016/j.egypro.2015.11.754

[10] Ahmad T., Chen H., Guo Y., Wang J., A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review, Energy Build., 2018, 165, 301–320, DOI: 10.1016/j.enbuild.2018.01.017

[11] Frayssinet L., Merlier L., Kuznik F., Hubert J.-L., Milliez M., Roux J.-J., Modeling the heating and cooling energy demand of urban buildings at city scale, Renew. Sustain. Energy Rev., 2018, 81, 2318–2327, DOI: 10.1016/j.rser.2017.06.040

[12] Grzenda M., Macukow B., Heat Consumption Prediction with Multiple Hybrid Models, In: Omatu S. et al. (Eds.), Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, IWANN 2009. Lecture Notes in Computer Science, vol 5518, Springer, Berlin, Heidelberg, 2009

[13] Killian M., Kozek M., Ten questions concerning model predictive control for energy efficient buildings, Build. Environ., 2016, 105, 403–412, DOI: 10.1016/j.buildenv.2016.05.034

[14] Thieblemont H., Haghighat F., Ooka R., Moreau A., Predictive control strategies based on weather forecast in buildings with energy storage system: A review of the state-of-the art, Energy Build., 2017, 153, 485–500, DOI: 10.1016/j.enbuild.2017.08.010

[15] Cigler J., Gyalistras D., Široký J., Tiet V.-N., Ferkl L., Beyond theory: the challenge of implementing Model Predictive Control in buildings, Proceedings of the 11th REHVA World Congress, CLIMA 2013, Prague, Czech Republic, 16–19 June 2013

[16] Prívara S., Cigler J., Váňa Z., Oldewurtel F., Žáčeková E., Use of partial least squares within the control relevant identification for buildings, Control Eng. Pract., 2013, 21, 113–121, DOI: 10.1016/j.conengprac.2012.09.017

[17] Hietaharju P., Ruusunen M., Leiviskä K., A Dynamic Model for Indoor Temperature Prediction in Buildings, Energies, 2018, 11, 1477, DOI: 10.3390/en11061477

[18] Hietaharju P., Ruusunen M., Leiviskä K., Enabling

Demand Side Management: Heat Demand Forecasting at City Level, Materials, 2019, 12, 202, DOI: 10.3390/ma12020202

[19] Hietaharju P., Ruusunen M., Peak Load Cutting in District Heating Network, Proceedings of the 9th EUROSIM Congress on Modelling and Simulation, Oulu, Finland, 12–16 September 2016

[20] Klein K., Herkel S., Henning H.-M., Felsmann C., Load shifting using the heating and cooling system of an office building: Quantitative potential evaluation for different flexibility and storage options, Appl. Energy, 2017, 203, 917–937, DOI: 10.1016/j.apenergy.2017.06.073

[21] Pan Z., Guo Q., Sun H., Feasible region method based integrated heat and electricity dispatch considering building thermal inertia, Appl. Energy, 2017, 192, 395–407, DOI: 10.1016/j.apenergy.2016.09.016

[22] Balaras C.A., The role of thermal mass on the cooling load of buildings. An overview of computational methods, Energy Build., 1996, 24, 1–10, DOI: 10.1016/0378-7788(95)00956-6

[23] Braun J.E., Load Control Using Building Thermal Mass, J. Sol. Energy Eng., 2003, 125, 292, DOI: 10.1115/1.1592184

[24] Reynders G., Nuytten T., Saelens D., Potential of structural thermal mass for demand-side management in dwellings, Build. Environ., 2013, 64, 187–199, DOI: 10.1016/j.buildenv.2013.03.010

[25] Kensby J., Trüschel A., Dalenbäck J.-O., Potential of residential buildings as thermal energy storage in district heating systems – Results from a pilot test, Appl. Energy, 2015, 137, 773–781, DOI: 10.1016/j.apenergy.2014.07.026

[26] Verbeke S., Audenaert A., Thermal inertia in buildings: A review of impacts across climate and building use, Renew. Sustain. Energy Rev., 2018, 82, 2300–2318, DOI: 10.1016/j.rser.2017.08.083

Kai Zenger* and Nguyen Khac Hoang

Optimal control maps for fuel efficiency and emissions reduction in maritime diesel engines

Abstract: The paper introduces an advanced modelling method and optimisation algorithm, by which ship diesel engines control parameters can be effectively designed. The fuel consumtion is minimised while at the same time fulfilling the NO_x emission constraints. The problem is non-trivial: the methodology introduced proves efficient, is fair and fulfils the regulations set by the International Maritime Organisation.

Keywords: diesel engine, NOx emissions, fuel efficiency, control map, optimisation, optimal design

***Corresponding Author: Kai Zenger:** Aalto University, School of Electrical Engineering, E-mail: kai.zenger@aalto.fi
Nguyen Khac Hoang: Aalto University, School of Electrical Engineering, E-mail: hoang.kh.nguyen@aalto.fi

1 Introduction

Oceangoing ships are the largest single cause of nitrogen oxides (NOx) emissions globally, and NOx is generally a major air pollutant in the atmosphere. Most of these emissions are released near the land, which causes a major pollutation problem and health risk to the people. It has been reported that outdoor air pollutation caused about three million premature deaths globally in 2010. Since ship transportation is constantly increasing, it is easy to understand that the International Maritime Organization (IMO) is setting more and more stringent restrictions to ship emissions [1]. Large tankers are major pollutants driven by diesel engines. Even if the number of diesel engines in automobile industry is foreseen to decrease fast in the future, such a trend cannot be foreseen for maritime engines, because replacement of large diesel engines as power source in maritime applications seems to be a hopeless task for several decennia to come.

The engine manufacturers are interested in developing more and more efficient engines with increasing efficiency, reduced fuel consumption and redured emissions. Unfortunately, considerable efficiency increase is already hard to establish, and reducing fuel consumption generally implies higher NOx emissions and vice versa. Because of this, IMO has set regulations (Tier II and Tier III) that set limits to NOx emissions in some operation points (speed and load) of the ship. However, only a few operation points have been set, which means that it is unclear how the ship emissions should be controlled over the whole operation range. Even worse, the current regulations give a possibility to "cheat" by setting the emissions low at the given operation points (high fuel consumption) but use all effort to save fuel in other operation points (high NOx emissions). The paper presents a method, where the Design of Experiments (DOE) method is used to model the fuel consumption and NOx emission at any given operation point. It then becomes possible to construct smooth functions to cover all operation points of the ship engine. At any operation point an optimization problem can be set and solved, where the fuel consumption is minimized under a given constraint of maximum NOx emission. The solution gives certain control parameters of the ship (common rail pressure, charge air pressure, start of ignition timing), which are to be used in the operation point in question for optimal performance. It now becomes possible to compare the fuel consumption and emission level under standard routes travelled by the ship. In addition to that it becomes possible to construct optimal operation parameters and allowed NOx levels under a large number of operation points, thus giving advice to IMO how te future regulations could be stated, in order to cover all operational areas and to avoid all possibility to cheat. The results of the paper have been obtained and confirmed using real diesel engine data from large engine manufacturers.

2 Research objectives

The fuel consumption and emissions of a diesel engine can be affected by a set of input parameters, which can be set and controlled before and sometimes even during a cruise. In this section key parameters are presented and their influence on the break specific fuel consumption (BSFC) and nitrogen-oxide (NOx) emissions are studied. Specifications by the International Maritime Organisation are shortly presented and later in Section 3 the Design of Experiments method is described as a

means to model the BSFC and NOx in a given operation point as a function of key operation parameters.

2.1 Key parameters of the engine

The intake pressure P_I, the common rail pressure P_{CR} and the start of injection SoI are considered to have a major effect on BSFC and NOx. Their effect is highly dependent on the operation point of the engine (speed and load), [2], [3]. The fuel consumption, here BSFC is defined by the fuel flow per the produced work [kWh]. The emissions, here specifically NO_x emissions are measured by the amount per produced work [g/kWh].

High intake pressure (boost pressure) has a major effect on engine performance. It leads to efficient combustion of the air/fuel mixture with reduced exhaust emissions of unburned fuel. At the same time however the amount of CO_2 and NO_x emissions can increase [4].

Common rail pressure affects the fuel spray into the cylinders and therefore has an impact on the ignition quality. High pressure reduces the amount of smoke but at the same time increases NO_x emissions. Moreover, there is a big relationship between high common rail pressure and engine load: under heavy load high pressure is beneficial, but the NO_x emission is increased. Under light load the fuel consumption ($BSFC$) increases.

The start of ignition (SOI) means the time at which fuel is injected into the cylinder to start the combustion process. It is measured as the crankshaft angle reaching the top dead center. Nowadays the heat release process is a major concept in the development of combustion engines in order to optimise fuel saving and emission reductions. Therefore the ignition process and specifically the (SOI) is particularly important.

Static maps are used to determine the control variables of the engine. The operation point (speed, load) pay a key role, because the optimal parameters are a nonlinear function of them. In Fig. 1 it is shown how the setpoints in engine control are set by the maps.

2.2 Emission reduction targets

The International Maritime Organisation (IMO) sets regulations for the allowed emissions of ships [5]. An example is given in Fig. 2.

According to Tier II conditions at the operation speed 1000 rpm the NO_x emission can be at most 9 g/kWh calculated over the whole trip. For practical use

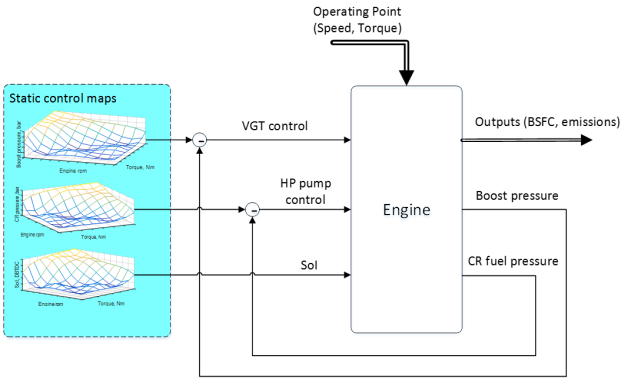


Fig. 1. Static maps to determine the control setpoints for intake (Boost) pressure, common rail pressure and (indirectly) fuel consumption. VGT=variable geometry turbocharger, HP=high pressure pump

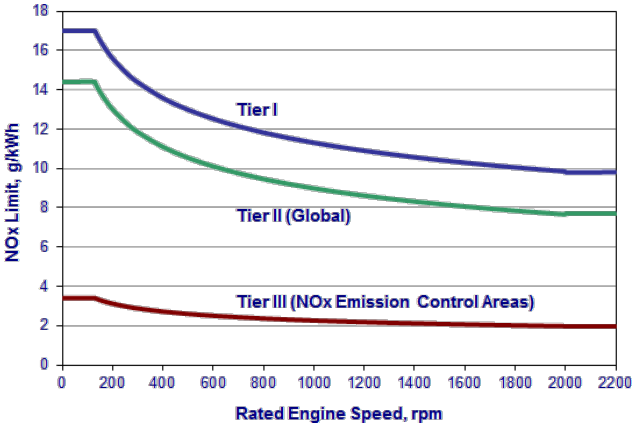


Fig. 2. NO_x emission limits under TierI, II and III

the requirement can be extended by defining the operation point as (speed, load) and by giving weights related to the time that a ship travels at a given load. For example, consider Fig. 3 where the ship is assumed to travel at a constant speed but with varying loads. The weights represent the assumed time fraction of the total trip that the ship travels under the load given in the table.

E2: Constant-speed main propulsion application including diesel-electric drive and all controllable-pitch propeller installations	Speed (%)	100	100	100	100
	Power (%)	100	75	50	25
	Weighting factor	0.2	0.5	0.15	0.15

Fig. 3. Setpoints with different weights

If the NO_x emissions corresponding to the operation points in the table are $n4$, $n3$, $n2$ and $n1$, respectively,

the hard constraint for the emissions during the cruise is

$$(0.15 * n1 + 0.15 * n2 + 0.5 * n3 + 0.2 * n4) \ < 9 \text{ g/ kWh} \quad (1)$$

It is important to realize that the IMO regulations are fulfilled as long as (1) is fulfilled. In other words there is no maximum value of NO_x emissions at any specific load. Under the contemporary regulations this gives a possibility to "cheat" by allowing high emissions on loads, under which the ship is normally not run. Also, minimizing the emissions at the load points 25%, 50%, 75%, 100% gives clearly an acceptable running policy, which unfortunately gives the possibility to high emissions (with good savings in fuel use) in "intermediate" loads.

It is clear that the design of optimal engine parameters to be used during a cruise of a ship must be designed such that both fuel use and emissions are set to minimal values, however such that the constraints on emissions at all loads are fair (no "cheating").

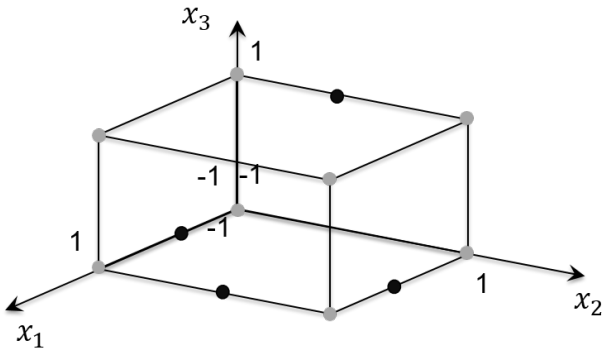


Fig. 4. Choosing the coded factors in the design of experiments

Factor	(-)	(0)	(+)
Boost, bar	1.1	1.5	1.9
CR pressure, bar	1100	1250	1400
SoI, DbTDC	-4	4,5	13

Fig. 5. Factor intervals (example data only)

3 Modelling by the design of experiments

In order to design the optimal control policy as function of the operation point of the ship the model for fuel usage (BSFC, break specific fuel consumption) and NO_x emissions is needed. To construct these models with as small number of measurements the method (*Design of Experiments, DOE*) was used. For the theory of *DOE* see e.g. [6]. The coded response variables are $BSFC$ and NO_x and the coded factors boost (intake) pressure x_1, common rail (CR) fuel pressure x_2 and start of ignition (SoI) x_3. The experiments are typically run at 500 rpm speed intervals at 8-16 load points. The idea is to use statistically relevant experiments to construct the relationship ot the output function with the input variables by using as small number as experiments as possible. In the current case the Box-Behnken design method with 15 experiments at each setpoint was used [4]. See Figs. 4 and 5. Each variable has been given three values denored as -1, 0 and +1 and experimental runs are done according to the Box-Behnken table.

The maps of the coded variables are obtained as a function of setpoint, see Figs. 6, 7, 8.

After the test runs have been done, least squares method is used at each setpoint to determine the coefficients in

$$\begin{aligned} BSFC = {} & a_0 + a_1 P_1 + a_2 P_{CR} + a_3 SoI \\ & + a_{12} P_1 P_{CR} + a_{13} P_1 SoI + a_{23} P_{CR} SoI \\ & + a_{11} P_1^2 + a_{22} P_{CR}^2 + a_{33} SoI^2 \end{aligned} \quad (2)$$

$$\begin{aligned} NOx = {} & b_0 + b_1 P_1 + b_2 P_{CR} + b_3 SoI \\ & + b_{12} P_1 P_{CR} + b_{13} P_1 SoI + b_{23} P_{CR} SoI \\ & + b_{11} P_1^2 + b_{22} P_{CR}^2 + b_{33} SoI^2 \end{aligned} \quad (3)$$

where the values for $BSFC$ and NO_x were obtained by measurements from a real engine.

Note that the models with one set of parameters a_i, b_i are valid in one operation point only. Also note that the variables P_1, P_{CR} and SoI are always restricted by lower and upper limits

$$\begin{aligned} P_{1l} \le P_1 \le P_{1u} \\ P_{CRl} \le P_{CR} \le P_{CRu} \\ SoI_l \le SoI \le SoI_u \end{aligned}$$

It would now be possible to consider the table E2 of IMO regulations, Fig. 3, and take the 4 operation points N_1 (100, 25), N_2 (100, 50), N_3 (100, 75), N_4 (100, 25) to

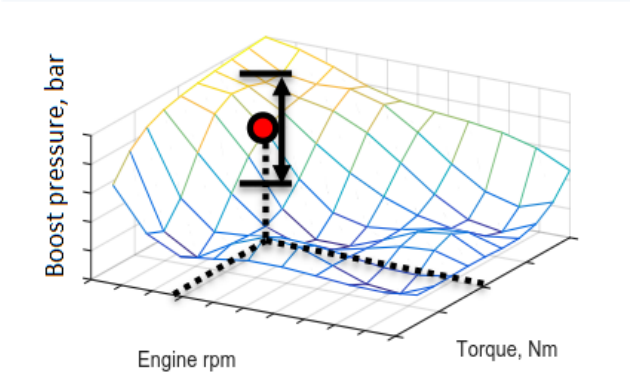


Fig. 6. Intake pressure variation and limits

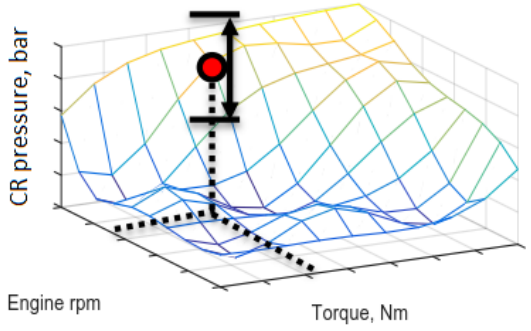


Fig. 7. Common rail variation and limits

calculate the optimal operational parameters by solving for all i

$$\begin{aligned} &\min BSFC(i) \\ &\text{s.t} \quad NO_x(i) < \alpha \\ &\quad P_{1l} \le P_1(i) \le P_{1u} \\ &\quad P_{CRl} \le P_{CR}(i) \le P_{CRu} \\ &\quad SoI_l \le SoI(i) \le SoI_u \end{aligned} \tag{4}$$

where α is the limit 9 g/kWh. However, acting like this would probably lead to a result that NO_x levels would be exremely high for other operation points than those four ones selected in the E2 table (to minimise BSFC). That would not be fair and it would be "cheating", although it would formally satisfy the IMO regulations. To solve this problem, more advanced algorithms are needed.

Optimization problems like in (4) can be solved numerically by e.g. the Sequential Quadratic Programming (SQP) algorithm. There the nonlinear problem is at each iteration embedded into a Quadratic Programming subproblem in order to form the new iteration of the solution vector [7].

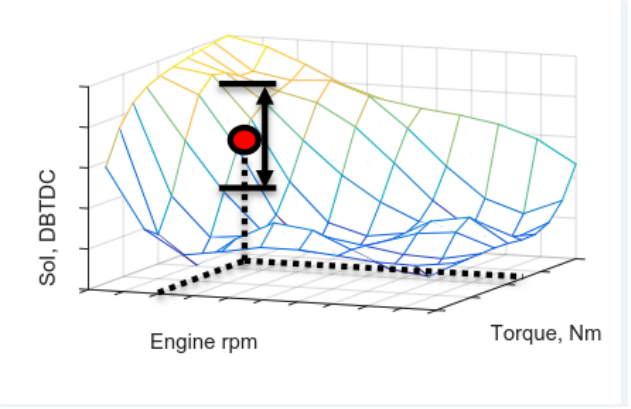


Fig. 8. Start of Ignition variation and limits

4 Optimisation by including weights and the cruise profile

Consider Fig. 9 which gives a load distribution over a

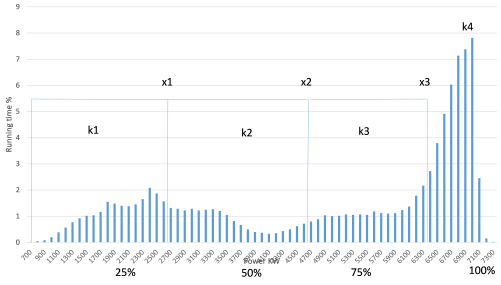


Fig. 9. Cruise characteristics (invented data)

whole cruise of a ship. It is possible to minimize the fuel consumption during the whole cruise while fulfilling the NO_x constraints as follows.

- Choose 4 random points n_1, n_2, n_3, n_4 for admissible values in (3)
- Set the points as a candidate for the NO_x curve as in Fig. 10.
- Check that the constraint (1) is fulfilled. If not, pick new values for $n1$, $n2$, $n3$ and $n4$.
- Interpolate by piecewise approximation for all load vs. NO_x data α_i available from measurements, as in Fig. 11.

– Minimise the fuel consumtion by

$$\min BSFC(i)$$
$$NO_x(i) < \alpha(i)$$

where $\alpha(i)$ is obtained from the piecewise linear NO_x curve.
– Pick new values for $n1$, $n2$, $n3$ and $n4$ and start again from the first step. After a desired number of trials the NO_x curve is chosen, which gives the minimum fuel consumption.

Note that the amount of iterations can be quite large. If three values for the intake pressure, three values for the common rail pressure and three values for the ignition time are chosen, each n_i:s has 27 possible values, and the whole grid of n_i has 27^4 possible values. Some of these are not admissible in that they do not fulfil the IMO E2 constraint, but anyway the search space is large.

The algorithm calculates 12 optimal engine parameters (intake pressure, common rail pressure and ignition time at 4 operation points). The E2 weighting quarantees that the IMO regulation is fulfilled (hard constraint).

The figure 10 shows the resulting NO_x curve. The 'circled' points correspond to the loads 25%, 50%, 75% and 100%. The curve fulfills the IMO regulations by construction (optimisation constraint) and by a piecewise linear interpolation it gives a reasonable target for allowable emissions at each load. See Fig. 11 for an example.

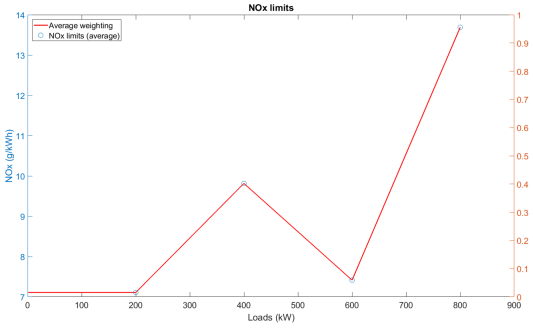


Fig. 10. NOx limits

The final objective function is the estimated fuel consumption over the whole cruise (e.g. Fig. 9) under assumption of a load profile and a fixed NO_x curve. The parameters obtained by searching over different load profiles are then used to apply over a real cruise, and total fuel savings can be evaluated. In Fig. 12 an

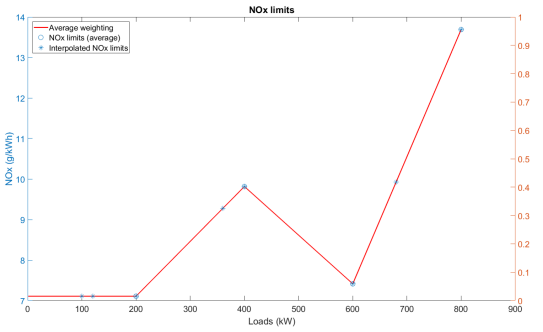


Fig. 11. NOx limits at interpolated load points

interpolated curve has been presented, where 8 operation points have been used during a given cruise. The corresponding estimated engine fuel use in a year was to be 576 tons, when nominal operation leads were 607 tons of fuel usage. The saving is approximately 5%.

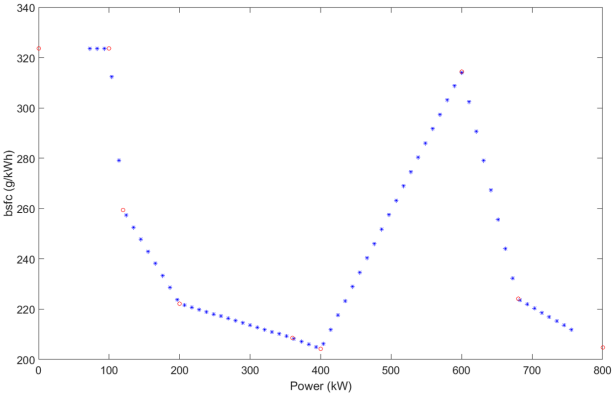


Fig. 12. Interpolated optimal bsfc values at based on eight operation points

5 Conclusion

The paper has discussed a systematic method to optimise engine control parameters of a ship diesel engine, with the goal to minimise fuel consumption but fulfilling the NO_x constraints set by the IMO. These regulations have so far been given in a way that gives room to interpretations and even possibility to unfair design, where NO_x emissions are only minimised where absolutely necessary. The paper introduces an advanced algorithm, which actually sets a NO_x curve that can be used in minimising the fuel consumption with constraint on any operation point. It is believed that the method

can be used for general greenhouse gas emissions also (not only NOx). Also, the design gives valuable information on the Tier regulations as such, specifically on how they should be considered over a whole cruise of a ship. It goes without saying that the optimization approach presented in the paper can lead to extremely useful improvements technically, economically and from the environmental viewpoint.

In spite of the fact that the results are promising, the method and algorithms are still under investigation.

ACKNOWLEDGMENT

The EU project Hercules 2 funded by the European Commions and the Intens project funded by Business Finland are greatly acknowledged.

References

[1] R. M. A. VI. Regulations for the prevention of air pollution from ships and nox technical guide 2008. *International Maritime Organisation*, 2009.

[2] L. Guzzella and A. Amstutz. Control of diesel engines. *IEEE Control Systems*, 18(5):53–71, 1998.

[3] J. B. Heywood. *Internal Combustion Engine Fundamentals*. Mc-Graw-Hill, 1988.

[4] N. K. Hoang and K. Zenger. Designing optimal control maps for diesel engines for high efficiency and emission reduction. European Control Association, 2019. (to appear).

[5] Imo marine engine regulations. *Emission Standards*. URL https://www.dieselnet.com/standards/inter/imo.php.

[6] P. G. Mathews. *Design of Experiments with MINITAB*. ASQ Quality Press, 2005.

[7] P. T. Boggs and J. W Tolle. Sequential quadratic programming. *Acta numerica*, 4:1–51, 1995.

Heli Karaila*, Lasse Järvinen, Ari Oksanen

Mass flow-based controls with solids measurements reduce sludge handling costs

Abstract: The Tampere Water Viinikanlahti wastewater treatment plant in Tampere, Finland has commissioned what is believed to be the world's first multi-variable predictive controller (MPC) of a centrifuge sludge dewatering operation based on multiple online measurements of solids content. The online measurements have replaced manual testing that was considered too slow or not timely enough for optimum real time control. In addition to the centrifuge mass flow-based control other unit operations such as primary clarifier sludge pump scheduling and optimization of anaerobic digester input solids are now based on mass flow values rather than volumetric values. The project objectives of minimizing the recirculation of material inside the plant and optimizing the solid amount in the dewatered dry cake were more than met in addition to achieving significant chemical and energy savings.

Keywords: wastewater, centrifuge, dewatering, optimization, MPC

***Corresponding Author:** Business manager Heli Karaila, E-mail: heli.karaila@valmet.com

Second Author: Lasse Järvinen
lasse.jarvinen@tampere.fi, Ari Oksanen
ari.oksanen@tampere.fi

1 Introduction

At the wastewater plants it is still typical that monitoring and control of the sludge process is done based on laboratory samples follow up and visual look of the process. In order to optimize the process there are needed online measurements with the control application. In this article is descriped Tampere Water Viinikanlahti wastewater treatment plant project with the results in Tampere Finland. In the project there was added multiple real time solids measurements to the sludge process in order to replace manual laboratory follow and optimize the process with the massflow based controls. At the centrifuge there was commissioned multi-variable predictive controller (MPC) of a centrifuge sludge dewatering operation based on online measurements of solids content.

2 Tampere Viinikanlahti process

Fig. 1. Seventy-five per cent of Tampere's wastewater is processed at the Viinikanlahti wastewater treatment

The biological and chemical wastewater treatment in Viinikanlahti is based on an activated sludge process coupled with phosphorus precipitation by ferric sulfate. The process consists of screening, grit removal, primary sedimentation, aeration and secondary sedimentation. Wastewater sludge is digested in anaerobic digesters after which the sludge is dewatered by centrifuge. The daily flow is about 70,000 m^3 (230,000 p.e.); producing approximately 63 m^3 of sludge is per day. Biogas from the digester is used to generate electrical energy and heating for the plant.

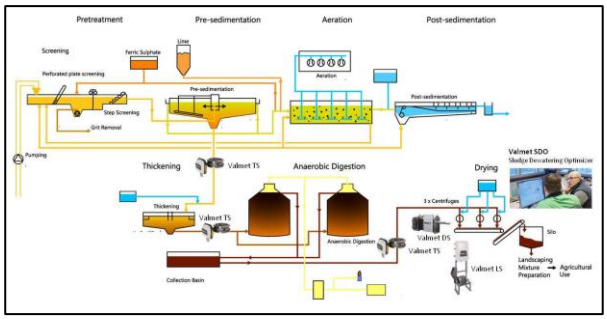


Fig. 2. Tampere Water Viinikanlahti wastewater treatment plant with Valmet's solids measurement (TS, LS and DS) locations and Valmet SDO

3 The need for reliable online measurements

Traditionally the only reliable online measurement available to wastewater process engineers has been flow. Laboratory test results only partially support the optimization process. This is because laboratory tests are carried out rather seldom and the results are archived after a considerable delay. As an example of the need of online measurement instead of laboratory measurement is shown in this example of a centrifuge centrate measurement made in an earlier trial at a wastewater plant in Southern Finland (Fig. 3). On the basis of the laboratory measurement, it is not possible to follow the process dynamics for the correct polymer dosage, because the situation changes immediately after the laboratory sample is taken.

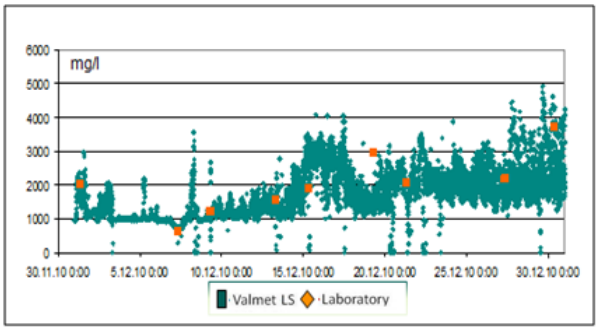


Fig. 3. Laboratory samples taken normally twice per week provide no information of the actual dynamic situation

4 Clarifier and thickener solids measurements

Valmet TS has been developed from a third generation microwave solids transmitter originally designed for use in the demanding environment of a pulp and paper mill. It uses patented microwave-based technology, which allows it to measure total solids content, unaffected by affected by flow rate or color of the process stream. Solids conduct microwaves faster than water so that shorter microwave transmission times correlate to higher solids content. The relationship is linear, making it is easy to calibrate the device regardless of what is being measured.

Fig. 4. *Valmet TS microwave total solids transmitter*

Three Valmet Total Solids transmitters (Valmet TS) are used to measure total solids after pre-sedimentation, after thickening and before the dewatering centrifuge. An immediate advantage after installation was now process parameters could now be seen based on solids mass rather than the volumetric total flow. The four circular primary clarifiers in the Viinikanlahti pre-sedimentation stage allow solids in the wastewater to settle to the bottom of the clarifier before being pumped to the sludge thickening tank. Prior to the Valmet TS installation, pumping from the four clarifiers was controlled in a timed sequence, which meant that from time to time a clarifier would be emptied of solids and only water would be pumped to the thickening tank. With Valmet TS measuring the total solids content, the pumping sequence from the four clarifiers is now controlled by the solids content to avoid excess water being pumped to sludge thickening.

A second Valmet TS monitors solids content after thickening and enables the optimization of the

thickener to increase solids content to the digester. The higher solids content to the digester reduces heating demand, increases digester residence time and produces more biogas.

5 Centrifuge measurements

The third Valmet TS is used to stabilize the mass flow to the centrifuge, now measured in kilos of sludge per hour. In the first phase of centrifuge optimization, performed in December 2015, this allowed dewatering polymer to be controlled as a ratio to the mass flow rather than the cubic meter-based flow rate previously used. With the mass flow under control, the second phase of optimization in January 2016 was to optimize solids in the centrate and moisture in the dry cake with a combination of torque control and polymer. A Valmet Low Solids Measurement, (Valmet LS), was installed in the centrate outflow and another specialized measurement, Valmet DS, measures the solid content of the dry cake as it falls to the conveyor.

The Valmet Low Solids Measurement (Valmet LS) measures a continuous sample flow through the system by utilizing an integrated centrifugal pump. The system has two LED light sources in a flow through measurement cell where absorption, scattering and depolarization signals from both of the light sources are measured. Valmet LS continuously measures both the entrained air index, which indicates overdosing of polymer, and the suspended solids content within the range of zero to 5,000 mg/l. The measurement cell utilizes an extremely strong sapphire glass with high optical properties and is cleaned automatically together with the sample lines at given intervals.

Fig. 5 *Valmet LS measures solids in the range of 0 to 5,000 mg/l.*

The Valmet Dry Solids Measurement (Valmet DS) extracts a continuous sample from the falling cake flow after a centrifuge or screw press and measures the solid content before returning the sample back to the process. Utilizing Valmet's proven patented microwave technology and requiring no special certification or safety procedures, it makes a stable and accurate measurement of cake solids up to 35%.

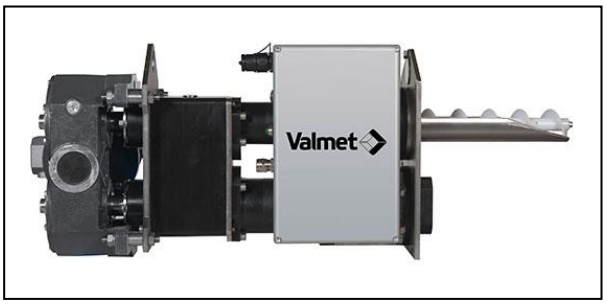


Fig. 6. *Valmet DS measures cake solids up to 35%.*

6 Centrifuge control

Traditionally the only online measurement at the sludge dewatering phase is flow; the process is controlled with laboratory based measurements or visually based on the color of centrate or dried cake appearance. When centrate color is dark, more polymer is added and operators normally overdose

polymer in order to limit the recirculation of reject (centrate) solids. However this is not only wasteful in terms of polymer but can lead to foaming in the centrifuge and inefficient centrifuge operation. Centrifuge torque also has an effect on centrate solids but is very seldom changed. The lack of information, only provided by infrequent laboratory samples with often a day's delay, means that centrifuge operation has been very much a "dark art" and very difficult to optimize.

This is where the multivariable model predictive control (MPC) as employed by the Valmet Sludge Dewatering Optimizer (Valmet SDO), a small-scale Valmet DNA control system, comes into play. With continuous centrate and dry cake solids information from Valmet LS and Valmet DS together with the stabilized mass flow to the centrifuge, the control can combine the optimum torque and correct polymer dosage. The result optimizes both energy and polymer while achieving the target reject solids and higher dryness in the dry cake from the centrifuge.

MPC uses process models (Fig 7) to predict the interactions between the modified variables (polymer and centrifuge torque). This allows the operator to make set point changes and follow up as for normal single loops. The interactions of the separate loops are taken care of automatically as a background task of the optimizer.

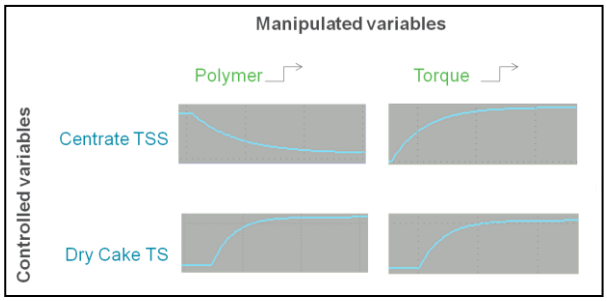


Fig. 7. *Polymer and Torque process models to centrate TSS and Dry Cake TS*

As the centrifuge torque is increased, more water is extracted from the sludge and the total solid (TS) content increases in the dry cake, but at the same time centrate total suspended solids (TSS) increase to be wastefully re-circulated through the plant reducing capacity. Increasing polymer dosage increases the solid content of the dried sludge and also reduces centrate solids. The dynamic relationship of these interactions is difficult, if not impossible, to control separately with single PID controllers but with MPC control it is easy to take care of the effects.

The MPC control principle is based on following control strategy (Fig 8.):

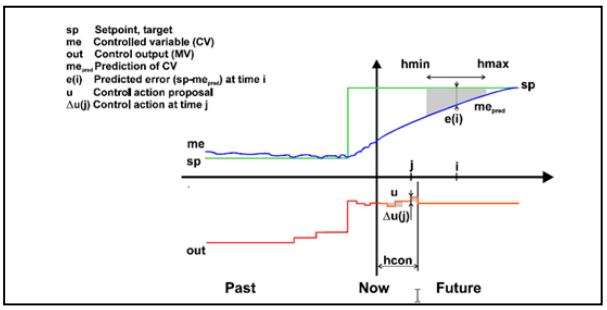


Fig. 8. Control principle of MPC for one controllable parameter

1. At each control execution moment, the controller performs a forecast of the process output, i.e., it predicts the future behavior of the controlled variables. Predictions are made over a certain time horizon *hmax* and are based on process models and known control actions (history).

2. The controller calculates the optimal subsequent *hcon* control actions, which keeps the number of errors occurring between setpoints and predicted process outputs as small as possible during the time period *hmin--hmax.* The calculation is based on an optimization of the cost function, which presents how the smallest possible error occurrence is achieved with minimal control actions.

3. First, one of the proposed control actions is applied to the process. All other actions are ignored and the whole procedure is repeated, leading to updated control actions with corrections based on the latest measurements.

7 Results

Performance of Online measurements

After the calibration of the measurements laboratory

samples were taken in order to follow up the performance of the measurements over a two month period (Fig. 9).

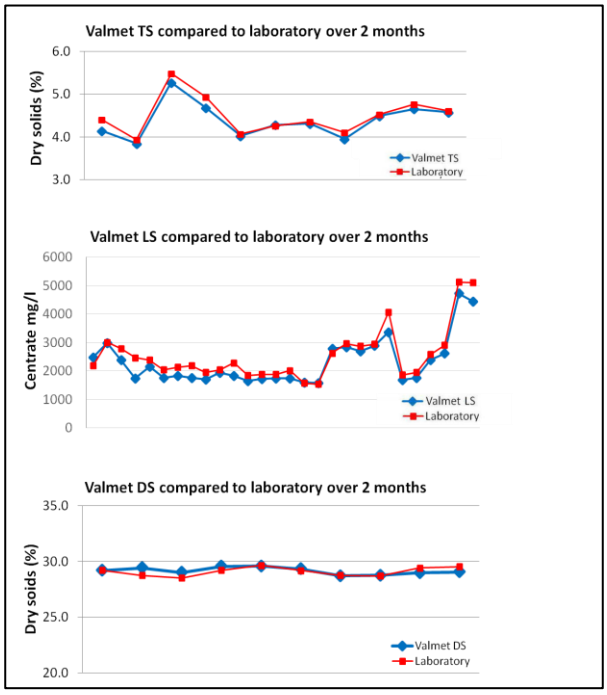


Fig. 9. Two month follow-up after initial calibration

Since startup, little or no maintenance has been needed. The optimization has been in use 24/7 and, apart from initial fine tuning of the DS dry cake measurement, all the measurements still use the same calibration parameters from start-up. Operators have found the control room display easy to understand and operate, providing them with a new window to the process (Fig 10).

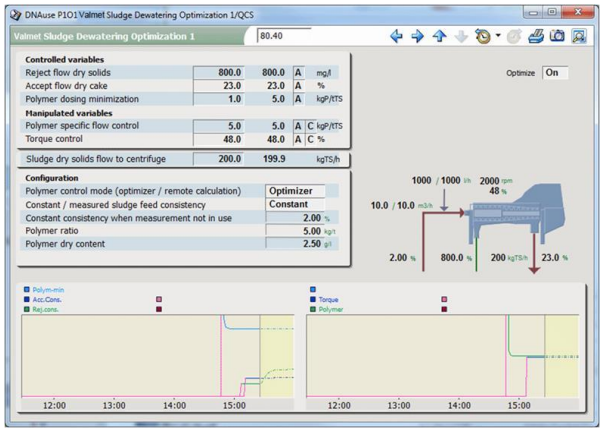


Fig. 10. Valmet SDO operator display provides a real-time picture of centrifuge operation.

8 Conclusion

In one year of operation, the measurements and control have proved to be very reliable. Measurement devices were acquired in order to have exact data from the process and this has enabled meaningful trials on the effectiveness and selection of polymers used in the plant. The controls have achieved the goals of polymer feed optimization, decreased energy consumption and savings in dry cake transportation costs. The project has also been a perfect opportunity to test new technology for the new Sulkavuori underground treatment plant being built in Tampere and estimated to start in 2023.

More than 140,000 €/year savings have been calculated as follows:

- Sludge pumping from clarifiers reduced from 76 m^3 to 50 m^3 an hour
  - Reduced excess water to thickening
  - Energy savings are approx. 37 %, 5,000 €/year

- Digester solids increased from earlier 3.5% to 5%
  - 32 % less sludge to treatment
  - Increased biogas production

- Solid content of centrate water is now 50 % lower and more stable
  - Polymer consumption has decreased almost 40 % from level 8 kg/ton
  - Savings with less material circulated, 10,000 €/year
  - Polymer savings, 49,000 €/year

- Dried cake solids content has increased by about 1-2 % from 29,7 % to over 31%

Saving in transportation costs approx. 80,000 €/a

Learning compliant assembly skills from human demonstration

Markku Suomalainen and Ville Kyrki

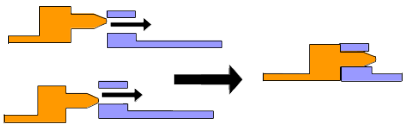
***Abstract*—Robotic assembly is mainly used inside factories where both the environment and the task for each robot stays constant and the batch sizes are large, with car factories presenting a prime example. However, in manufacturing Small and Medium-sized Enterprises (SMEs) or construction yards the level of automation is very low, mainly due to the changing environment causing two major problems for robots: firstly, the programming of robots is often difficult and thus it can take too long to make the same robot perform multiple tasks interchangeably. Secondly, the use of robots with traditional control methods requires an accurate model of the environment, which can be either costly to acquire and prone to accidental changes in the real environments (SMEs) or simply infeasible (construction). To enable the use of robots in new environments, robots must be easy to teach and able to adapt to small changes in the environment. In this paper we propose methods to use Learning from Demonstration (LfD) with compliant motions to facilitate the usage of robots in new environments.**

I. INTRODUCTION

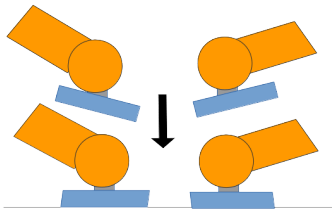
The strenuousness of programming a robot to perform different tasks is a major reason holding back the widespread use of robots in industry and at people's homes. Industrial robots are mainly used only when the same product is manufactured for long periods of times. One of the next places where the usage of robots can really increase is enterprises where production batches can be small. But to enable this step, domain experts must be able to teach the robots the required task, such that a robotics expert is not required at the stage every time the robot needs to learn a new task or fails at completing a taught task.

Learning from demonstration (LfD) is an established paradigm in robotics, where the goal is easily programmable robots. In short, the idea is to show the robot an example of a skill, which the robot learns to reproduce and generalize into other locations and similar situations. Methods to show an example include *e.g.* kinesthetic teaching (holding a gravity-compensated robot and leading it through the motions) and teleoperation. However, traditional LfD techniques struggle with compliant motions, which are required in many industrial assembly tasks.

In this paper we propose to use LfD with compliant motions to overcome the aforementioned problems. In LfD the user can show the robot how to perform a required task, using either teleoperation or kinesthetic teaching where the teacher physically holds a gravity-compensated robot and leads it through the desired task. We developed methods to ease the use of compliance on three different levels in programming a robot: on the control level, on the primitive level and on the motion sequencing level. On the control level, we propose using impedance control for cases where both the manipulator and object are ground based. On the primitive level we present a new impedance control– based motion primitive which can be used to learn and encode motions that use the environment to mitigate pose uncertainties– humans naturally have the skill to exploit contact forces in insertion tasks, and we want to convey the skill from human to robot in an efficient way. On the motion sequencing level we first show how a complex human demonstration can be segmented into phases, each of which can be modelled with the primitive. Then we present how the primitives can be sequenced online to successfully reproduce the task. Additionally, we show that the presented motion primitive can also be applied effectively for bimanual assembly tasks. Finally, we present how to learn from human teachers search motions similarly as a human inserting a plug into a socket in darkness, which can be used as efficient exception strategies in assembly. To conclude, this paper presents a framework that can accelerate the degree of automation in tasks where currently the use of robots is infeasible.



(a) Position alignment.



(b) Orientation alignment

Fig. 1: Compliant motions can be used for aligning both position and orientation of a workpiece [1]

This paper is an overview of seven publications by the authors following the aforementioned paradigms, and the mathematical details of the methods can be found from those publications. In [2], we learned desired direction and axes of compliance for a motion by assuming we can directly measure the direction of the force which the human teacher applies to the robot in kinesthetic teaching. However, we noticed that this assumption only holds true for certain force/torque sensor configurations, and hence we wanted to solve also the more general problem, where we can only measure the force between the end-effector and the environment. We solved this problem in [3], with the observation that in a compliant sliding motion there is always a certain sector of directions from which the robot can apply force to perform the observed motion. We managed to take the intersection of these sectors in

a 3-D motion, over one or more demonstrations, and thus learn the parameters for a dynamically linear compliant motion. In [1] we generalized the task to work with rotational motions as well. Furthermore, in [4] we learned how to sequence these motions to perform a full task, such as pipeline assembly. To make the robots more independent even in case of changes in the environment, in [5] we looked into whether a robot could learn to search using contact forces, similarly as a human tries to fit a key into the keyhole in darkness. Finally, in [6] we showed that our method can be applied to dual-arm tasks and examined the role of compliance in dual-arm assembly with a little more detail. Additionally, to show that the method from [3] is robust enough to work with systems where errors in measurements can be higher, we combined the method with a stability-guaranteed Virtual Decomposition Control- based impedance controller for a heavy-duty hydraulic manipulator with a 475kg payload [7].

II. METHOD

To make a robot execute a task, the task must be represented in a manner that is understandable for the robot, often called a *policy* consisting of *primitives* each of which models a *phase* of the task, such as shown in Fig. 2b where moving the block to touch the table is one phase and moving it into the corner is another. In this chapter we consider modelling and learning from a human demonstration a single phase of such a policy. This is the *continuous* level of a hierarchical policy, and the simplest examples of this sort of behaviour for modelling a trajectory are splines [8] or Bezier curves [9]. If the task requires contact with the environment, the simplest approach is to augment the trajectory with a *force profile* consisting of forces the robot should apply to the environment at each position. The obvious downside of this kind of representation is that even small changes in the environment or in the robot's coordinate system can easily cause the task to fail. Thus simply recording the trajectory and forces from a human demonstration and replaying them is not a valid LfD strategy. There exist more general and popular primitives than the simple trajectory encoders mentioned that are used nowadays to represent tasks learned from human demonstrations. However, especially when trying to learn how to take advantage of the environment with compliant motions, there are certain downsides in the currently popular motion primitives.

Perhaps the most recognized primitives used currently in LfD are Dynamic Movement Primitives (DMP) [10] and Gaussian Mixture Model (GMM) with either Gaussian Mixture Regression (GMR) [11] or Stable Estimator of Dynamical Systems (SEDS) [12]. Strengths of the DMPs include the ability to be learned online and that they can be coupled with wrench or impedance profiles. These attributes, along with the simplicity, make DMPs a popular choice for learning and encoding complicated trajectories. Additionally, with correctly chosen gains DMPs can be shown to be stable, and DMPs have been shown to be generalizable through *task-parametrization* to new situations [13], [14] by simply modifying a parameter

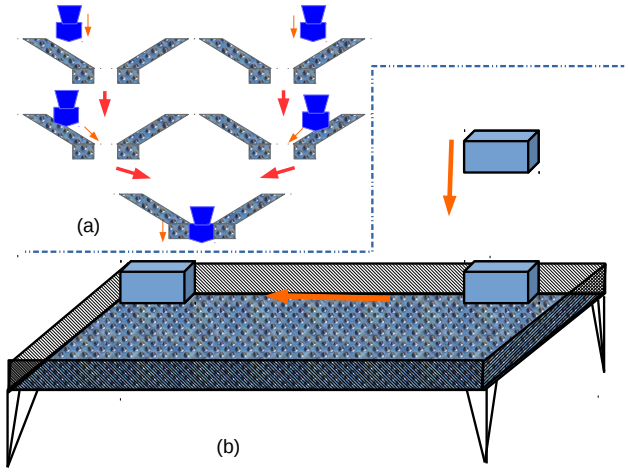


Fig. 2: Compliant motion policy used (a) to align workpieces and (b) to place a box at the corner of the table [4].

relevant to the current task. A downside of DMPs is that to learn from multiple demonstrations, tools such as Dynamic Time Warping (DTW) [15] need to be used to temporally align the demonstrations. Especially with more than two demonstrations this becomes a tedious task.

Similarly to DMPs, GMMs can be task-parametrized and augmented with a wrench profile. SEDS uses GMMs as well to model the task, but due to the use of a dynamical system, the stability can be guaranteed with correctly chosen parameters, unlike when using the statistical methods in GMR. SEDS has also been used to produce impedance to allow compliant motions [16]. The main downside of all these methods when used for compliant motions is the tight coupling of the trajectory and force profile, which makes the methods susceptible to pose uncertainties and makes taking full advantage of compliant difficult. Thus we propose new methods to efficiently learn robust compliant skills from human demonstrations.

A. Linear Motion with Compliance

In this section we present the approach called Linear Motion with Compliance (LMC), which was gradually developed in [1]–[3] and used successfully as a component in [4], [6], [7]. The key idea is that we model a task as a sequence of linear motions with compliance such that we take advantage of the environment to guide and align the tool, as shown in Figs. 2 and 1. We assume that many workpieces have a mechanical gradient such as depicted in Fig. 3, which can be used to guide workpieces into alignment. The problem we address is the following: how to learn from human demonstration a task such that the *convergence region*, *i.e.* the set of starting poses from which alignment with a same set of parameters is successful, is maximized. Thus the uncertainties related to the relative pose between the workpieces to be aligned can be efficiently mitigated. Such uncertainties can rise from, for example, small modifications to the environment or simply the uncertainty of grasping an object. To clarify this even further, let us consider the situation in Fig. 3 and assume that the goal is to slide the

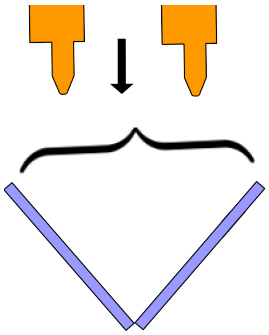


Fig. 3: Illustration of the theoretical convergence region (black brace) of the algorithm in a pure translational case [1].

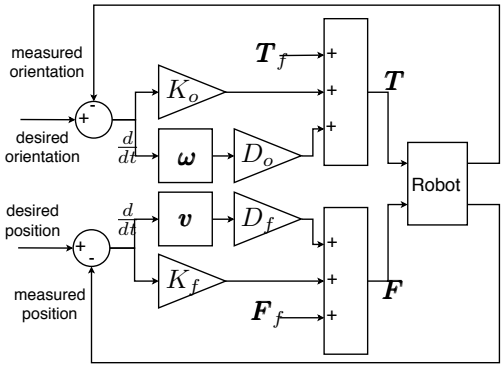


Fig. 4: An impedance controller with F/T feed-forward is used to reproduce the search motions [5].

tool to the bottom of the valley. The error in translation is the horizontal difference between the orange tools tip and the center bottom point of the valley. Now, if the tool is sliding directly downwards from anywhere within the convergence region and the tool is compliant perpendicular to this motion, it will slide along the surface all the way to the bottom.

We use impedance control to reproduce the motions. Impedance control is defined as

$$\begin{aligned} \boldsymbol{F} &= K_f(\boldsymbol{x}_d - \boldsymbol{x}) + D_f\boldsymbol{v} + \boldsymbol{F}_f \\ \boldsymbol{T} &= K_o(\boldsymbol{\beta}_d - \boldsymbol{\beta}) + D_o\boldsymbol{\omega} + \boldsymbol{T}_f \end{aligned} \tag{1}$$

where $\boldsymbol{F}, \boldsymbol{T}$ are the force and torque used to control the robot, $\boldsymbol{x}_d$ the desired position, $\boldsymbol{x}$ the current position, $\boldsymbol{\beta}_d$ the desired orientation, $\boldsymbol{\beta}$ the current orientation, K_f and K_o stiffness matrices and $D_f\boldsymbol{v}$ and $D_o\boldsymbol{\omega}$ linear damping terms. Parameters $\boldsymbol{F}_f$ and $\boldsymbol{T}_f$ are the superposed feed-forward force and torque, which can be used if additional implicit force on top of the standard impedance controller is required. The block diagram of the controller is shown in Fig. 4.

Thus, now if we analyse (1), we see two parameters on both translations and rotations that must be adjusted to provide this sort of behaviour: the stiffness matrices K_f, K_o and the desired position and orientation $\boldsymbol{x}_d, \boldsymbol{\beta}_d$. As observed, with efficient exploitation of compliance linear motions suffice to achieve assembly goals in scenarios such as in Fig. 3: thus, we can write the the desired position $\boldsymbol{x}_d$ and orientation $\boldsymbol{\beta}_d$ in a feed-forward manner as

$$\begin{aligned} \boldsymbol{x}_{d,t} &= \boldsymbol{x}_{d,t-1} + \nu\Delta t\hat{\boldsymbol{v}}_{\boldsymbol{d}}^{*} \\ \boldsymbol{\beta}_d, t &= \boldsymbol{\beta}_{d,t-1} + \lambda\Delta t\hat{\boldsymbol{\omega}}_{\boldsymbol{d}}^{*} \end{aligned} \tag{2}$$

where $\hat{\boldsymbol{v}}_{\boldsymbol{d}}^{*}$ and $\hat{\boldsymbol{\omega}}_{\boldsymbol{d}}^{*}$ are the desired directions describing the human teachers intended motion in translation and rotation, Δt the sample time of the control loop and ν and λ the translational and rotational speeds.

B. Learning the LMC primitive

In this section we explain the process of learning the parameters $\hat{\boldsymbol{v}}_{\boldsymbol{d}}^{*}$,$\hat{\boldsymbol{\omega}}_{\boldsymbol{d}}^{*}$, K_f and K_o such that the conditions described in the previous section are met. The data required for learning are the measured Cartesian poses of the end-effector and the corresponding wrenches measured by the Force-Torque sensor (F/T sensor) according to Fig. 5 from a human demonstration.

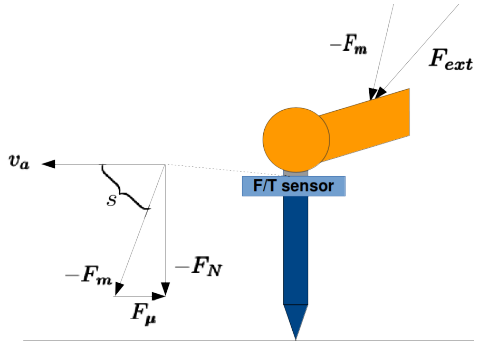


Fig. 5: Forces recorded by an F/T sensor when sliding along a surface. v_a is the actual velocity, F_{ext} the force applied by the teacher and s the sector of desired directions [1].

A general flow of the learning process is shown in Fig. 6: the pitch, *i.e.* the relation between mean translational speed of the demonstration ν and rotational speed λ, must be computed first from the raw data. After this, the same steps are taken for both translational and rotational motions: the first thing is to check if the teacher keeps either translations or rotations 3-Degree of Freedom (DoF) compliant, *i.e.* either position or orientation can change freely without affecting the execution of the task. Whenever this is not the case, the next thing is to check whether the teacher intentionally translated or rotated the tool, *i.e.* check the existence of $\hat{\boldsymbol{v}}_{\boldsymbol{d}}^{*}$ and $\hat{\boldsymbol{\omega}}_{\boldsymbol{d}}^{*}$. If either or both exist, the compliant axes must be deduced such that they are perpendicular to the desired direction, otherwise the compliant axes can be directly deduced from the data. Finally, if both $\hat{\boldsymbol{v}}_{\boldsymbol{d}}^{*}$ and $\hat{\boldsymbol{\omega}}_{\boldsymbol{d}}^{*}$ exist, ν and λ must be set according to the pitch, finalizing the LMC primitive.

The intuition into 3-DoF compliance can be observed from Fig. 7: the teacher tries to only rotate the tool, but due to contact forces, rotational force causes also translation in the wrist. In this case observed translation is caused completely by the environment and the corresponding degrees of freedom need to be set compliant (i.e. 3-DOF compliance). To numerically

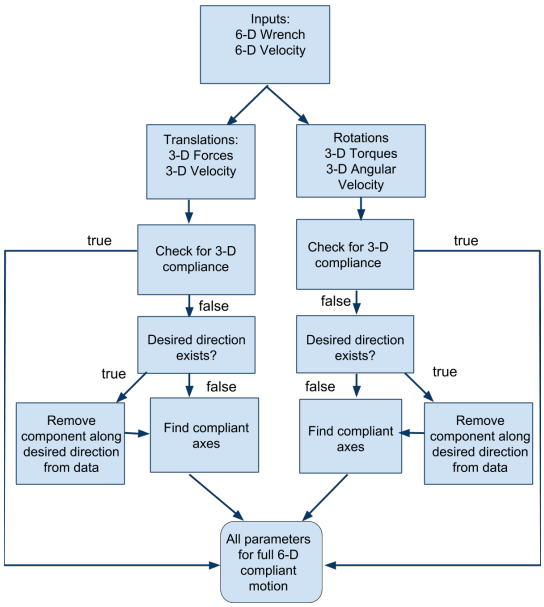


Fig. 6: A flowchart describing the whole process of finding the 6-D compliant primitive to reproduce a demonstrated motion [1].

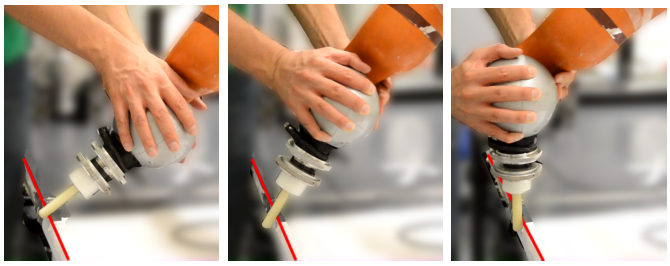


Fig. 7: A demonstration of rotating the peg around the edge of the table, where the teacher only rotates the tool and the translational motion in the wrist is caused by the contact forces. The edge of the table is highlighted in red [1]].

detect this, we looked whether more *work* was done by the environment or the teacher, where work in physics is defined as

$$W_x = \boldsymbol{F}_m \cdot \Delta \boldsymbol{x}$$
$$W_\beta = \boldsymbol{T}_m \cdot \Delta \boldsymbol{\beta} \quad (3)$$

where W is the work, $\Delta \boldsymbol{x}$ the change in translation, $\Delta \boldsymbol{\beta}$ the change in angle and $\boldsymbol{F}_m$ and $\boldsymbol{T}_m$ the force and torque measured by the F/T sensor. The idea is that when the work measured at an interval is positive, the environment has done the work since the forces and torques are caused by the environment. By comparing the amount of positive and negative work measured we can deduce whether the teacher or the environment did more work to move the tool.

The idea for learning the desired direction stems from Fig. 5, where the forces acting on a tool sliding along a surface are shown. The key is to observe sector s, which marks the 2-D sector between the actual direction of motion $\boldsymbol{v_a}$ and the

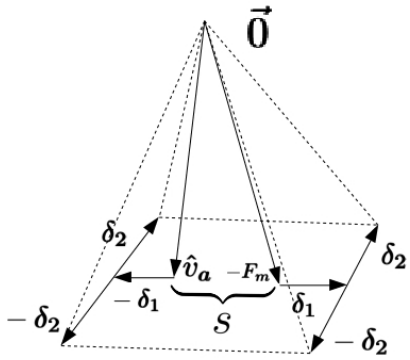


Fig. 8: Illustration of expanding 2-D sector s from Fig. 5 for translations into 3-D set of directions. Continuous lines represent the vectors and dotted lines highlight the pyramid shape [1].

negative of the force measured by the F/T sensor, $-\boldsymbol{F_m}$. Firstly, we observe that the width of this sector depends on the friction force $\boldsymbol{F_\mu}$ such that with higher friction force the sector s becomes more narrow. Secondly, if the external force, $\boldsymbol{F_{ext}}$ is applied from anywhere within sector s, the observed motion will be exactly the same, along the direction of $\boldsymbol{v_a}$. This gives us, at every time instant, a range of directions from which the robot can apply force to achieve a certain motion. We hypothesize that by computing an intersection of these sectors s during a demonstration consisting of sliding motions, we can find the desired direction $\hat{\boldsymbol{v}}_{\boldsymbol{d}}^*$ where the human intends to push the tool, and the same logic can be applied to rotations to find $\hat{\boldsymbol{\omega}}_{\boldsymbol{d}}^*$.

To transfer this intuition into 3-D, several steps are required. Firstly, there is always uncertainty in a humans demonstration, even if the teacher tries to draw a straight line. Thus, there is a risk that taking an intersection results in an empty set. In Fig. 8 is shown how we propose extending the sector s from Fig. 5 into a pyramid shape in 3-D. These pyramid shapes are then projected into 2-D and an intersection is calculated with suitable outlier rejection. If more than one demonstrations are supplied, the intersection is simply computed from a concatenation of the demonstrations.

The next step in Fig. 6 is finding the compliant axes. The first question in this process is, separately for translations and rotations, whether the desired direction was observed. If it was, this already reduces the dimensionality of the possible compliant axes by one, since the compliant axes must be orthogonal to desired direction as per Condition **??**. The key idea for finding the compliant axes for reproducing the observed motion from the remaining DoFs is the following: if motion along other directions besides the desired direction was observed, it must have been caused by the environment, and thus compliance is required to replicate the motion. Without a desired direction, we assume that all motion is caused by the environment. Thus, when the desired direction exists, we first subtract it from the raw motion data before advancing, and then the compliant axes are computed similarly for the

cases where desired direction exists and where it does not. To compute comparable values, we take inspiration from the Bayesian Information Criterion (BIC) [17] to choose the correct number of compliant axes.

C. Learning search motions

In Section II-A we made the assumption that there is a physical gradient, such as a chamfer, that can guide the robot's tool into the goal pose. However, it is often the case that this sort of guidance is not available and the environment surrounding the goal pose cannot be exploited to mitigate uncertainties. Even in such a case a human still can have task-dependent intuition of how to efficiently locate the goal– a human might, for example, use a different strategy for fitting a key into a lock than inserting an electric plug into a socket. It would be highly useful if this sort of task-dependent information could be directly conveyed to the robot from a human demonstration– in industry this sort of search could be used as an *exception strategy* in case of a failed task due to error in the environment model.

The existing work on exception strategies for assembly tasks is very limited. Abu-Dakka *et al.* [18] used random walk in case an assembly task failed and searching had to be done. Jasim *et al.* [19] used an Archimedean spiral, which is guaranteed to find the goal with the correct resolution and starting position. However, the spiral is limited to 2-D case and randomness is something to be used as a baseline for better methods. Kronander [20]Chapter 5 used incremental learning, where the human assists the robot during insertion if the robot gets stuck. However, none of these methods took advantage of any intuition a human may have on the tasks at hand.

The approach in this paper is to not assume any contact that can help guide the search, either for guiding or localization purposes, meaning that no earlier experience with this particular plug pose is expected and no visual or auditory sensor input is allowed. Gathering such demonstrations from a human is non-trivial, and we managed this by blindfolding the teacher and varying the relative pose between the start of the demonstrations and the goal. Essentially, there are two things we can learn from the human from such a demonstration without further assumptions: the area in which to search, and the dynamics of the in-contact motions. We call the area to search the *exploration distribution*, which we learn by fitting a Gaussian to the teacher's search trajectory. As the proposed method is for use in environments where the location information can be erroneous, we regard the information conveyed by the exploration distribution and recorded forces as more important than the starting location of the search. Thus we set each search into a common coordinate frame, which in this thesis is called the *search frame*. Furthermore we choose to align the demonstrations based on their origin, not the goal, even though in the world coordinate system they share the goal but not the origin. With this choice we can better learn the *search* strategy of the teacher in situations where localizing the tool w.r.t the world frame is impossible.

D. Learning Sequence of Motions

In this chapter we address segmenting a demonstrations into phases, each of which is used to learn a single LMC primitive, and then during execution a correct primitive must be chosen at a correct moment– a single primitive is often insufficient to encode a whole task, and thus the primitives must be *sequenced* to be useful in real-life tasks. We show that LMC primitives can be successfully combined to perform assembly skills, such as attaching hose couplers together. As LMC depends neither on time nor pose, the length of a single primitive does not need to be known beforehand, which brings more flexibility and error tolerance– however, the price of that flexibility is that an additional algorithm is required for learning to detect these changes.

Intuitive approaches for segmenting human demonstrations into phases are often simple heuristics, such as Zero-Velocity Crossing (ZVC) (*i.e.* a change of direction in velocity) or threshold values in contact force. However, this sort of simple heuristics are often not error-tolerant and have to be manually designed for each task. At the other end of spectrum in complexity are usage of multimodal inputs including vision, which can then yield impressive results in, for example, success detection in screwing [21]. However, we use strictly pose and wrench signals acquired from a demonstration since we do not want to depend on vision due to challenges in occlusion and accurate detection of contact.

In this paper the goal is to segment synchronized pose and wrench recordings from human demonstrating an assembly task into segments, each consisting of a single LMC primitive presented in Section II-A. Equations (2) and (1) defining LMC inspired us to model the state dynamics of a single phase by a linear Gaussian model, in which the next state depends on current state $\mathbf{s}$ (pose or a subset), measured interaction (wrench or a subset) $\mathbf{a}$ concatenated with value 1, and current phase $\boldsymbol{\rho}$, where each phase consists of a single LMC– an example of what a phase can look like is seen in Fig. 2. Thus, we write $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \boldsymbol{\rho}_t)$. The distribution of the next state is then

$$\mathbf{s}_{t+1} \sim \mathcal{N}\left(A_{\boldsymbol{\rho}_t}\mathbf{s}_t + B_{\boldsymbol{\rho}_t}\mathbf{a}_t, \Sigma_{\boldsymbol{\rho}_t}\right) \tag{4}$$

where $A_{\boldsymbol{\rho}_t} \in \mathbb{R}^{m\times m}$ represents the uncontrolled dynamics, $B_{\boldsymbol{\rho}_t} \in \mathbb{R}^{m\times d}$ models compliance through interaction forces and constant offset velocity through the concatenated 1, and $\Sigma_{\boldsymbol{\rho}_t} \in \mathbb{R}^{m\times m}$ is the covariance matrix corresponding to phase $\boldsymbol{\rho}_t$. Thus, the model assumes linear system dynamics and $B_{\boldsymbol{\rho}_t}$ is used to model contact interaction effects and constant desired direction of motion. This model is used for detecting the correct phase, but the actual control is performed with LMC.

An inspiration for the segmenting approach of this work was the work of Kroemer *et al.* [22], who use an autoregressive state space model together with an Hidden Markov Model (HMM) to segment a demonstration: essentially, their idea is that each phase depends on the previous phase and previous state, *i.e.* $p(\boldsymbol{\rho}_t|\mathbf{s}_t, \boldsymbol{\rho}_{t-1})$. However, as we are modeling compliant in-contact motions while assuming pose uncertainties,

depending on the pose for phase changes is not desirable. Thus, we use the graphical model depicted in Fig. 9 with $p(\boldsymbol{\rho}_t|\mathbf{a}_t, \boldsymbol{\rho}_{t-1})$ such that each phase depends on the previous phase and previous action to better take into account the compliant nature of the task being modelled. For learning the model, we adapt the Expectation-Maximization (EM) algorithm from [22] to learn from multiple demonstrations the model parameters $\theta = (w, A, B, \Sigma)$. During reproduction, we take advantage of the joint probabilities to choose the correct primitive.

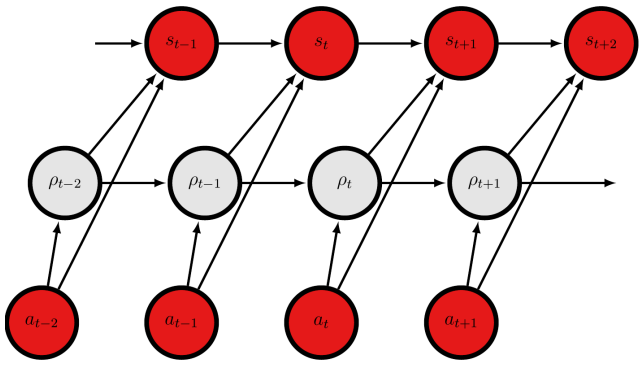


Fig. 9: The graphical model we used for segmenting the demonstration [4].

III. EXPERIMENTS AND RESULTS

We tested our approaches on various tasks and setups. On the hose-coupler setup in Fig. 12 we performed experiments on the LMC primitive and combining primitives. On the peg-in-hole setup shown in Fig. 15 we tried the LMC primitive both on single and dual arms and also the search. Additionally, we performed search experiments on a plug-and-socket setup shown in Fig. 10 and used LMC with a heavy-duty hydraulic manipulator shown in Fig. 11.

In the hose-coupler setup we defined both the Tool Center Point (TCP) and the Center of Compliance (CoC) in the flange of the robot to achieve rotational compliance around the flange and to observe the translations occurring at the flange when the orientation of the tool changes. In this task there is a high likelihood of orientation error when commencing the task due to difficulties in pose estimation, with examples shown in Fig. 12. We showed that with one desired direction and a correctly identified stiffness matrix the hose couplers can be aligned with the same set of parameters starting from both Figs. 12a and b and ending up in Fig. 12c after two demonstrations from different starting positions.

In the hose-couple alignment task, for translations a desired direction is found, but for rotations it is not– visualization of the rectangles representing the limits of desired direction at each time interval are shown in Fig. 13, where the red rectangles are from a demonstration starting from the pose of Fig. 12a and the blue ones from Fig. 12b. It can be observed that for translations the rectangles from both demonstrations are aligned, but for rotations the two demonstrations are clearly

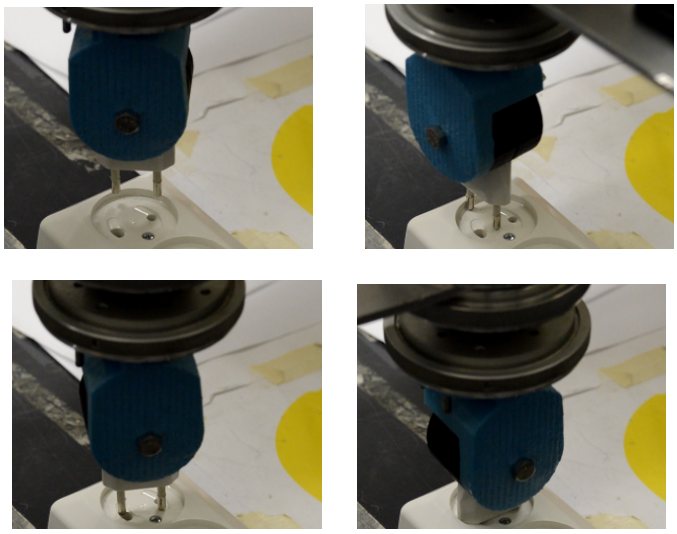


Fig. 10: An example sequence of a robot inserting a plug into a socket without vision sensing [5].

separate, leading to the conclusion that a desired direction for translations is required but for rotations there is not a desired direction.

Finding the number of compliant axes is visualized in Fig. 14, where each blue cross represents the mean directions of motion of a demonstration and the red axes are the axes of Principal Component Analysis (PCA) performed on all mean directions of a demonstration. Since for translations there exists a desired direction, it is plotted in cyan (overlapping the first PCA axis, as expected) and subtracted from the mean direction of motion data, *i.e.* the blue crosses are projected into the plane of the other principal axes, resulting in the green crosses. Now it can be observed that one of the principal components connects the green crosses, thus explaining the observations and resulting in choosing one compliant axis along that component. As the TCP was set in the flange, translation is required to perform the alignment. For rotations, the analysis is done directly on the PCA data, as there is no desired direction. It can be observed that the rotations are close

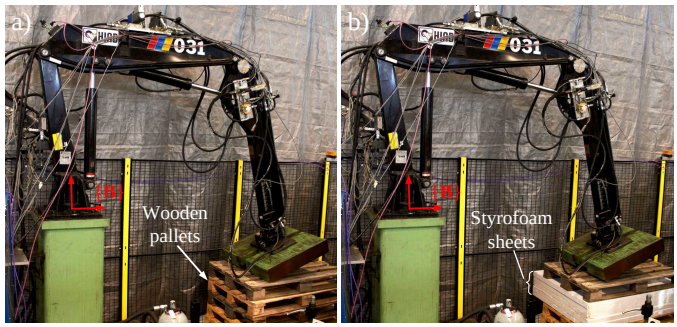
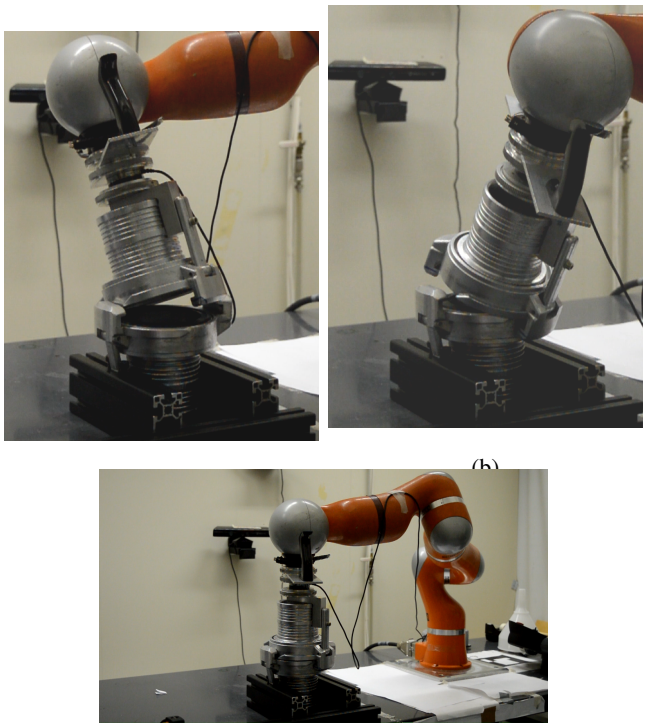
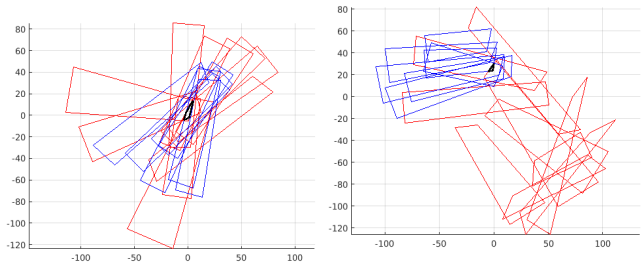


Fig. 11: a) Experiment setup with wooden pallets. b) Experiment setup with styrofoam sheets and wooden pallets. The manipulator's position in the figures show the starting point of the test trajectories (same starting position in the both cases) [7].



(c)

Fig. 12: Two possible starting poses and the final pose of the hose-coupler alignment task [1].

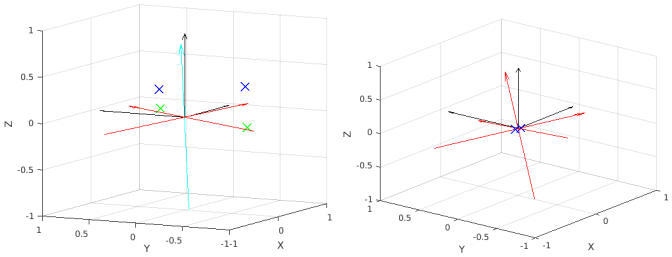


(a) Translations (b) Rotations

Fig. 13: Visualization of finding the desired direction, shown for translations and rotations of the hose-coupler alignment task. The red and blue colors indicate the two separate demonstrations of the task and the black rectangle is the intersection, the set of all desired directions in the projection coordinate

to the origin, but still far enough that one compliant axis was detected, as required to align the tools.

We experimentally verified that we can successfully reproduce the alignment motions. Additionally, we showed successful learning and reproduction of a peg-in-hole task with a varying starting orientation error. Screenshots from a reproduction are shown in Fig. 15. Moreover, we also showed that this primitive can be successfully used with teleoperated demonstrations, which are shown to be noisier than by kinesthetic teaching [23] with a heavy-duty hydraulic



(a) Translations (b) Rotations

Fig. 14: Illustrations of choosing the directions of compliant axes on the hose-coupler alignment experiment. The black arrows are coordinate axes, the red ones the eigenvectors U, the blue crosses the average motions of each demonstration and the green crosses their projections to the first principal component. In (a) the desired direction is plotted in cyan (overlapping the third eigenvector as expected). In both (a) and (b) 1 compliant axis is chosen [1].

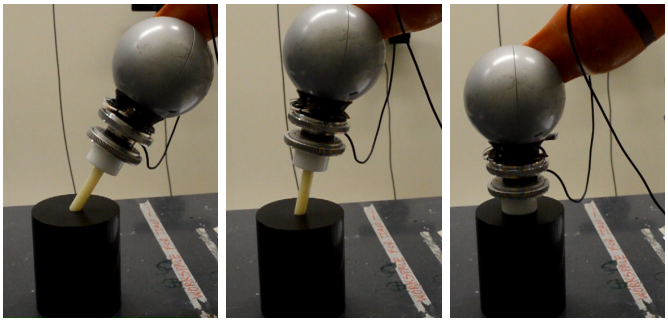
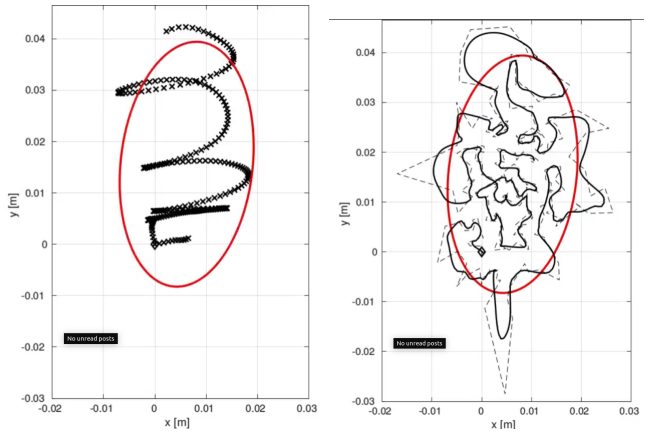


Fig. 15: Screenshots from a reproduction video of the P-I-H motion. The motion starts from the leftmost picture, and the peg is rotated and pushed to the bottom. The peg has radius 16.5 mm, length 80 mm and a rounded tip, and the hole's radius is 0.25 mm more than the peg's [1].

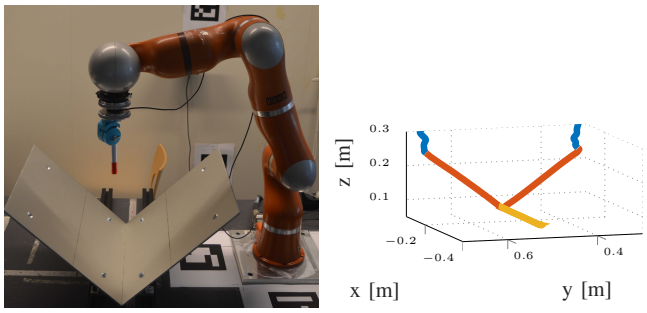
We performed the search motions on the peg-in-hole setup with 85% accuracy and the plug-in-socket task with 67% accuracy, which we consider good considering the difficulty of the tasks (essentially a near-blind search in 2-D or 3-D). In Fig. 16a is shown how the exploration distribution is learned from human demonstration, and in Fig.16b how a search trajectory is created from the exploration distribution by sampling.

We tested segmenting and sequencing of motions on both the hose-coupler setup (Fig. 12) and valley setup seen in Fig. 17a. In the hose-coupler setup, the algorithm correctly identified lowering the coupler as one LMC phase and interlocking the couplers as another and reproduction was successful. In the valley setup, the algorithm correctly identified that sliding down either side is the same phase, as seen in Fig. 17b, thus showing that the robot learned to take advantage of the guidance of either chamfer.



(a) Demonstration and exploration distributions. (b) Exploration distribution and search trajectory.

Fig. 16: Visualizations of creating a search trajectory from one or more human demonstrations.



(a) (b)

Fig. 17: The physical valley setup (a) and the phases learned from a demonstration of sliding to the bottom of the valley and then towards the camera (b).

IV. Conclusions

We successfully showed that we can learn from human demonstrations various tasks requiring compliance. The results from [2]- [7] can be used to greatly advance the usage of robots in SMEs by three very important factors: firstly, the usage of LfD makes teaching the robot new tasks easy and efficient, thus allowing the robot to perform varying tasks when production batch sizes are small. Secondly, by the use of compliance, small changes in the workplace due to *e.g.* vibrations may not cause the task to fail. Thirdly, even if the task fails, if a proper exception strategy is learned with the search, the robot can recover even from errors by itself and carry on it's task without need of an employee to re-teach everything. We believe that these results have the potential to significantly boost the usage of robots in Finland.

References

[1] M. Suomalainen and V. Kyrki, "Learning 6-d compliant motion primitives from demonstration," *Autonomous Robots*, 2019, submitted. arXiv:1809.01561.

[2] M. Suomalainen and V. Kyrki, "Learning compliant assembly motions from demonstration," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 871–876, IEEE, 2016.

[3] M. Suomalainen and V. Kyrki, "A geometric approach for learning compliant motions from demonstration," in *Humanoid Robots (Humanoids), 2017 IEEE-RAS 17th International Conference on*, pp. 783–790, 2017.

[4] T. Hagos, M. Suomalainen, and V. Kyrki, "Segmenting and sequencing of compliant motions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6057–6064, IEEE, 2018.

[5] D. Ehlers, M. Suomalainen, J. Lundell, and V. Kyrki, "Imitating human search strategies for assembly," *2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019, Accepted for Publication. arXiv:1809.04860.

[6] M. Suomalainen, S. Calinon, E. Pignat, and V. Kyrki, "Improving dual-arm assembly by master-slave compliance," in *2019 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, Accepted for Publication.

[7] M. Suomalainen, J. Koivumäki, S. Lampinen, J. Mattila, and V. Kyrki, "Learning from demonstration for hydraulic manipulators," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3579–3586, IEEE, 2018.

[8] A. Ude, "Trajectory generation from noisy positions of object features for teaching robot paths," *Robotics and Autonomous Systems*, vol. 11, no. 2, pp. 113–127, 1993.

[9] J.-H. Hwang, R. C. Arkin, and D.-S. Kwon, "Mobile robots at your fingertip: Bezier curve on-line trajectory generation for supervisory control," in *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1444–1449, IEEE, 2003.

[10] S. Schaal, "Dynamic movement primitives-a framework for motor control in humans and humanoid robotics," in *Adaptive Motion of Animals and Machines*, pp. 261–280, Springer, 2006.

[11] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 286–298, 2007.

[12] S. M. Khansari-Zadeh and A. Billard, "Learning stable nonlinear dynamical systems with gaussian mixture models," *IEEE Transactions on Robotics*, vol. 27, no. 5, pp. 943–957, 2011.

[13] J. Lundell, M. Hazara, and V. Kyrki, "Generalizing movement primitives to new situations," in *Conference Towards Autonomous Robotic Systems*, pp. 16–31, Springer, 2017.

[14] M. Hazara and V. Kyrki, "Model selection for incremental learning of generalizable movement primitives," in *Advanced Robotics (ICAR), 2017 18th International Conference on*, pp. 359–366, IEEE, 2017.

[15] M. Muhlig, M. Gienger, S. Hellbach, J. J. Steil, and C. Goerick, "Task-level imitation learning using variance-based movement optimization," in *2009 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1177–1184, IEEE, 2009.

[16] N. Figueroa and A. Billard, "A physically-consistent bayesian non-parametric mixture model for dynamical system learning," in *Proceedings of The 2nd Conference on Robot Learning*, vol. 87 of *Proceedings of Machine Learning Research*, pp. 927–946, PMLR, 2018.

[17] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[18] F. J. Abu-Dakka, B. Nemec, A. Kramberger, A. G. Buch, N. Krüger, and A. Ude, "Solving peg-in-hole tasks by human demonstration and exception strategies," *Industrial Robot: An International Journal*, vol. 41, no. 6, pp. 575–584, 2014.

[19] I. F. Jasim, P. W. Plapper, and H. Voos, "Position identification in force-guided robotic peg-in-hole assembly tasks," *Procedia Cirp*, vol. 23, pp. 217–222, 2014.

[20] K. J. A. Kronander, *Control and Learning of Compliant Manipulation Skills*. PhD thesis, EPFL, 2015.

[21] Z. Su, O. Kroemer, G. E. Loeb, G. S. Sukhatme, and S. Schaal, "Learning manipulation graphs from demonstrations using multimodal sensory signals," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2758–2765, IEEE, 2018.

[22] O. Kroemer, H. Van Hoof, G. Neumann, and J. Peters, "Learning to predict phases of manipulation tasks as hidden states," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4009–4014, IEEE, 2014.

[23] K. Fischer, F. Kirstein, L. C. Jensen, N. Krüger, K. Kukliński, T. R. Savarimuthu, *et al.*, "A comparison of types of robot control for programming by demonstration," in *Proceedings of the 2016 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 213–220, 2016.

Timo Malm*, Timo Salmi, Ilari Marstio and Iina Aaltonen

Are collaborative robots safe?

Abstract: Collaborative robots (cobots) market is supposed to increase rapidly. Although cobots are safer than their ancestors, industrial robots, there are some safety issues. The risk assessment for cobots is difficult since the cobots are working beside humans. In most of the applications the cobots can be safe. Sharp tools or objects and risk of impact to head are excluded from the common applications due to their risks. The common idea is to apply power and force limitation to ensure safe performance of cobots. The speed of the robot has huge effect on the impact force. Speed reduction with adequate safety controls can decrease the impact force to acceptable level in most of the applications. If there are severe risks in the robot cell, then it is possible to apply adequate separation distance between human and robot or safety-rated monitored stop to ensure safety and relatively close vicinity between human and the robot. One issue is that the stopping performance is complex and currently it is difficult to predict. Quite often, the validation requires impact force/pressure measurements or impact modelling.

Keywords: functional safety, collaborative robots, safety requirements, levels of collaboration

***Corresponding Author: Timo Malm:** VTT,
P.O. Box 1300, FI-33101 Tampere, Finland
E-mail: timo.malm@vtt.fi

Timo Salmi, VTT
E-mail: timo.salmi@vtt.fi

Ilari Marstio, VTT
E-mail: ilari.marstio@vtt.fi

Iina Aaltonen, VTT
E-mail: iina.aaltonen@vtt.fi

1 Introduction

Collaborative robots (cobots) have been under discussion for some years. They have some properties, which make it possible to work safely beside the robot. They are expected to open up new possibilities for flexibility, productivity and user friendliness. Also fenceless production cells are often mentioned.

Currently the world market of cobots is about 600 million € and at 2027 the market is supposed to be 7500 million € [1]. Collaborative robots are typically small and their reach is usually below 1.3 m and due to the size, their applications are often related to handling of small size objects. However, new applications are expected to appear.

The industrial robots are typically stronger, faster and more accurate than collaborative robots, but also cobots have some advantages. According to Kildal et al., the expectations of cobots are often fulfilled and in many cases, they are easier to program [2]. According to Bender et al., increased operational efficiency is the most frequent reason to choose a cobot. However, the payback time requirement for cobots is longer than for industrial robots. Very often cobots are evaluated by applying other than monetary terms. The most important values have been related to ergonomics, quality and flexibility. [3]

One advantage of the collaborative robots is that, usually, they are easier to program and the robot workspace does not have as many objects as the workspace of an industrial robot. On the other hand, collaborative robots are used in applications, which change more often than industrial robot applications. Continuous changes make it challenging to maintain adequate level of safety.

Although cobots are safer to use than industrial robots, according to Kildal et al. common opinion is that risk analysis is more difficult to make for cobots [2]. It can be understandable, since cobots are used in applications where human and robot are close to each other and it makes the risk assessment more difficult.

One special difficulty is related to force limiting of cobots, since the validation of force limits is difficult and the impact effect depend on the situation and how a person feels an impact. According to measurements at VTT a change of a parameter, especially, speed has an effect on impact force. Also Kirchner et al. points out that a parameter change of the robot cause unpredictable impact forces for many collaborative robots. Measurements can reveal the impact forces [4]. The aspect how a person feels an impact is difficult since Machinery Directive (2006/42/EC) is partly dedicated to enable trade by declaring uniform safety requirements. From the trade point of view, there should be applicable safety limits. The aspect how a person feels an impact is related more to user organization and related requirements (Work

Equipment Directive 2009/104/EC).

This text aims to point out factors, which need to be considered to estimate the safety of a collaborative robot. Many safety features have conditions, when their performance is adequate according to safety standards. Section 2 describes the collaboration of human and the robot. Section 3 points out safety requirements, which are important for cobots. Section 4 points out safety issues and measures, which are important for cobots. Section 5 presents safety design process model for collaborative robots. The process model is focusing on impact hazards and related safety measures. The idea of the process model is to point out factors, which have remarkable effect on safety measures to be used. Section 6 shows remarks related to safety of cobots.

2 Collaboration

Collaboration between human and robot can be realized in many ways and the separation distance, collaborative workplace locations and applied forces can vary from one application to another. The safety of collaboration can be based on inherently safe structures, guards, sensors, motion control, safe procedures and functional safety.

The ISO 10218-2 standard presents conceptual applications of collaborative robots, which are [5]:

- Hand-over window. Autonomous operation, reduced speed near the window, fixed or sensitive guards
- Interface window. Autonomous operation, except at the interface window the robot stops, fixed or sensitive guards, hold-to-run control.
- Collaborative workspace. Autonomous operation, person detection system, reduced speed according to distance.
- Inspection. Autonomous operation, person detection system or enabling device, reduced speed according to distance
- Hand-guided robot. Moving by hand guiding, hold-to-run control, reduced speed according to distance.

The above list does not show applications, which applies power and force limitation as a safety measure. Force limitation could be applied at any of the mentioned conceptual application and the collaborative workspace could allow more intensive collaboration.

Aaltonen et al. present four levels of collaboration. Here are presented the levels and comments how separation and speed control, like the dynamic safety system can be related to the level. [6]

- No coexistence: physical separation.
- Coexistence: human works in (partially or completely) shared space with the robot with no shared goals.
- Cooperation: human and robot work towards a shared goal in (partially or completely) shared space.
- Collaboration: human and robot work simultaneously on a shared object in shared space. Physical contact is allowed, possibility for hand-guiding.

The level of safety depends on the level of collaboration, due to the exposure time and separation distance. If there is no coexistence, the risk for a person is not so high since the person is not exposed to danger. It is more difficult to say the difference of the risk between the three other collaboration levels, although collaboration seems more risky due to obvious vicinity of the robot.

3 Safety requirements

Most of the collaborative robots are designed according to inherently safe principles. The collaborative robots are designed so that they should not exceed the defined force, at least with slow speed. In old robot safety standard (ISO 10218-1:2006) there has been a general force limit (150 N), but now the limit is specific for each body part of the human according to ISO TS 15066 [6]. The power and force limiting, brings new kind of thinking, since the contact is now a designed feature and not just a rare mishap. The designer needs to estimate, which body parts can be exposed to an impact of the robot and then limit forces accordingly.

The collaborative operations apply at least one of the means: safety-rated monitored stop, hand-guiding, speed and separation monitoring or power and force limiting by inherent design or control. The means are described more at the section 4.

One issue is that according to ISO 10218-2 section 5.2.2 safety related parts of the robots must comply with PL d and Cat 3 requirements of ISO 13849-1 [7]. This is related, among others, to stop, speed, area, power and force control. Many of the current robots do not comply with the requirements and therefore one have to consider, can e.g. a speed limit be applied to guarantee safety.

The ISO 13849-1 standard is related to functional safety of safety functions and associated control systems. The requirement levels are associated Performance Levels (PL) from “a” to “e” (highest level). There are both qualitative requirements, for example

associated to software, and quantitative requirement i.e. the average probability of dangerous failure per hour. For PL d the probability value have to be below 10^{-6}. The factures, which affect the calculation of average probability of dangerous failure per hour (PFH_D) are: Mean Time to Dangerous Failure ($MTTF_D$), designated architecture (Category) and Diagnostic Coverage (DC). In addition, Common Cause Failures (CCF) are considered, but only to fulfil adequate requirements, not to for calculation. The category 3 is associated to duplicated structure (two channel system, one-out-of-two, 1oo2, hardware fault tolerance = 1), which is able to reveal single dangerous failures, since the duplicating part can perform the safety function, if one channel fails. The category 2 has also redundancy, but instead of duplication, diagnostics and alarms are applied to ensure safety. The two-channel system or failsafe (single fault safe) structure is considered more reliable than one-channel system with diagnostic, although the probability of dangerous failure per hour can be the same. The reason is that when applying two-channel system, the input data is not so sensitive to mistakes i.e. PL d can be achieved with lower $MTTF_D$ and DC values.

Fig. 1 shows safety measures and related standards according to ISO 10218-2. The basic design is made according to ISO 12100 [10] and then guards, safety distances and protective devices are selected and designed according to the relevant requirements.

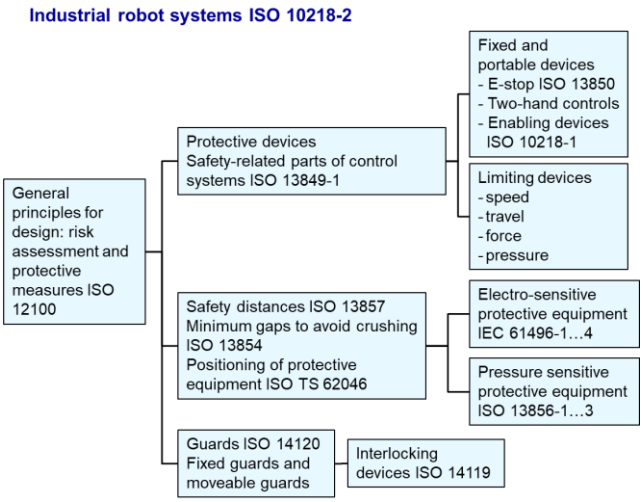


Fig. 1. Safety measures and standards of ISO 10218-2.

4 Safety issues and measures

Collaborative robots can be safer than heavy industrial robots. However, the collaborative robots are used in applications, where human and robot are close to each other and there is a high probability of impact. The intention is that human and robot work together. If the robot has sharp or dangerous tools, then similar safety measures as applied in industrial robot cells, are required. The ISO 10218-1 [9] describe the safety measures to collaborative robots from which at least one measure must be chosen. The measures are: safety-rated monitored stop, hand-guiding, speed and separation monitoring and power and force limiting by inherent design or control.

4.1 Safety-rated monitored stop

The stopping of the robot is monitored continuously and unauthorized movement case protective stop, which cut the power from servomotors. Protective stop resembles emergency stop (which is initialized by a person) and it provides, typically, safe stopping performance, but restarting requires manual start-up. Safety-rated monitored stop does not require manual restart, if the start-up can be made safely and persons at the restricted area are detected. Some robots have requirements fulfilling internal monitoring system, but monitoring can be done also by applying external monitoring system.

4.2 Hand-guiding

The robot is operated by applying controls near the end-effector. The controls include also emergency stop and enabling device. The robot applies safety-rated monitored speed.

4.3 Speed and separation monitoring

The position of the robot and humans are measured and speed is controlled according to the separation distance. The separation distance is calculated according to ISO 13855 (or ISO TS 62046), takes into account: stopping time of the robot, delays related to detection, communication and action, human speed, human reach towards danger point and uncertainty related to accuracy [11]. The speed can be reduced down to zero to avoid impact hazard. The robot should have safety-rated monitored speed function in order to realize flexible solution. Without safety-rated monitored speed reduction the separation distance would be long, i.e. often over 3 m, depending on the stopping time.

4.4 Power and force limiting by inherent design or control

Some collaborative robots have power and force limiting, which is based on either lightweight construction and/or quick, requirements fulfilling impact detection and stopping. If the load is heavy, also speed reduction is needed to fulfil the allowed max.

force/pressure limits. The force limits of the robot are difficult to realize accurately, since so many factors affect the result (e.g. speed, axis, horizontal/vertical movement, load). Due to complexity of the robot stopping performance, measurements are currently the only way to verify the achieved force limits related to quasi-static impacts (clamping) [4]. Transient impact forces (open space) are difficult to measure and therefore the forces are verified by applying calculation model. An example model is presented at ISO TS 15066. In the model an average hand is applied and for smaller hand, transient force would be smaller than the model value, since due to smaller mass the impact is more flexible (recoil).

In addition to the impact force limit values, there are also pressure limit values. Pressure values are difficult to measure and model, since it is difficult to estimate the effective impact area of the robot and the human. The impact area 1 cm^2 gives roughly similar speed limit values as force limit calculations. Smaller impact area (sharp edge) can give very high pressure values and therefore very small speed limit values.

4.5 Speed effect on impact force

Speed reduction reduce effectively the impact forces caused by the robot. Fig. 2 shows how speed affect the impact force in transient fully inelastic collision to hand and chest according to calculation model. Fig. 2 shows also that an impact with similar force is achieved for hand with higher speed than for chest. This means that when fulfilling the force limit requirements an impact to hand may be done with higher speed than for chest. One may conclude also that higher speeds can be applied if an impact only to hands is relevant. The robot weight in the model is 100 kg (heavy robot) and load 10 kg. Calculation is made by applying energy-based calculation model of ISO TS 15066 (equation A.5). In reality, some energy turns to heat and therefore the model gives a slightly pessimistic value. On the other hand, inelastic impact gives lower values than elastic impact, but partly elastic impact would be hard to calculate or compare to real impacts. The impact to hand is more flexible (reduced mass is small in the model) than to chest and therefore the force values for chest are higher at the same speed. For both hand and chest, the allowed limit force is 140 N. The heavy 100 kg robot is chosen because the values are almost the same for heavier robots and therefore it represents a worst-case scenario. For smaller robots, the mass and load affect the model results strongly. Fig. 3 shows that the robot speed 1.3 m/s or below cause acceptable impact force to the hand. For smaller robots, the speed value is a little bit higher. The maximum speed of cobots (tool centre point) is usually close to 1.5 m/s, depending on the applied tool. If the impact is to the chest then the speed need to be below 0.4 m/s. To fulfil the impact force requirements, reduced speed is required if a person is working so close to the robot that it could hit the human body.

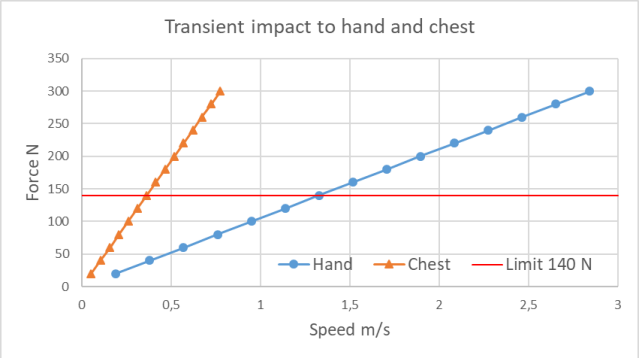


Fig. 2. Figure shows how similar transient impact force is achieved with higher speed for hand than for chest.

4.6 Other safety measures and issues

In addition, the mentioned obligatory measures, it is possible to use external tactile sensors or soft pads to meet the force limit requirements.

One safety issue related to collaborative robots is that they are applied in applications, which change often. This means that also risk assessment should be made often and it may be difficult to find new risks, if only quick risk assessment is done.

One aspect to be considered is that the robots may have exceptions to fulfil requirements. The robot system designer must check from the manual the conditions, when the robot can fulfil safety requirements. Following aspects are examples of conditions stated by manufacturer of the cobot:

- some conditions may hinder detecting impacts (e.g. specific positions, slow speed),
- there can be stopping performance parameters (e.g. stopping category and performance level)
- high speed or force limits may cause different safety requirements (e.g. overriding),
- control of singularity points may require additional tasks to maintain safety and
- acceleration/deceleration parameters affect stopping performance (slower deceleration can increase impact force).

One obvious issue is the applied tools. A sharp tool is usually dangerous and the robot work area may have corners or other machines, which cause potential hazard if human body part is clamped against it. In addition, grippers may be hazardous, but there are also models, which take into account the human presence.

5 Safe design process model for collaborative robots

Safety design process for collaborative robots (Fig. 4) is part of machinery safety design process (see Fig. 3). According to the design process, first risk assessment is applied to find out, which parts of the machine need safety measures. Basically, risk assessment is required to identify risks. Risk identification is made by applying, usually, hazard list of ISO 10218-2. The next phase risk estimation can be made according to harmonized standard, if the risk is described there. If the risk differs from the harmonized standard, then risk estimation and evaluation need to be done and documented carefully. Support to risk evaluation and reduction (safety measures) can be found in the safety design process for collaborative robots (Fig. 4). In addition, the safety measure can be selected by applying, for example, Machinery Directive (mandatory requirements), other standards and state-of-the-art knowledge. Arguments are needed to prove the solution, which is not according to the relevant harmonized standard.

Here in the collaborative robot design model risks are related to impact, clamping, shearing and stabbing. Other risks are considered by applying robot safety standards (ISO 10218-1 and ISO 10218-2). Risk reduction is made first by removing risk by applying inherently safe design, secondly by safeguarding and thirdly informing user about the risks [10]. The inherently safe design means, usually, selecting and using so small collaborative robots that they cannot hurt human. Robot selection is not here part of the process, but it is made before the collaborative robot design process (Fig. 4). The safety design process for collaborative robots is related mainly to safeguarding, which includes safety function evaluation and control and limitation of power, force, speed, stopping and area, which can be associated to Fig. 5 and phase 6 of Fig. 4. External safety devices are related to phase 3 and additional measures like enabling devices to phase 5 of Fig. 4 and Fig. 7.

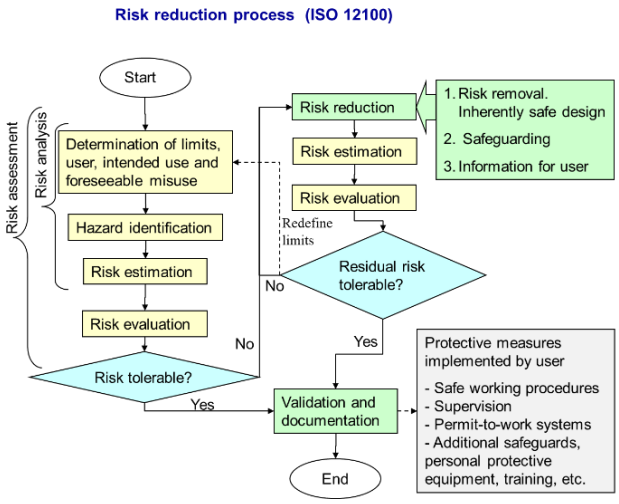


Fig. 3. Safety design process according to ISO 12100. [10]

Fig. 4 describes safety process for helping risk evaluation and reduction i.e. selecting safety measures. Fig. 5 presents the safety process related to internal safety functions. Light green means question and the track branches to two tracks (Yes/No). Light blue colour refers to action and other colours refer to specific colour coded phases. Here are explanations to the numbers/phases related to the figures:

1. Beginning of the process. There is a collaborative robot, with safety functions i.e. robot for the application is already selected. In addition, risk analysis is already made for the robot cell. First, consider impact to the head and are there sharp edges or tools, which cause hazards.
2. Do the safety functions fulfil the ISO 10218-2 section 5.2.2 requirements (PL d and Cat 3)?
3. If internal safety functions are not adequate, then apply external safety devices. These can be related to e.g. dynamic safety system, external tactile sensors, external safety-rated monitored stop or area restrictions and isolation (see Fig. 6).
4. Use PL assignment (risk assessment) for the application to see, if it gives lower requirement than PL d (see Fig. 8).
5. Can additional measures justify e.g. PL d, Cat 2. After phase 5 return, back to previous question, and furthermore to relevant phase (see Fig. 7).
6. Internal safety functions can be applied, if they fulfil safety requirements. Internal safety functions are related to e.g. impact forces, restricted area, speed or safety-rated monitored stop (see Fig. 5).

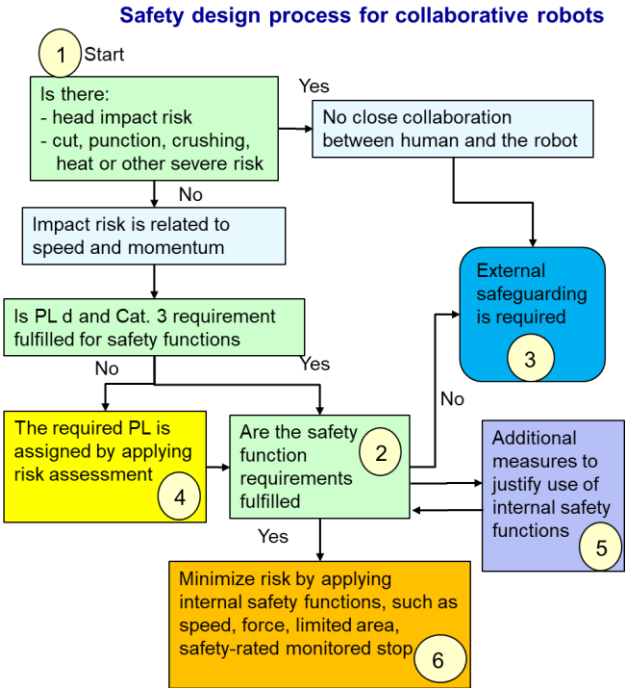


Fig. 4. Basic safety process for collaborative robots to select safety measures.

Fig. 5 shows the situation when Fig. 4 process leads to phase 6, which is related to internal safety functions.

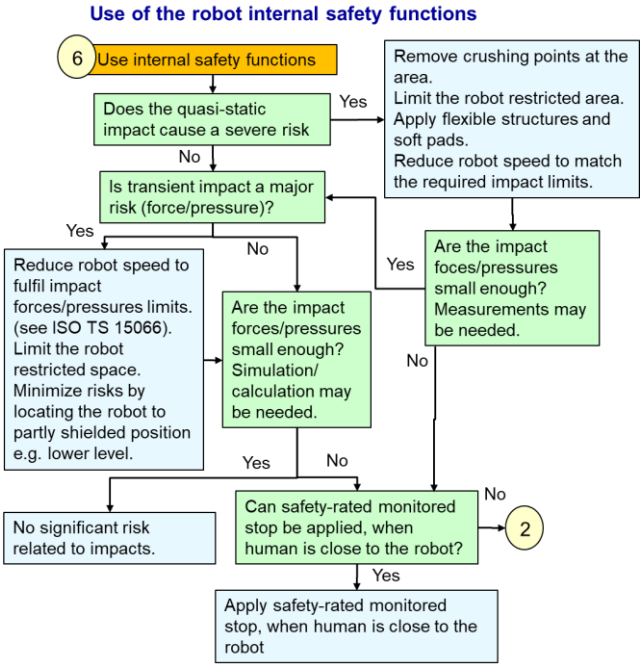


Fig. 5. Safety process for collaborative robots related to mainly internal safety functions (phase 6).

Fig. 6 shows external measures, which can be used to ensure safety. It is associated to phase 3 of Fig. 4. The measures can be valid also for industrial robots, except for external tactile sensors. Separation distance monitoring requires typically safety system, which monitors both robot and human movements, with adequate safety measures. An example of safety system is VTT dynamic safety system [12, 13]. Isolation of the robot restricted space is typical means for industrial robots and it provide only limited collaboration between human and the robot.

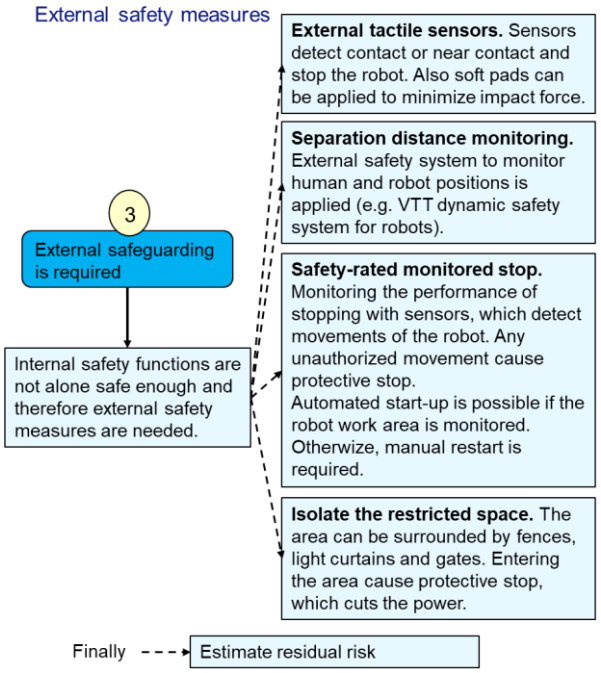


Fig. 6. External safety measures for the robot (phase 3).

Fig. 7 describes additional measures to reduce risk in order to decrease safety requirement level for the primary safety function. This phase is applicable only if the risk reduction need is small. In this phase, electronic safety functions are not applied, since PL is defined before safety functions can be applied. If safety functions are applied in this phase, they need to fulfil safety requirements of ISO 10218-2 (PL d and Cat. 3). Enabling device must fulfil the requirements stated at ISO 10218-1 [9].

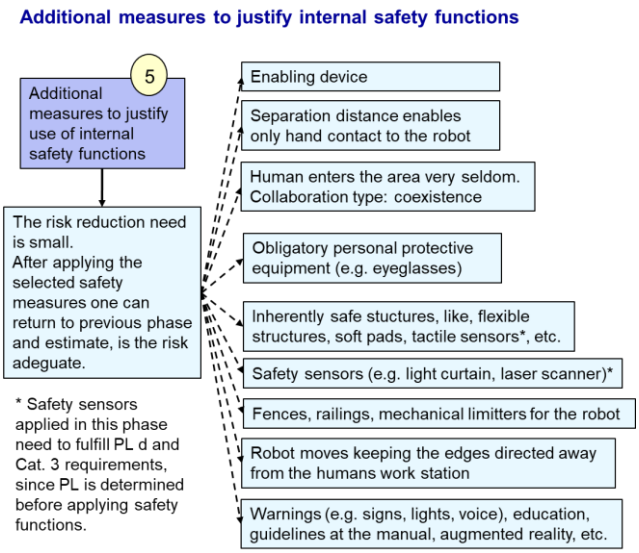


Fig. 7. External safety measures for the robot (phase 5).

Usually, the PLs of ISO 10218-2 are applied for safety functions. They are valid for typical robot applications. However, in some applications, the risks can be different and the PLs need to be reassigned. In practice, it means that severity is low and the robot cannot hurt a person. In phase 4, some severe risks are already ruled out and they are dealt at phase 3. Phase 4 is described at Fig. 8. In this phase the risk graph of ISO 13849-1 is applied to assign the PL. It would be possible to apply other functional safety standards in assigning PL or SIL, but apparently, ISO 13849-1 is here the most applicable. In some cases, there are other machinery standards, which gives performance level for specific safety functions (e.g. stability), but then one should consider how well they are applicable for the robots. After phase 4 one need to return back to phase 2.

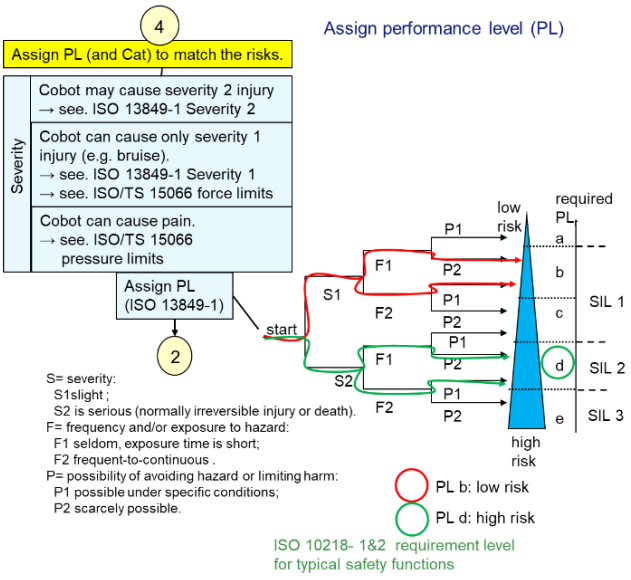


Fig. 8. Assign performance level (PL) (phase 4).

6 Discussion

It was mentioned already at the introduction that risk assessment is difficult for cobots, since close collaboration between human and robot is expected. It is relatively straightforward to isolate a system, but when safety-rated monitored stop or separation distance is applied, then also functional safety requirements are essentially relevant. When power and force limitation is applied, then, in addition, measurements or calculation models are needed to validate the applied impact force limits.

The stopping performance of the cobot is complex and, therefore, simple force limits of the robot controller do not give accurate results. According to Braman, power and force limiting is the main form of collaboration [14] and therefore it is important to consider the force limits. Currently the ISO TS 15066 provides force/pressure limits to validate cobot application. However, many aspects affect the measurement results and the measurement conditions are not yet defined in standards. The force limits face, currently, several problems: what is the right force limit for each body part, how do persons feel the impact force (sensitive vs. robust persons), how to measure the force, how the robot manufacturers can realize exact force limits in all situations. Same and, actually, more difficult problems are related to pressure limits.

The cobots can be placed also on a mobile platform and then they are mobile robots. Mobile robots can be also without additional robot on a mobile platform/robot. There are not yet standards for the safety of mobile robots and therefore the requirements need to be found from other standards, Machinery Directive and risk assessment. The amount of risks is typically larger for mobile robots than cobots, since mobile robots can be applied in many places during one work cycle.

One specific problem is related to impact to the head. According to ISO TS 15066 impact to head or sensitive body regions shall be prevented whenever reasonably practicable [7]. In most of the applications human could stick his head into dangerous impact position - The question is: What is reasonably practicable head impact prevention.

Apparently, the cobots are developing and some current issues may be solved in the near future. Currently functional safety level is not adequate for many robots, but in the near future inherently safe structures or adequate safety functions will solve the problem. The impact forces/pressures may be measured, with simple cheap device or expected impact forces could be simulated accurately. Currently, some cobots have long delays in stopping performance

and it cause long separation distance between human and robot. Long stopping time affect also impact forces. True collaboration between human and cobot require quick stopping, which is related to good brakes or motion control. Quick stopping may affect structure durability, cobot's stability and load stability, and therefore the stopping performance needs to be optimized also in the future.

Although cobots are often considered to be safe, risk assessment is required to ensure safety. Hazard identification is obligatory phase, but if the risk is similar to the risk described in harmonized standard, then the risk estimation and risk reduction can be adopted from a harmonized standard i.e., usually, ISO 10218-2. All phases of the risk assessment need to be done, one way or another.

VTT Technical Research Centre of Finland Ltd. is developing in the NxtGenRob project the optimum ways to utilize next generation robotics in Finnish industry by developing solution models, design practices and (by evaluating) demonstrations from different perspectives. The main funder of the project is Business Finland Oy. In addition, seven companies have supported the project.

References

[1] Hämäläinen M. Robotti nostaa palkkaa (In Finnish). Metallitekniikka 11/2018. p. 35.

[2] Kildal J., Tellaeche A., Fernández I., and Maurtua I., "Potential users' key concerns and expectations for the adoption of cobots," Procedia CIRP, vol. 72, pp. 21–26, 2018.

[3] Bender M, Braun M, Rally P, Scholtz O. Lightweight robots in manual assembly – best to start simply. Examining companies' initial experiences with lightweight robots. In: Bauer W, editor. Report. Fraunhofer Institute for Industrial Engineering IAO; 2016.

[4] Kirschner D, Schlotzhauer A, Brandstötter M, and Hofbaur M. Validation of Relevant Parameters of Sensitive Manipulators for Human-Robot Collaboration. International Conference on Robotics in Alpe-Adria Danube Region. ResearchGate. 2018. DOI: 10.1007/978-3-319-61276-8_27

[5] ISO 10218-2:2011. Robots and robotic devices - Safety requirements for industrial robots - Part 2: Robots. 72.

[6] Aaltonen I., Salmi T., Marstio I. Refining levels of collaboration to support the design and evaluation of human-robot interaction in the manufacturing industry. In: 51st CIRP Conference on Manufacturing Systems. Published by: Elsevier B.V. 2018. 6.

[7] ISO/TS 15066:2016. Robots and robotic devices — Safety requirements for Industrial robots — Collaborative operation. 33

[8] ISO 13849-1:2015. Safety of machinery. Safety-related parts of control systems. Part 1: General principles for design. 86.

[9] ISO 10218-1:2011. Robots and robotic devices - Safety requirements for industrial robots - Part 1: Robot systems and integration. 43

[10] ISO 12100:2010. Safety of machinery. General principles for design. Risk assessment and risk reduction. 77

[11] ISO 13855:2010. Safety of machinery. Positioning of safeguards with respect to the approach speeds of parts of the human body. 40

[12] Salmi T., Marstio I.; Malm T.; Montonen J. Advanced safety solutions for human-robot-cooperation. In: 47th International Symposium on Robotics, ISR Proceedings, (21 - 22 June 2016, Munich, Germany), Mechanical Engineering Industry Association (VDMA), Information Technology Society (ITG) within VDE, 2016. 610-615.

[13] Malm T., Salmi T., Marstio I., Montonen J., Safe collaboration of operators and industrial robots. In: Automaatio XXII proceedings (23 – 24 March 2017, Vaasa, Finland), Finnish Society of Automation, 2017, 6

[14] Braman R. 2019. The basics of Designing for Safety with Collaborative robots. MachineDesign. Uploaded from website 1.2.2019. https://www.machinedesign.com/motion-control/basics-designing-safety-collaborative-robots

Matti J Huotari*, Kari Määttä, Teemu Myllylä, and Juha Röning

Photoplethysmography signal analysis to assess sauna exposure, arterial elasticity, and recovery

Abstract: The basic biomedical information on illnesses is increasing, however, diseases like arteriosclerosis (AS) is becoming a common cardiovascular disorder among elderly people, especially in females. It is predicted that the negative impacts of AS on young people can be greater than on the elderly people in the long run. Degenerative changes in the vascular tree have many causes in addition to the life style. Arterial elasticity (AE) would provide a direct indicator for cardiovascular healthiness and predict AS risk. Photoplethysmography (PPG), and especially its response pulse wave decomposition, envelope analysis, and its second order derivative (SDPPG) opens us to determine the instantaneous heart rate (IHR) which is used to seeing on fitness equipment, sports watches, and consumer heart rate devices. It is measured in beats per minute. We do not remove ectopic or anomalous beats. A physical exposure of human to sauna bath has been shown to improve endothelial function in patients with risk factors and also heart failures. Namely, sauna exposure promotes relaxation and wellbeing, which can be recommended to prevent the development of diseases also in healthy adults.

Keywords: arterial elasticity, photoplethysmography, pulse wave analysis, sauna exposure

***Corresponding Author: Matti Huotari:** University of Oulu, E-mail: matti.huotari@oulu.fi

1 Introduction

The hemodynamic responses to the sauna exposure have specific effects which are not caused by one single stimulus. The responses, which are caused by the sauna exposure, can depend on thermoregulatory responses, age, gender, the circulatory and respiratory system, as well as traditions related with the exposure time and temperature of the Finnish sauna [2].

PPG measurement receives interest because of the simplicity, but the difficulty of adjusting parameters restrict applications. However, its second order derivative (SDPPG) opens us to determine the instantaneous heart rate (IHR) which is $60/(t_{An-1}-t_{An})$ where t_{An} is the n^{th} A peak of SDPPG. The sauna exposure can be recommended mainly in order to easier recover after physical exercise, and the various pain problems. The sauna exposures in long term effect on the motion system increases arterial elasticity. It reduces the viscosity of the blood so that blood flows easier and increases diameters of blood vessels and the joints' mobility. The IHR and variation of arterial elasticity with blood pressure are caused by the sauna exposures on healthy subjects. (Fig. 1 upper, before sauna, lower, after sauna).

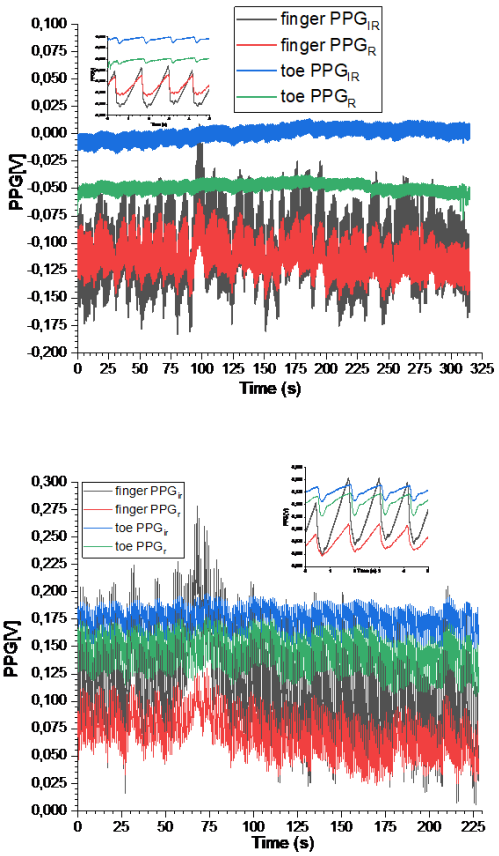


Fig. 1. Resting state PPG signal before sauna exposure for 32 years healthy male subject: finger PPG_{ir} & PPG_{red} and toe PPG_{ir} & PPG_{red} (before sauna upper, and after sauna lower). Inserts the first 5 s.

Regular sauna bathing has been shown to be protective

from cardiovascular disease according to litterature [3, 4, 5]. Our vascular system responds to interval training, especially by the sauna exposures the body to heat alternating between sauna heat and normal temperatures, the arteries are stimulated to expand and after sauna to contract. Theoretically sauna can have a great influence on the basic hemodynamic parameters such as rising the heart pulse rate, lowering blood pressure, photoplethysmography and also arterial elasticity [1]. However, the arteries expand and contract also in the resting state (Fig. 1). The sauna exposure is associated with short-term improvement in cardiac function indicated by photoplethysmographic measurements from the left forefinger and the left second toe. However, we have healthy subjects in this study. Informed consents were obtained from the test subjects, who abstained from alcohol, caffeine, and strenuous exercise in the 24 h up to the day of the tests.

2 Method and Subjects

Accurate determination of start and peak of a PPG signal plays a central role in arterial stiffness, instantaneous heart rate, and its variability. For analysis of four PPG signals correspond to each other perfectly at a given frequency, as in the case of finger IR (infra red) and red LED. In PPG technology, the main difficulty is its quantitative analysis. PPG based on phase sensitivity technic has proved very good. In our measurements the light intensities and wavelengths (red 640 nm & infrared 920 nm) are fixed. In practice, the arterial pulse waveform is based on the propagating pulse wave from the left ventricle. It travels through the arterial circulatory system and arrives the multiple peripheral, parallel capillary arteries (in Figure 1 to finger and toe). The elasticity index was calculated as the relation of the peak time of percussion wave to the peak time sum of the other waves..

The clinical patient measurements were conducted in Oulu University Hospital (OUH, Finland) as a clinical device test (test II), where we studied 17 patients of who 10 subjects exhibited normal arterials and 7 patients with peripheral arterial disease according to the ankle brachial index (ABI). Additionally, young female and male volunteers with good heart health status were included from the University departments in the sauna test group of 28 subjects. The volunteer subjects in the clinical device test and in the sauna test participating in the study were examined in supine in the Oulu University, Workshop sauna facility (I test), and OUH (II), Blood Surgery Clinic, with PPG probes. The groups contain data from 17 (II) (5 Europeans, 3 African Africans) and 28 (I) clinical subjects. The studies were approved by the ethical review boards of the Oulu University, OUH, and the Finnish National Supervisory Authority of Health and Welfare (VALVIRA).

3 Results and discussion

In Figure 2, PPG pulse waveform is decomposed to its primary wave, percussion, tidal, dicrotic, repercussion, and retidal for one heartbeat.

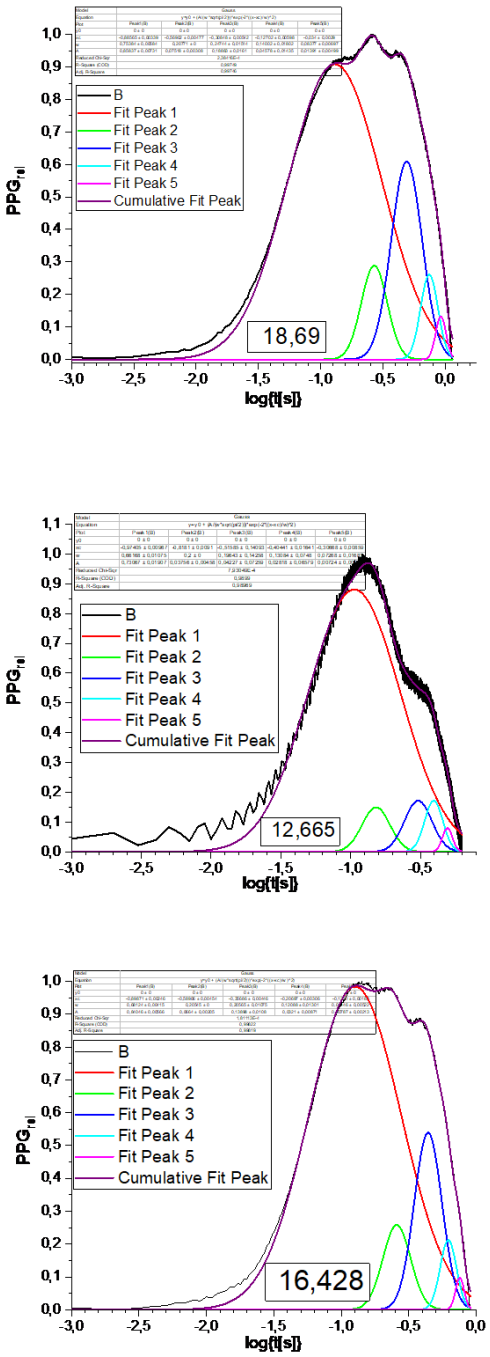


Fig. 2. PPGs from a 32 years male decomposed to its component before (upper), in sauna (middle), and after sauna (lower). Index is 18,690, 12,665 and 16,428.
The PPG pulse waves are decomposed so that they produce five primary components: percussion wave, tidal wave, dicrotic wave, repercussion wave, and retidal wave, Figure 3.

The envelopes were determined before the decomposition. In pulse waveform analysis, SDPPG waveform envelopes produce the distinction of five sequential waves called the initial positive wave, the early negative wave, the late upsloping wave, the late down sloping wave, and the diastolic positive wave, the last one is during the diastole, Figure 3. The heart rate is based on SDPPG beats between the peaks A, Figure 3.

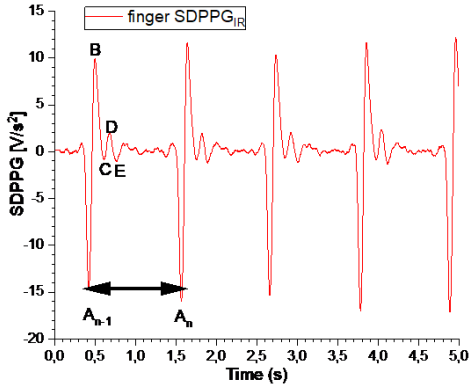


Fig. 3. SDPPG for 5 s from the left index finger by infrared led from a 32 years male healthy cohort, where A, B, C, D, and E are characteristic points of SDPPG from Figure 1 PPG. The IHR is shown with the arrow.

The wave components (A, B, C, D, and E) can be used for further calculations /2/. Further study would be needed to make clear the different causes of response patterns. Biophotonics and applications would be helpful in the future studies. However, full understanding of PPG waveforms and their physiology are still lacking. After investigation the intrinsic connection between infrared and red PPGs from finger and toe, and derived an elasticity index to describe the relationship between biophysical parameters, is based on fewer assumptions than other methods.

SDPPG of the pulse wave is also an indicator to evaluate the elastic properties of blood vessels. By calculating the maximum and minimum values of the SDPPG waves, we found that they contain possibly the elastic properties of the blood vessel wall. In order to validate and generalize the sauna exposure results, a study with larger number of healthy subjects and a more comprehensive reference method (e.g. ultrasound) would be needed, including both cardiovascular patients and healthy control subjects.

IHR in beats per min increases as in an exercise during the sauna exposure, Figures 4, 5, 6, 7, and 8. This phenomena supply blood to all the peripheral locations of body because the arteries and veins dilate. It is known the diastolic blood pressure drops with the systolic blood pressure. The increased blood circulation stimulates also sweat glands.

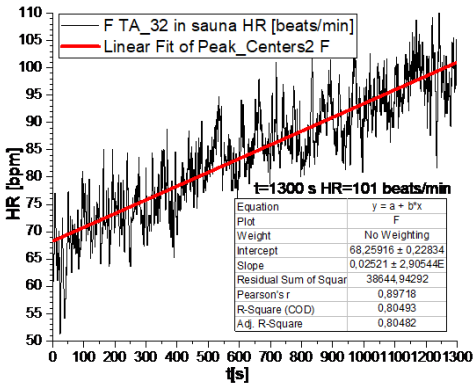


Fig. 4. During the sauna exposure for 1300 s at 80°C, the IHR increased from 67 to 101 beats/min in a 32 years male healthy cohort. The slope value is 0,0252.

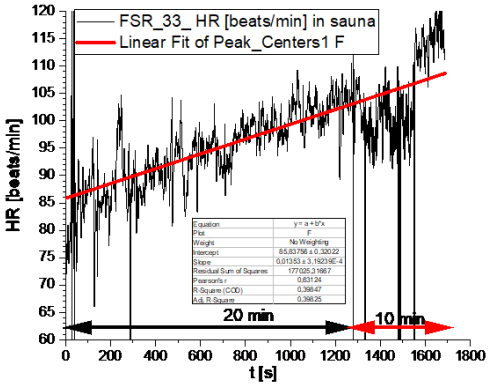


Fig. 5. During the sauna exposure for 1700 s at 80°C, the IHR increased from 85 to 120 beats/min in a 34 years male healthy cohort. In sauna the heart rate increased during the first 20 min linearly having the slope value 0,0135. After it can be change of state.

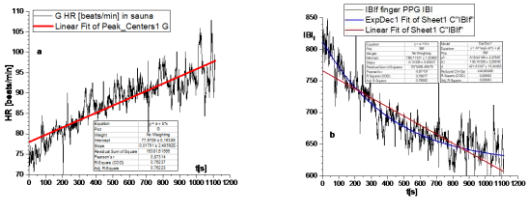


Fig. 6a and b. During the sauna exposure for 20 min, the IHR increased from 75 to 100 beats/min in a 58 years male healthy cohort with the correlation coefficient for exponential grow 0,810 and the corresponding exponential decaying IBI with correlation coefficient 0,840. During the first 20 min heart rate increased linearly with the slope value 0,0179.

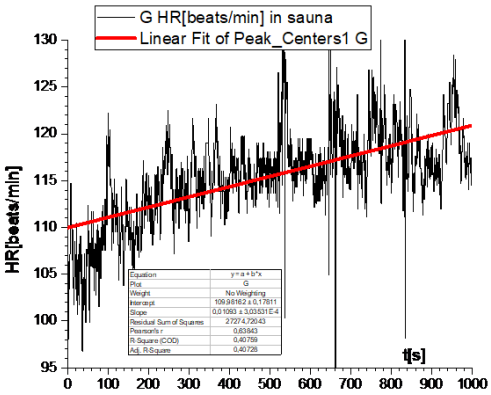


Fig. 7. During the sauna exposure for 15 min, the 10 min would be enough in this case based on the IHR increased from 105 to 120 beats/min in a 58 years female healthy cohort. During the first 15 min heart rate increased nonlinearly with the slope value 0,0109.

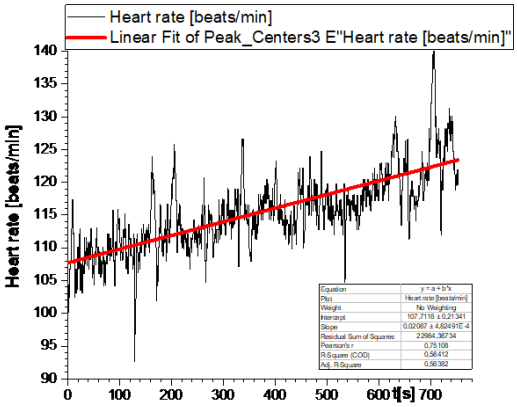


Fig. 8. During the sauna exposure for 15 min, the 10 min would be enough in this case based on the IHR increased from 107 to 125 beats/min in a 28 years female healthy cohort. During the first 15 min heart rate increased linearly with the slope value 0,0209.

There are many illnesses and diseases for what sauna exposure can be helpful and recommended, but not shown yet. The following diseases were identified: hypertension, hypotension, cardiac failure, peripheral circulatory disturbance, sensitivity to cold, arterial obstructive disease, collagen disease, Raynaud's disease, anemia, ischemic organ derangement, all kinds of rheumatoid arthritis, rheumatism, hypercholesterolemia, hyperlipidemia, hyperuricemia (including gout), glucose intolerance (diabetes mellitus, insulin resistance) [3].

IHR increased almost linearly during the first 20 min in sauna exposures. The promising results encourage us for further studies related to the PPG amplitude and rate measurements and their usage also in clinical diagnosis or screening of vascular changes. Also instantaneous heart rate variability (IHRV) would be interesting signal. HRV represents a promising marker of the autonomic nervous system (ANS) regulation exposed in the sauna. Our needs to analyze the signal on a case by case basis so long as we don't have the proper automatic measurement and analysis system. For automated clinical diagnosis based on PPG would be also important in the future healthcare.

4 Conclusions

The heart rate increased almost linearly in many cased during the sauna exposures. In each case, it is important to have the same resting period. It would be difficult that all the subjects arrive at 8 in the morning to the sauna, and we have no parallel measurement system. However, in Japan and in Finland, the frequent use of sauna or onsen exposure times during a week is important factor. The test subjects who had Finnish sauna for 7 times per a week were in the healthiest conditions for the heart and brains [4, 5]. PPG measurement gives big data records unless proper analysis procedures, like decomposition of pulse waveforms, or the linear heart rate function. The PPG pulse waveforms in sauna were difficult to decompose. The Japanese Onsen and the Finnish Sauna would be interesting in comparison measurements by PPG.

References

[1] Peltokangas, M, Vehkaoja, A, Huotari, M, Verho, J, Mattila, VM, Röning, et al. (2017). Combining finger and toe photoplethysmograms for the detection of atherosclerosis, Physiological Measurement, 38, (2).

[2] Wessapan T, Rattanadecho P (2015) Heat transfer analysis of the human eye during exposure to sauna therapy, Numerical Heat Transfer; Part A: Applications.

[3] Kamioka H, Mori Y, Nagata K, Iwanaga S, Uzura M, Yamaguchi S (2019) Relationship of daily hot water bathing at home and hot water spa bathing with underlying diseases in middle-aged and elderly ambulatory patients: A Japanese multicenter cross-sectional study, Complementary Therapies in Medicine, Volume 43, Pages 232-239.

[4] Laukkanen, JA, Laukkanen, T, Khan, H, Babar, M, Kunutsor, SK (2018) Special Article, Combined Effect of Sauna Bathing and Cardiorespiratory Fitness on the Risk of Sudden Cardiac Deaths in Caucasian Men: A Long-term Prospective Cohort Study, Progress in Cardiovascular Diseases, Volume 60, Pages 635-641.

[5] Luurila, OJ (1980) Arrhythmias and other cardiovascular responses during Finnish sauna and exercise testing in healthy men and post-myocardial infarction patients, Academic dissertation, ISBN 951-99256-6-X.