

ADVANCED APPROACH OF MATERIAL REGION DETECTIONS ON FIBRE-REINFORCED CONCRETE CT-SCANS

Marek PECHA^{1,2}, Martin CERMAK², Vaclav HAPLA², David HORAK^{1,2},
Jiri TOMCALA²

¹Department of Applied Mathematics, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava, Czech Republic

²IT4Innovations National Supercomputing Center, VSB–Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava, Czech Republic

marek.pecha@vsb.cz, martin.cermak@vsb.cz, vaclav.hapla@vsb.cz, david.horak@vsb.cz, jiri.tomcala@vsb.cz

DOI: 10.15598/aeec.v15i2.2319

Abstract. *Detections of material regions on CT-scans of solids are commonly treated manually by an expert. Although such manual detections have many advantages, some amount of human error is also incorporated. Moreover, expert opinions may vary significantly. We present an application of the k -means++ clustering as an alternative option to manual way of material area detections. k -means++ clustering is derived from k -means (the method of vector quantization, originally from signal processing), popular for cluster analysis in data mining and image processing communities. The algorithm's main advantages are its simple implementation and fast convergence to a local optimum of an objective function. We benchmark the suggested approach on transverse CT-scans of a fibre-reinforced concrete solid. Moreover, we introduce a technique for processing air distribution, such that the appropriate pixels detected as the pixels of air are converted into pixels representing concrete. The technique is based on the connected component algorithm. Benchmark and results of proposed method conclude the paper.*

Keywords

CT-scan, fibre-reinforced concrete, k -means++, material region detection.

1. Introduction

Image processing techniques form an essential part of modern simulation methods especially in pre-processing, post-processing and visualizations. In this

paper, we present an application of the commonly known k -means++ algorithm [1] as a solver for finding k maximal homogenous parts of transverse CT-scans of a solid. We call these parts material regions, which correspond to specific material areas in a CT-scan image. We suggest the technique as a semi-automatized alternative to the manual thresholding approach [2], which is used mainly in material science communities. Then, we focus on a problem formulation, which can be solved by k -means type algorithms, commonly known as Lloyd algorithms [3] and [4]. For practical purposes, we propose to use the k -means++ variant of this algorithm. Standard k -means and k -means++ are presented. Benchmark of proposed method concludes the paper.

2. Fundamental Notations and Definitions

Let $\Omega \subset \mathbb{R}^2$ be the image function area. It is suitable to define the image function $f(x, y)$ as the mapping $f : \Omega \rightarrow \langle 0, 255 \rangle$ in case of a CT-scan image.

Definition 1 (Region of Interest). The *Region Of Interest* $\Omega_{ROI} \subset \Omega$ is a part of an image function area which is selected by the segmentator/classifier (a program or human user). Region Ω_{ROI} is typically a part of image foreground and $\Omega \setminus \Omega_{ROI}$ is interpreted as the background of the image.

The *Region of Interest* is often denoted by the acronym *ROI*. For the purposes of this text, the *ROI* definition is sufficient; see [5, pp. 340] for further information. For extracting the *ROI* from an image function area Ω , a *binary mask* is used in image processing

field. In general, a binary mask corresponds to a *characteristic function of a set*.

Definition 2 (Binary mask). Let Ω be the image function area, Ω_{ROI} be *ROI*, and f be the image function. An image binary mask $f_{mask} : \Omega \rightarrow \{0, 1\}$ determines points $p_1, p_2, \dots, p_n \in \Omega$ which belong to Ω_{ROI} and it is defined as the *characteristic function* of Ω_{ROI} :

$$f_{mask}(p_j) := \begin{cases} 1 & \text{if a point } p_j \in \Omega_{ROI}, \\ 0 & \text{if a point } p_j \notin \Omega_{ROI}, \end{cases} \quad (1)$$

where $j \in \{1, \dots, n\}$.

3. Formulation of the Problem of Reinforced Concrete CT-Scan Image Scene Partitioning

Our goal is to find a suitable partitioning of the CT-scan image scene area Ω into *ROIs* $\Omega_{ROI}^1, \Omega_{ROI}^2, \dots, \Omega_{ROI}^k$, such that $\Omega = \Omega_{ROI}^1 \cup \Omega_{ROI}^2 \cup \dots \cup \Omega_{ROI}^k$. We consider that the regions are mutually disjoint, i.e. $(\forall g, h \in \{1, 2, \dots, k\}) (g \neq h) : \Omega_{ROI}^g \cap \Omega_{ROI}^h = \emptyset$. Unless otherwise stated, assume $i \in \{1, 2, \dots, k\}$.

Further, properties of pixels are maximally similar within each determined region and maximally dissimilar among the regions. More formally, the length of the interval I containing values of the function $f|_{\Omega_{ROI}^i}$ is minimized, whereas values of $f|_{\Omega \setminus \Omega_{ROI}^i}$ are as far as possible from the centroid c_i of the interval I . Other region requirements are not treated such as region compactness, or the total length of the regions boundary. The problem formulated above implies that partitioning of the image scene area is not locally specified; it treats k region homogenizations of pixel value distributions and formulates them as a clustering problem which can be solved by standard clustering algorithms. We propose Lloyd based algorithms, commonly known as algorithms of k -means type, to solve this problem.

4. The k -means Algorithm

Partitioning of input data into k subsets commonly called clusters, relabelled as regions in this text, is NP -hard problem in general. Let us stepwise denote input points, referred to as pixel values of gray-scale CT-scans, p_1, p_2, \dots, p_n , belonging to input set P , and desired k regions S_1, S_2, \dots, S_k , belonging to region set S . Often,

the k -means is denoted as *hard-clustering* type algorithm, i.e.

$$(\forall p_j \in P) (\exists! S_i \in S : p_j \in S_i). \quad (2)$$

In general, it is an image partitioning technique which decomposes an image area into reasonable non-compact parts; by definition, determination of *ROI* is not a local specified problem. However, the k -means is an optimizer, hence, the solution returned by the solver, is a relax of the optimal solution. Thus, let c_1, c_2, \dots, c_k be centroids of the appropriate image regions S_1, S_2, \dots, S_k . The regions are built so that they minimize a sum of squares within each region, i.e. the k -means solves the minimization problem

$$\operatorname{argmin}_S \sum_{i=1}^k \sum_{\bar{p} \in S_i} \|\bar{p} - c_i\|_2^2, \quad (3)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. In general, we can divide k -means algorithm procedure into two significant steps, i.e. an *initialization step* and an *update step*, as follows:

- **Initialization step:** The algorithm determines the nearest centroid from the set of centroids $\{c_1, c_2, \dots, c_k\}$, which we denote by C , of the appropriate regions S_1, S_2, \dots, S_k for each input point $p_j \in P$ as follows: *it compares stepwise Euclidean distance between the point and each centroid $c_i \in C$, and the point is assigned to region $S_i \in S$, in which the Euclidean distance between the point and appropriate centroid is minimal.* The initial state is typically a random assignment of points p_1, p_2, \dots, p_n to regions S_1, S_2, \dots, S_k or a random selection of centroids from the input point set P and then an assignment of input points by the approach mentioned above. Mersenne twister [6] is typically used as the pseudo-random integer generator.
- **Update step:** The algorithm computes new coordinates of region centroids as an arithmetic mean of the points within an appropriate region.

These two steps are repeated until the maximum iteration count is exceeded or point region membership changes are relatively small. By the following pseudocode, we simply describe the entire k -means procedure. In the implementation of the k -means algorithm, a point assignment to an appropriate region is implemented so that the point memberships are only stored in membership vector \vec{U} . This approach is efficient for a computer memory management, because there is no point reassignment overload.

Algorithm 1: k -means

Input : input set $P = \{p_1, p_2, \dots, p_n\}$, desired region count k , reasonable small $\varepsilon \in \mathbb{R}$, maximum iteration count M .

Output: regions S_1, S_2, \dots, S_k .

1. Set $t = 0$.
2. Choose randomly k starting centroids from P and assign them to starting centroid set C^t .
3. Assign points to initial regions $S_1^t, S_2^t, \dots, S_k^t$ by their Euclidean distance to the nearest centroid $c_j^t \in C^t$ and set the initial membership vector U^t as follows:
for $p_j \in P : U\{i^t\} = j^t$, where j^t is index of the point and i^t is the nearest centroid index.
4. Set iteration counter $b = 0$ and $\delta = 0$.
5. Compute new region centroids as the arithmetic mean of points within-region, i.e.

$$c_i^{t+1} = \frac{1}{|S_i^t|} \sum_{p_j \in S_i^t} p_j. \tag{4}$$

6. For each point $p_j \in P$ find the nearest centroid $c_i^{t+1} \in C^{t+1}$ by the *Euclidean distance* between them. Further, assign points to appropriate regions $S_1^{t+1}, S_2^{t+1}, \dots, S_k^{t+1}$ and set membership vector U^{t+1} ; if $U\{i^{t+1}\} = U\{i^t\}$ do nothing, otherwise $\delta = \delta + 1$.
7. If $\frac{\delta}{|P|} < \varepsilon$ or $b > M$ stop, otherwise set $\delta = 0$, $b = b + 1$ and continue to step 4.

5. The k -means++ Algorithm

The k -means algorithm quickly converges to a local optimum of the potential function ϕ , but there are many practical examples for which it generates arbitrarily bad clustered regions, i.e. $\frac{\phi}{\phi_{opt}}$ is unbounded even when $n := |P|$ and k are fixed. It becomes apparent that the reasons of this fact are the random choices of starting centroids and appropriate assignment of input points to starting regions.

Therefore, in 2007, David Arthur and Sergei Vassilvitskii proposed an efficient technique for selecting k -means starting centroids from an input point set P , and they called this technique k -means++ [1]. Its general principle is a uniformly random choice of starting centroids from input point set at probability distribution, which is computed as the squared Eu-

clidean distance to the centroid that has already been chosen, as follows:

Algorithm 2: k -means++

Input : input set $P = \{p_1, p_2, \dots, p_n\}$, desired region count k , reasonable small $\varepsilon \in \mathbb{R}$, maximum iteration count M .

Output: regions S_1, S_2, \dots, S_k .

1. Choose uniformly randomly a centroid c_1^0 from the input point set and assign c_1^0 to the set of centroids C .
2. For each $p_j \in P$, compute the best label distance $D(x)$ as the minimal Euclidean distance between point and the centroids that have already been chosen.
3. Choose next starting centroid c_i^0 randomly at probability distribution $D(x)^2$.
4. Repeat 2 and 3 until the k starting centroids have not been chosen, then continue with the standard k -means algorithm.

By this augmentation of the k -means, the k -means++ algorithm becomes $\mathcal{O}(\log k)$ -competitive, where k is the desired region count. Moreover, it can be proved that $E[\phi] \leq 8(\ln k + 2)\phi_{opt}$ [1], i.e. the solution returned by the solver is at most $8(\ln k + 2)$ worse by factor than the optimal solution.

6. Non-Local Material Detections in CT-Scan Value Distributions

Industrial CT-scans are typically 8-bit grayscale images; it implies that appropriate intensities of desired materials are coded into at most 256 values, i.e. $0, 1, \dots, 255$. From *Institute of Geonics of the CAS*, we have obtained a 1491 large CT-scan dataset of the Fiber-Reinforced Concrete (FRC), the 591st image is shown in Fig. 1. All the presented binary masks are associated with this CT-scan.

In the FRC dataset, a Fiber-Reinforced Concrete consists of 3 most significant material types, i.e. air, fiber, and concrete. The material detection problem is not locally specified by definition, therefore it is appropriate to use any of the Lloyd type clustering algorithms to solve the material detection problem. As noted in the previous section, the k -means++ algo-

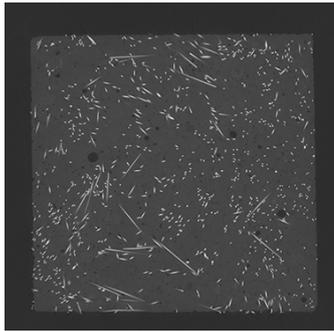
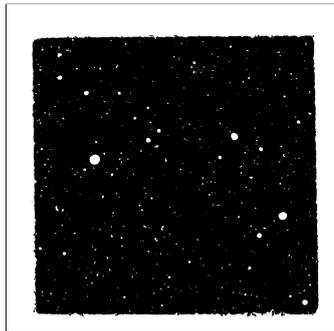
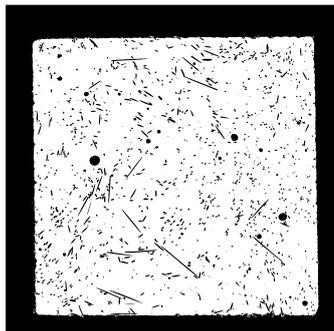


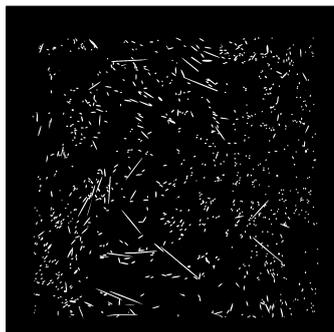
Fig. 1: 591th CT-scan from RFC dataset.



(a) Air region binary mask.



(b) Concrete region binary mask.



(c) Ferro-fiber region binary mask.

Fig. 2: *k*-means++: achieved RFC material detection solutions.

rithm is bounded. Therefore, we propose to use it instead of the unbounded standard *k*-means algorithm.

Hence, we normalize the pixel values by the standard image processing approach, i.e. remap linearly from $\{0, 1, \dots, 255\}$ to $(0, 1)$, and set *k*-means++ parameters $k = 3$ (i.e. number of material to be determined), $\varepsilon = 10^{-4}$ (tolerance), and $M = 10000$ (maximum iterations). For the 591th dataset CT-scan, the achieved solution binary masks are depicted in the following figures, i.e. the air region Fig. 2(a), the concrete region Fig. 2(b), region of ferro-fibers Fig. 2(c).

In practical applications, the outside-air region (the hatched part in Fig. 3) is not suitable for additional processing, e.g. construction of mesh material parts from determined binary masks. Therefore, we propose to detect image connected components [7] and remove a connected component from an outside-air region.

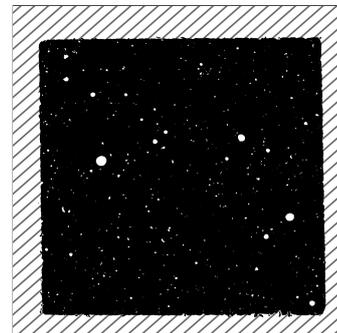


Fig. 3: Hatched outside-air region.



(a) Part of an air region binary mask.



(b) Part of a reduced air region binary mask.

Fig. 4: Reduction of a small air bubble distributions.

Moreover, small parts of an air region, i.e. small air bubbles in concrete, are typically not appropriate for additional computations, e.g. plasticity [8] and [9] or plasticity with damage [10] computations on the FRC. Hence, we propose to remove air bubbles with a size below a certain threshold from the air region and consider them as parts of the concrete region. In Fig. 4, the threshold is set to 15 pixels.

For implementation, we recommend to use C++ OpenCV framework [11]. It is an open-source image processing framework frequently used in image-processing communities. Moreover, the algorithms above, i.e. k -means, k -means++ and image connected components algorithm, are implemented in the OpenCV framework, see [12].

We benchmark the presented approach on FRC dataset containing 1491 CT-scans. As already mentioned, such k binary masks determine as homogenous as possible k parts of geocomposite material parts. Furthermore, the binary masks can be used for another processing, e.g. building of hull and volume meshes by the Marching cubes algorithm [13].

7. Future Work

In case of industrial simulations such as plasticity computation on a heterogeneous solid, it is suitable to convert the real-world solid into a finite element model enabling highly accurate simulations. Image partitioning/segmentation plays an important role in detection of material regions of the CT-scan image scene and set up their binary masks. Furthermore, the determined binary masks are used for building hull-mesh by the marching cube algorithm as we present in the text above. For a tetrahedral volume mesh geometry generation from hull-mesh boundary representation, the NETGEN library [14] is typically used.

The plasticity computations will be proceeding in PERMON [15] (Parallel, Efficient, Robust, Modular, Object-oriented, Numerical) software toolbox developed at IT4Innovations National Supercomputing Center, Ostrava, Czech Republic. PERMON is a set of solvers combining quadratic programming and domain decomposition methods. It makes use of and extends the PETSc [16] framework for numerical computations. Among the main applications are contact problems of mechanics. They can be decomposed by means of TFETI (Total Finite Element Tearing and Interconnecting) [17] non-overlapping domain decomposition method implemented in the PermonFLLOP package. The resulting quadratic programming problems can be solved by the PermonQP package efficiently.

8. Conclusions

In this paper, we introduce an advanced tool for detections of material regions on CT-scans of fiber-reinforced concrete. The essential part is based on the k -means++ clustering for detecting binary masks of desired material areas on the CT-scan image scene. We benchmark the presented approach on FRC dataset containing 1491 CT-scans. Three significant material regions are detected, i.e. air, concrete, fibers. Moreover, we introduce a technique based on an image connected component algorithm for reduction of the air bubble distribution. Based on the achieved results, we propose the automatized presented technique as a suitable alternative to the manual thresholding approach, which is a standard way to solve such problem in material engineering communities. Additionally, the most significant benefits of an automatized technique like ours, which is being presented in this paper, is its possibility to be integrated into a pipeline, in which an industrial process is simulated. The optimization and tuning of the pipeline runned on high performance computers are subjects of our further research.

9. Acknowledgments

This work was supported by the Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project IT4Innovations excellence in science (LQ1602), and from the Large Infrastructures for Research, Experimental Development and Innovations project IT4Innovations National Supercomputing Center (LM2015070); by the internal student grant competition project SP2017/169 "PERMON toolbox development III" and by the Czech Science Foundation (GACR) project no. 15-18274S. Work is also partially supported by Grant of SGS NO. SP2017/122, VSB-Technical University of Ostrava, Czech Republic.

References

- [1] ARTHUR, D. and S. VASSILVITSKII. k -means++: The advantages of careful seeding. *Theory Stanford* [online]. 2007. Available at: <http://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf>.
- [2] BLAHETA, R., R. KOHUT, A. KOLCUN, K. SOUCEK, L. STAS and L. VAVRO. Digital image based numerical micromechanics of geocomposites with application to chemical grouting. *International Journal of Rock Mechanics and Mining Sciences*. 2015,

- vol. 77, iss. 1, pp. 77–88. ISSN 1365-1609. DOI: 10.1016/j.ijrmmms.2015.03.012.
- [3] LLOYD, S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. 1982, vol. 28, iss. 2, pp. 129–137. ISSN 0018-9448. DOI: 10.1109/TIT.1982.1056489.
- [4] FORGY, E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*. 1965, vol. 21, iss. 3, pp. 768–769. ISSN 1541-0420.
- [5] BRINKMANN, R. *The Art and Science of Digital Compositing*. 1st ed. San Francisco: Morgan Kaufmann, 2008. ISBN 978-0123706386.
- [6] MATSUMOTO, M. and T. NISHIMURA. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*. 1998, vol. 8, iss. 1, pp. 3–30. ISSN 1049-3301. DOI: 10.1145/272991.272995.
- [7] DEVIJVER, P. A. and C. RONSE. *Connected components in binary images: the detection problem*. 1st ed. New York: John Wiley & Sons, 1984. ISBN 0-86380-016-5.
- [8] SYSALA, S., M. CERMAK, T. KOUDELKA, J. KRUIS, J. ZEMAN and R. BLAHETA. Subdifferential-based implicit return-mapping operators in computational plasticity. *ZAMM Zeitschrift fuer Angewandte Mathematik und Mechanik*. 2016, vol. 96, iss. 11, pp. 1318–1338. ISSN 0044-2267. DOI: 10.1002/zamm.201500305.
- [9] MARKOPOULOS, A., V. HAPLA, M. CERMAK and M. FUSEK. Massively parallel solution of elastoplasticity problems with tens of millions of unknowns using PermonCube and FLLOP packages. *Applied Mathematics and Computation*. 2015, vol. 267, iss. 1, pp. 698–710. ISSN 0096-3003. DOI: 10.1016/j.amc.2014.12.097.
- [10] BLAHETA R., P. BYCZANSKI, M. CERMAK, R. HRTUS, R. KOHUT, A. KOLCUN, J. MALIK and S. SYSALA. Analysis of Aspö Pillar Stability Experiment: Continuous thermo-mechanical model development and calibration. *Journal of Rock Mechanics and Geotechnical Engineering*. 2013, vol. 5, iss. 2, pp. 124–135. ISSN 1674-7755. DOI: 10.1016/j.jrmge.2012.10.002.
- [11] Open Source Computer Vision Library. *GitHub, Inc.* [online]. 2017. Available at: <https://github.com/itseez/opencv>.
- [12] The OpenCV Reference Manual. *OpenCV* [online]. 2017. Available at: <http://opencv.org/>.
- [13] LORENSEN, W. E. and C. HARVEY. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*. 1987, vol. 21, iss. 4, pp. 163–169. ISSN 0097-8930. DOI: 10.1145/37401.37422.
- [14] Netgen Mesh Generator. *Source Forge* [online]. 2017. Available at: <https://sourceforge.net/projects/netgen-mesher/>.
- [15] HAPLA, V., Z. DOSTAL, D. HORAK, M. CERMAK, R. SOJKA, M. PECHA, J. KRUZIK and J. TOMCALA. PERMON (Parallel, Efficient, Robust, Modular, Object-oriented, Numerical). *IT4Innovations* [online]. 2017. Available at: <http://permon.it4i.eu>.
- [16] SATISH, B., A. SHRIRANG, M. ADAMS, J. BROWN, L. DALCIN, T. ISAAC, M. KNEPLEY, D. MAY, T. MUNSON, K. RUPP, B. SMITH, S. ZAMPINI, H. ZHANG, H. ZHANG, P. BRUNE, K. BUSCHELMAN, S. FARLEY, V. EIJKHOUT, W. GROPP, D. KARPEEV, D. KAUSHIK, L. C. MCINNES, V. MINDEN and J. SARICH. Portable, Extensible Toolkit for Scientific Computation. *PETSc* [online]. 2017. Available at: <http://www.mcs.anl.gov/petsc>.
- [17] DOSTAL, Z., D. HORAK and R. KUCERA. Total FETI – an easier implementable variant of the FETI method for numerical solution of elliptic PDE. *Communications in Numerical Methods in Engineering*. 2006, vol. 22, iss. 12, pp. 1155–1162. ISSN 1069-8299. DOI: 10.1002/cnm.881.

About Authors

Marek PECHA was born in Zilina, Slovak republic. He has received his M.Sc. degree in Computer Science and Applied Mathematics from VSB–Technical University of Ostrava (VSB–TUO), Czech Republic (2016 Faculty of Electrical Engineering and Computer Science dean’s award for diploma thesis). Since 2014 and 2016, he has been a Research assistant at IT4Innovations National Supercomputing Center and a Ph.D. student at the Department of Applied Mathematics, respectively, both institutions being within VSB–TUO. His research of interest covers numerical analysis, machine learning, computer graphics, material sciences, and massively parallel programming. He is a co-author of the PermonSVM library for linear support vector machines problems, and also a co-author of massively parallel version of the PRAGTIC software for fatigue analysis.

Martin CERMAK was born in Hranice, Czech

Republic. He received his M.Sc. from Computational Mathematics in 2008. He finished Ph.D. in 2012 from Computational Mathematics at VSB-TUO. Since 2012, he has been a Researcher at IT4Innovations National Supercomputing Center, VSB-TUO. His research interests include scalable algorithms for solving elastoplastic problems, material nonlinearities, contact problems, quadratic programming, non-overlapping domain decomposition, and parallel implementation of these methods.

VACLAV HAPLA was born in Havirov, Czech Republic. He has been a researcher at IT4Innovations National Supercomputing Center since 2011. He is the main developer of the open source PERMON libraries based on top of PETSc. His research is focused on massively parallel implementation of scalable Domain Decomposition Methods (DDM) and Quadratic Programming (QP) algorithms, and applications of QP such as contact problems or machine learning via Support Vector Machines. He is the author of the PRACE video tutorials about PETSc (available on YouTube), and lead full-day PETSc tutorials at PRACE Spring School 2012 in Cracow, Poland, and at PRACE Autumn School 2016 in Linz, Austria. He was awarded two national prizes in the field of computational sciences: 1st prize in Joseph Fourier Prize 2014, and 3-4th prize in Babuska Prize 2016.

David HORAK was born in Bohumin, Czech

Republic. He has received his M.Sc. and Ph.D. degrees in Computational Science and Applied Mathematics and Assoc. Prof. degree in Applied Mathematics from VSB-TUO. In 2007, he was awarded the Prof. Babuska Prize for the best Ph.D. thesis in Computational Sciences. Since 2017, he has been an Assoc. Prof. at the Department of Applied Mathematics and since 2012 a Researcher at IT4Innovations National Supercomputing Center, both at VSB-TUO. His professional interest covers development and implementation of the scalable quadratic programming solvers using domain decomposition methods of FETI type, integral-discrete transforms. He is co-author of the PERMON library.

Jiri TOMCALA was born in Frydek-Mistek, Czech Republic. From 1998 to 2016, he worked as a System Administrator at a middle-sized building company. In 2016, he received his MSc. degree in Computer Science and Applied Mathematics from VSB-TUO. Since 2016, he has been a Researcher at IT4Innovations National Supercomputing Center, VSB-TUO. His professional interest covers C/C++ programming and development and implementation of scalable solvers using domain decomposition methods of FETI type. He is a co-author of the PERMON library for the massively parallel solution of quadratic programming problems.