UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Principal Component Analysis Visualization and State Discovery with Soil Data

Sirola, Miki; Koskinen, Markku; Polvinen, Tatu; Pihlatie, Mari

2023-09-09

http://hdl.handle.net/10138/569578

Sirola, M, Koskinen, M, Polvinen, T & Pihlatie, M 2023, Principal Component Analysis Visualization and State Discovery with Soil Data. in Proceedings of the 12th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems. Proceedings of the IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems, IEEE, New York, IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems, Dortmund, Germany, 07/09/2023.

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

The 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
7-9 September, 2023, Dortmund, Germany

Principal Component Analysis Visualization and State Discovery with Soil Data

Miki Sirola [1], Markku Koskinen [1], Tatu Polvinen [1], Mari Pihlatie [1]
[1]University of Helsinki, P.B. 56, FI-00014 University of Helsinki, Helsinki, Finland, miki.sirola@helsinki.fi,
https://www.mv.helsinki.fi/home/miksirol/index.html

***Abstract*—Data exploration helps to gain understanding of the dataset and the system itself. There are methodologies to handle large number of sensors as well. In this paper operational states are defined to interpret physical behaviour in a soil ecosystem. Dimensionality reduction is achieved with Principal Component Analysis (PCA) method giving another view to the soil dataset from spring term. K-means algorithm groups data densities by clustering the data. This grouping is the basis for defining operational states in the system. Soil data as a part of an ecosystem involves specific features. In the applied approach dynamic visualization including animations constitute an important exploration view. All experiments are realized in Jupyter programming environment with Python 3 programming language. Related literature about data visualization is reviewed. Combining methods and tools with this data as a result soil ecosystem features are recognized.**

Keywords—Principal Component Analysis; Visualization; State Discovery; Soil Data

I. INTRODUCTION

Data exploration is done to gain understanding for the data set and the system itself, and to cope with a large number of sensors. In this paper the approach is to describe operational states and dynamic behaviour of the system.

Dimensionality reduction gives another view to the data and enables e.g. two or three dimensional visualizations. Here, Principal Component Analysis (PCA) method is used. The states are defined from the PCA result by identifying data densities, clusters that can be named as states. K-means algorithm is used in clustering.

In our earlier papers about state discovery with datasets from autonomous, self-healing data centers [1][2], we also have tried to predict state transitions with Hidden-Marcov Model (HMM). In this paper no predictions for state transistions are made, but the idea here is to define the states and explore state transitions to gain understanding of the ecosystem.

Here, soil pit data is explored. Also here physical interpretations of the defined states are important to gain better understanding of the system. Data exploration includes both static and dynamic visualization. Animations are used in visual exploration to illustrate the state compositions and transitions. Jupyter programming environment and Python 3 programming language are used in all realizations of experiments in this paper.

The paper structure is organized in follwing way. After defining the problem, introducing the background and shortly discussing the methods and tools in Introduction Section (I), related literature about visualization is examined in Section II. In Section III methods and tools are presented. In Section IV soil pit data is introduced. The core of this paper: state discovery exploration and visualization results are presented in Section V. Discussion Section (VI) and Counclusion Section (VII) wrap up the core contents of this paper before the Reference list.

II. RELATED WORK

Data visualization from related point of views is discussed in the following references in literature. Visualization and machine learning in data center management has been studied and discussed in [3]. Visualization and machine learning has been discussed also in [4]. No corresponding soil data studies were found.

It is interesting also to find out how Principal Component Analysis (PCA) techniques have been utilized in animation research. The approach in this paper is data exploration, but also a more general view is within our scope. It seems that for instance facial animation [5][6][7][8] with PCA methods is rather common topic. Human motion animation with PCA [9] is another common field. PCA techniques have also been utilized in modelling emotions [5].

Paper [3] presents a novel tool for data center management including data visualization and data management capabilities. In [4], design space for visualization of multidimensional comparative data analytics is defined.

In [5], PCA model using shape correctives based on incremental PCA learning is used in facial animation. Combination of dense depth maps and texture features around eyes and lips are used in recognizing emotional nuances. In [6], parametrization of mouth images for an image-based facial animation system is described. Visual parametrizations with Local Linear Embedding (LLE) and PCA are compared. In [7], PCA is used in calculating face image features. Neural networks are utilized as conversion

model. In [8], PCA is used to reduce the number of parameters in three-dimensional facial animations.

In [10], PCA based compression techniques are presented in 3D animations. Virtual human motion animations using PCA techniques are discussed in [9]. A geometry compression method based on clustered principal component analysis is introduced in [11]. Data-driven approach is used. In [12], PCA based compression scheme in soft-body 3D animations is described. Motion data is experienced with PCA methods also in [13]. Paper [14] presents affine transformation matrix in PCA analysis to compress data in 3D animation models.

In recent studies PCA is used in the flowing contexts. Compression of 3D mesh animation data of spatial-temporal segments with PCA is described in [15]. Object-based compression of three-dimensional animation geometry is discussed in [16]. PCA method is used. Sparse PCA to decompose original motions into smaller components learned by particular constraints in computer animation is presented in [17].

In character animation PCA is utilized in reduced linear regression problem [18]. PCA and Higher Order Singular Value Composition (HOSVD) are utilized in matrix and tensor-based approximation of 3D face animations [19]. Weighted Linear Discriminant Analysis (LDA)-PCA based colour quantization method suppressing saturation decrease is used in [20]. PCA based compression techniques are discussed also in [21].

PCA based 3D city model generalization for electricity simulation is presented in [22]. In [23], Hidden Markov Model (HMM) and Deep Neural Network (DNN) techniques are used in synthesizing facial animation. Features for an interactive agent implementation are described and the results are compared with conventional PCA analysis to perform objective evaluation.

III. Methods and Tools

Research methodologies used in this study are data exploration, statistical analysis, visualization with chosen assistant methodologies, and building exploration animation prototypes.

The main assistant methodology is Principal Component Analysis (PCA) [24]. PCA method compresses N variables to defined number of projections. In visualition usually two or three first components are used. The first component contains most variance in data, next component contains most of the remaining variance, etc. In dimensionality reduction some information is always lost, but at the same time completely new view opens to the data.

PCA method describe well data structure, properties, states and state transitions. In addition to variance distribution for each component, it is possible to check how much each variable effects to each PCA component. This component dominance measure is called PCA loadings.

Another important assistant methodology is K-means clustering algorithm [25]. K-means algorithm illustrates well local accumulations and densities in data. K-means algorith reveals data densities, clusters, states and state transitions illustratevely. Although this method is rather sensitive to outliers, it is mostly reliable and stable.

In addition to common PCA visualizatios PCA animations are used to illustrate dynamic behaviour in data. The animations show clearly the state composition in time including illustrative desciptions of state transitions. The animations help also in detecting anomalies in data. Sometimes failure states can be recognized.

In all realizations and experiments in this paper Jupyter tool and programming environment, and Python 3 programming language are used.

IV. Data

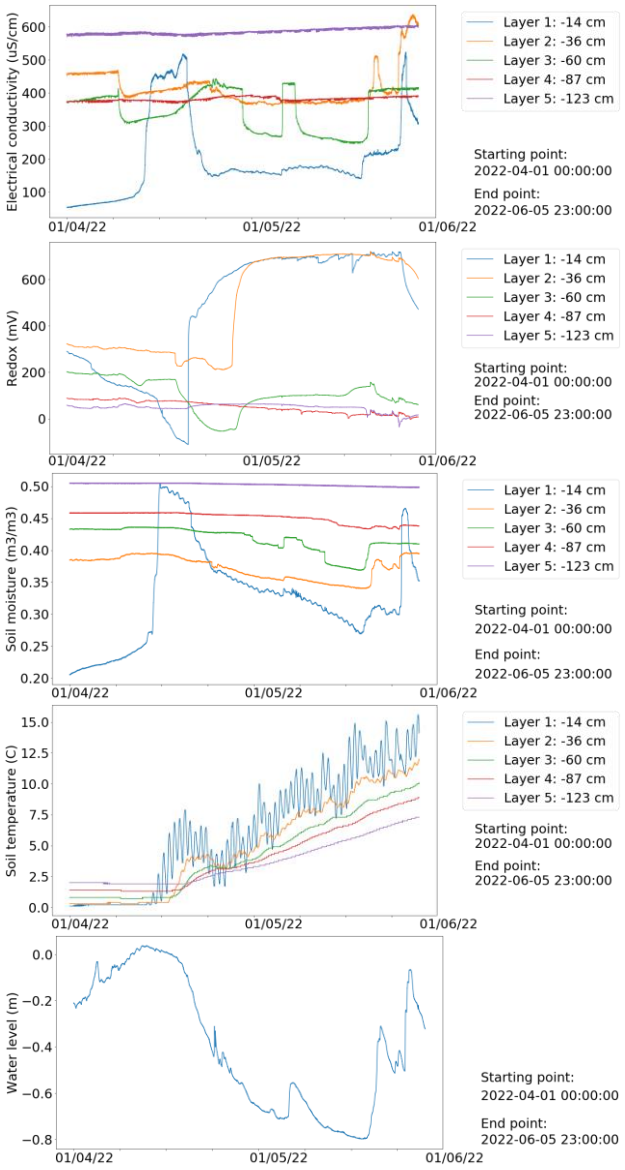

Figure 1. Main variables in the soil pit data dataset during spring term 2022.

Data is soil pit data from spring term 21.3.22 13.05 – 5.6.22 23.00. Data is collected from Viikki measurement station in Helsinki, and it is part of INAR (Institute for Atmospheric and Earth System Research) in Finland, which is connected to a big international network of research stations allover the world [26]. The measurement station is named SMEAR-Agri, so the used soil measurement results represent agricultural field environment in spring time.

The dataset includes such variables as Electrical Conductivity, Redox Potential, Water Content, Temparature, Matric Potential and Water Level. All these varaibles, except Water Level, are measured in five different depth levels is the ground. The leyers are 14 cm, 36 cm, 60 cm, 87 cm and 123 cm below the soil surface.

The dataset includes 22191 measurement samples and 42 variables. In the analysis only 33 first variables are used, because the rest of the variables are less important and include some binary variables that could not be used in PCA analysis.

In Figure 1 there are Electric Conductivity, Redox Potential, Water Content and Temeperature in all five leyers and Water Level. Note that in the figure the period begins from 1st of April, and the first about ten days samples in the dataset are missing.

V. State Discovery Exploration and Visualization Results

The PCA analysis compresses the main features in this ecological process into three main components, see the artificial time series expression in Figure 2. Density function expressions in Figure 3 opens another view to the same data.

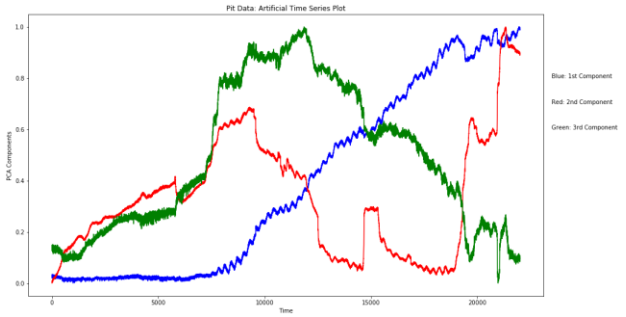

Figure 2. Three first PCA components in the analysis: artificial time series expression.

In the 1st PCA component (blue colour in Figure 1) the soil temperatures are dominating factors. In the 2nd PCA component (red colour in Figure 1) water content in soil and the rather strongly correlating electrical quantities are dominating. In the 3rd PCA component (green colour in Figure 1) soil water level and correlating quantites are dominating.

Two different expressions of density functions in the PCA analysis are seen in Figure 3. In Figure 4 there is 3-dimensional scatter plot of the PCA result including already also clustering information, here divided into three clusters.

PCA loadings, see Table 1, shows the dominance of each variable in each PCA component. In the table you can read also the variance distributions for three first PCA components. In this analysis the 1st PCA component is rather dominating having 73% of the whole variance while the two following components have 10% and 7 % of it. The table names the most dominating variables in each three first PCA components

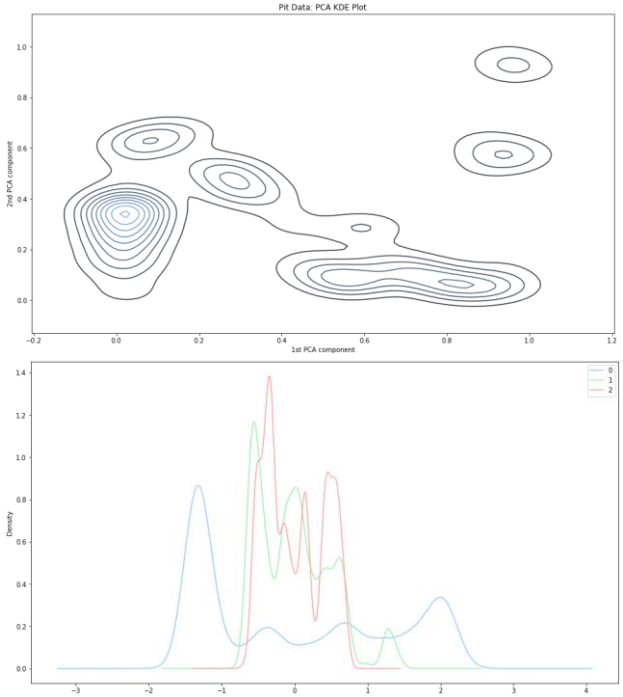

Figure 3. Density function in the form of KDE plot (upper plot) and drawn PCA component curves (lower plot). In the lower plot 1st PCA component is in blue colour, 2nd PCA component in green colour and 3rd PCA component in orange colour.

TABLE I. PCA Loadings

pca.components_	1st PCA component	2nd PCA component	3rd PCA component
> 0.4		WC_2	RX_3*, WC_1
> 0.3		EC_1, EC_3, WL*	MP_1
> 0.2	RX_2, RX_4*, T_{1-5}, WC_4*	RX_2*, RX_5*, WC_1, EC_2, WC_3	EC_1, EC_2*, WL
Relative variance	0.73	0.10	0.07

WC = Water Content
RX = Redox Potential
EC = Electrical Conductivity
WL = Water Level (Under ground)
MP = Matric Potential

T = Temperature
* = - (reverse correlation)
indeces 1-5 = layers in the ground

K-means clustering is the basis here for defining the states in this process. In Figure 4 you can see three states and in the Figure 6 seven states. State is a defined physical appearance of the process. In this case the number of states and the interpretation of them is not self-evident at all.

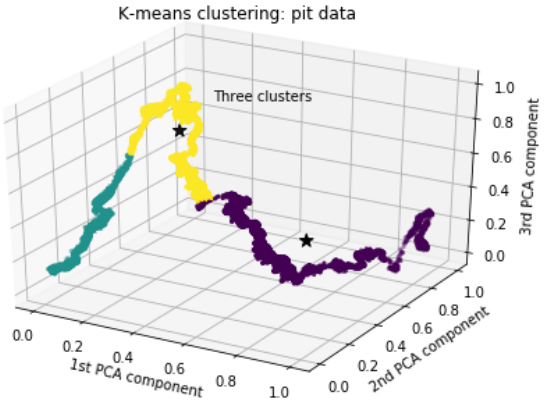

Figure 4. Three-dimensional scatter plot of the PCA result including grouping into three clusters.

The density functions, especially KDE plot, see Figure 3, supports seven states, while Davies-Bouldin Index gives the highest score for three clusters, see Figure 5. Seven clusters is a local maximum in score optimization though.

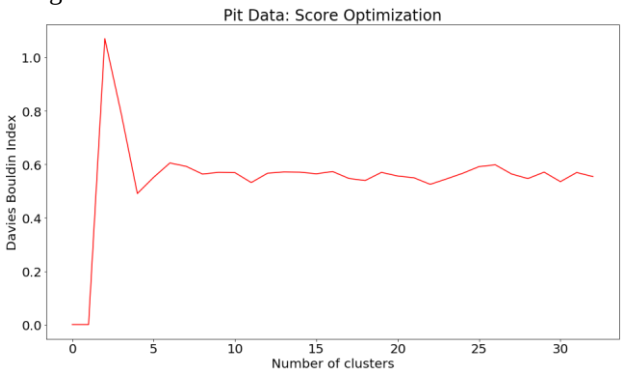

Figure 5. Score optimization with Davies-Bouldin index.

Three states are rather easy to name, but in seven state interpretation it is not so clear to find interpretations for each state separately. The soil samples are from spring term in Finland so the chronologically first state is clearly frozen ground (ground frost). The next state is melting ground. When the ground has melted the temperatures begin to rise little by little.

When ground frost melts the water content increases, electrical connectivity in soil increases and water level first goes up, but begins to decrease after the melting is complete and the temperatures beging to rise up. The electrical quantites strongly correlates with the water content.

The rest of the states comes out from rising and falling values in water content due to rainy and dry periods. Here the electrical quantites again correlate. So, in seven state interpretation you can question is every new entering state really a new one, or just repeating some structure and coming back some of the previous states. Because the temperatures mostly rise up little by little, that at least is a clear separating factor between some otherwise possibly kind of repeating states.

The state structure here is clearly chronological as the data is from spring term and summer time is approaching. This is not always the case. For example taking a whole year there would come more of a circulating structure of states where also frost ground would come back in the picture again.

Note that the water content first increase rapidly when ground has melted, but then begin to decrease. In this spring there comes another temporarily rather moisture period in the late spring.

To summarize the main state separation factors are clearly ground frost, melting ground (including changes in many dominating variables), rising temperatures, changes in water content. Also the electrical quantites have an important role here, but they also correlate strongly with the water content.

This kind of analysis can be used for identifying failure states as well. In this case we cannot talk about failure states with this data. All states are real in the ecosystem process. A measurement error could be a cause of a failure state though, but not such a phenomenom is recognized in this data.

In our previous paper [2] we express animations as practical and useful exploration tools in defining and understanding states and making physical interpretations in the process out of them. In Figure 6 you can see similar animation structure with current dataset, where seven different states are composed this time in chronological order. This figure is one statistic appearance of the whole dynamic animation.

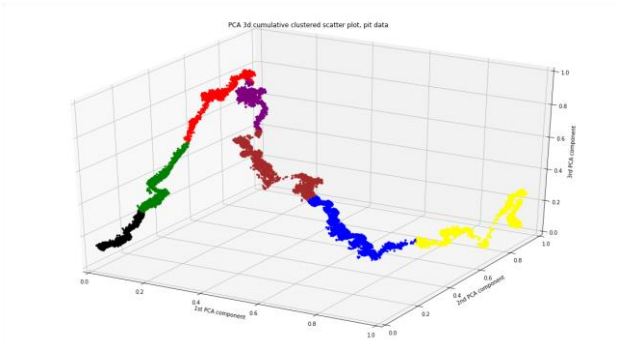

Figure 6. Exploration animation of the ecological process in spring term where seven different states follow each other in chronological order. The animated structure origins from the clustered PCA result of our dataset in examination for this paper.

The animation in Figure 6 beging from the down-left corner, where the first black coloured state represents clearly frost ground. Then little by little the soil melting process starts, and after that all known phenomena in spring term follow each other and come in the figure in an expressive manner. The animation ends in the down-right corner of the plot, where the state in yellow colour already approaches early summertime behaviour.

This animation srolls through all 22191 measurement samples in one minute and 51 seconds. The technical realization has some adjustable parameters e.g. to make the animation a bit faster or slower. The clustered PCA result is an illustrative example scenario showing out some spring term features measured from a soil pit.

VI. DISCUSSION

This paper is an attempt to show how explorative state discovery and visualization techniques work out also with soil pit data. Animating clustered PCA analysis results gives us an opportunity to how different operational states evolve in time in an ecological process.

The literature about visualization techniques and animation applications is reviewed quite intensively. No applications with soil data were found, but implementations with many other types of data was discovered. An interesting view to PCA is presented and described in [27] to complete the basic PCA references metioned in Section III about methods and tools.

The methodologies and tools used are well-known and common. PCA analysis reveals very well the data structure, which is a good basis for defining states from the system behind the data. K-means clustering helps in detection of data densities and finding grouping structures for various operational states. In this paper no prediction methods are utilized. Jupyter programming tool and Python 3 programming language is used in all realizations and presented experiments.

The state definitions in an ecologiacal system is somewhat problematic as well. Some states are obvious thinking of for instance variating seasons in northern climate. Some important variables depend on weather variations, and cannot be predicted reliably. Some kind of operational states can help in understanding the system and phenomena connected to it though.

Static visualization from different prespectives is somewhat easier than dynamical. Dynamical visualization including animations is difficult to concretize in a scientific paper, where it is only possible to use text, statistics, tables and static figures. The aim is to do this as clear as it is possible in this kind of forum.

There are available different scalers for normalizing data. In this analysis min-max scaler is used. In many cases it is justified to use standard scaler instead. The weakness in min-max scaler can occur e.g. if data includes many outliers, and therefore the scaling may get distorted and e.g. filter out peaks. With this data this was not a problem.

In the analysis in this paper no failure states were found. This was due to correctness in measurements. It would be interesting to track a measurement error state with this method. We might have data available for that, because last summer in another pit a reaper cut the cable of a Redox sensor, and after that the corresponding measurement output was kind of interesting in this perspective. The output had nothing to do with the real measurement result anymore, and therefore it would be a good example of a failure state caused by a measurement error. It would be interesting to see how this would affect the PCA result with a bigger amount of data. This case is left for a future work for furher analysis.

Another interesting issue for future work would be to add to the analysis possible interconnection of soil modes (states) and exchange of greenhouse gases between soil and the lower part of the atmosphere. We actually have good possibilities for extending our view including this point as we collect also meteorology data and Eddy Covarience data in the SMEAR-Agri measurement station. Especially the latter one would be useful here. Now our analysis utilizes only pit data. Adding some important and significant variables in this context from Eddy Covariance data into our analysis might open completely new aspects for our future studies.

Although the methodology and tools used here are well-known, the contribution to science in this paper comes from utilizing this combination of methods on a completely new application area. Also our steps used to utilize animations in data exploration constitute quite an unique view. Numerical or quantitative assessment in general and also comparison to other studies or methodologies with similar aims is very difficult, because the benefits in our study have mostly qualitative character.

VII. CONCLUSION

This paper shows how explorative data analysis can help in understanding ecological system by defining state structures that reflect the important phenomena in the process and utilizing visualization techniques including animations of grouped PCA analysis results. This kind of animation prototype illustrates how the states form and state transition occur and evolve with time in this case during spring season in agricultural soil.

The contribution of this work is in the domain case study that shows clearly some basic symptoms typical of this kind of structures. The PCA analysis animations are found to be effective in observing details in this context.

ACKNOWLEDGMENT

We acknowledge the possibility to use SMEAR-Agri soil data in our study. All SMEAR data will come publicly available.

REFERENCES

[1] O-P. Rintakoski, M. Sirola, L. Nguyen, and J. Hollmen, “State discovery and prediction from multivariate sensor data,”

Workshop on Advanced Analytics and Learning Temporal Data (ECLM PKDD), 2021.

[2] M. Sirola, O-P. Rintakoski, L. Nguyen, and J. Hollmen. "Principal component analysis visualizations in state discovery by animating exploration results," *IEEE International Workshop on Sensors and Smart Cities (SSC SmartComp 2022)*. Espoo, Finland, 2022.

[3] A. Chircu, E. Sultanov, D. Baum, C. Koch, and M. Sebler, "Visualization and machine learning for data center management." *Informatik workshops, Lecture Notes in Infomatics (LNI)*, Bonn 2019, pp 23-35.

[4] B. Schneider, D.A. Keim, and M. El-Assady, "DataShiftExplorer: Visualizing and comparing change in multidimenional data for supervised learning," *15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAP 2020)*. Valletta, Malta, 2020.

[5] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctiveness," *ACM Transactions on Graphics, Proceedings of the 40th ACM SIGGRAPH Conference and Exhibition*, 2013.

[6] K. Liu, A. Weissenfeld, and J. Ostermann, "Parametrization of mouth images by LLE and PCA image-facial animation," IEEE International *Conference on Acoustics Speech and Signal Processing Proceedings*, 2006.

[7] Y. Saito, T. Nose, T. Shinozaki, and A. Ito, "Conversion of speaker's face image using PCA and animation unit for video chatting," *International Conference on Intelligent Infromation Hiding and Multimedia Signal Processing (IIH-MSP)*, 2015.

[8] K. Goudeaux, T. Chen, S-W. Wang, J-D. Liu, "Principal component analysis for facial animation," *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, 2001.

[9] P. Glardon, R. Boulic, and D. Thalmann, "PCA-based walking engine using motion capture data," *IEEE Conference*, 2004.

[10] L. Vasa, and V. Skala, "COBRA: Compression of the basis for PCA represented animations," *Computer Graphics Forum, Vol. 28, Issue 6, 2009, pp. 1529-1540*.

[11] M. Sattler, R. Sarlette, and R. Klein, "Simple and efficient compression of animation sequences," *Proceedings of ACM SIGGRAPH / European Symposium on Computer Animation*, 2005, pp. 209-217.

[12] Z. Carni, and C. Gotsman, "Compression of soft-body animation sequences," *Computers & Graphics, Vol. 28, Issue 1, 2004*, pp. 25-34.

[13] K. Forbes, and E. Fiume, "An efficient search algorithm for motion data using weighted PCA," *Proceedings of ACM SIGGRAPH / Eurographic Symposium on Computer Animation*, 2005, pp. 67-76.

[14] P-F. Lee, C-K. Kao, J-L. Tseng, B-S. Jong, and T-W. Lin, "3D animation compression using affine transformation matrix and principal component analysis," *IEICE Transactions on Information and Systems Vol. E90-D, No. 7*, 2007, pp.1073-1084.

[15] G. Luo, X. Zhao, Q. Chen, Z. Zhu, and C. Xian, "Dynamic data rehaping for 3D mesh animation compression," *Multimedia Tools and Applications*, 2021.

[16] S. Das, and P.K. Bora. Object-based compression of 3D animation geometry. *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2018.

[17] K. Do, X. Nguyen, and H. Yu, "Learning and transferring motion style using sparse PCA," *Journal of Science: Comp. Science & Com. Eng., Vol. 35, No. 1, 2019, pp.1-10*.

[18] T. Nguyen, D. Nguyen, and V. Dinh, "PCA-based 3D facial reenactment from single image," *IEEE International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, Ha Noi, Vietnam, 2020.

[19] M. Romanszewski, A. Sochan, and K. Skabek, "Matrix and tensor-based approximation of 3D face animations from low-cost range sensors," *Communications in Computer and Information Science book series (CCIS, volume 935)*, 2018.

[20] S. Kojima, M. Harada, Y. Ueda, N. Suetake, "Weighted PCA-LDA based color quantization method suppressing saturation decrease," IEICE Tranactions on Fundamental of Electronics, Communications and Computer Sciences, 2020.

[21] G. Luo, Z. Deng, X. Jin, X. Zhao, W. Zeng, W. Xie, and H. Syo, "3D mesh animation compression based on adaptive spatio-temporal segmentation," *Proceedings of ACM SIGGGRAPH Symposium on Interactive 3D Graphics and Games (I3D'19)*, 2019.

[22] Y. Fan, L. Ming, L. Zhang, Z. Han, C. Wang, and Y. Tang, "PCA based 3D city model generalization for electricity simulation," *Procedia Computer Science, Vol. 122, 2017*, pp. 603-608.

[23] K. Sato, T. Nose, and A. Ito, "Synthesis of photo-realistic facial animation from text based on HMM and DNN with animation unit," *Smart Inoovation, Systems and Technologies book series (SIST, volume 64)*, 2016.

[24] I. T. Jolliffe, "Principal component analysis," 2nd *ed. Springer*, 2002.

[25] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory 28.2*, 1982, pp. 129-137.

[26] P. Hari, M. Kulmala, "Station for measuring ecosystem-athmosphere relations (SMEAR II)," *Boreal environment research 10: 315-322*, 2005.

[27] Y. Bodyanskiy, A. Deineko., A. Bondarchuk, and M. Shalamov, "Kernel Online System for Fast Principal Component Analysis and its Adaptive Learning," *International Journal of Computing, 20(2)*, 2021, pp. 175-180.