Principal Component Analysis Visualizations in State Discovery by Animating Exploration Results

Sirola, Miki

IEEE
2022-07-14

Sirola , M , Rinta-Koski , O-P , Le Ngu Nguyen & Hollmen , J 2022 , Principal Component Analysis Visualizations in State Discovery by Animating Exploration Results . in 2022 IEEE INTERNATIONAL CONFERENCE ON SMART COMPUTING (SMARTCOMP 2022) . IEEE , pp. 257-262 , 8th IEEE International Conference on Smart Computing (SMARTCOMP) , Espoo , Finland , 20/06/2022 . https://doi.org/10.1109/SMARTCOMP55677.2022.00064

http://hdl.handle.net/10138/350299
https://doi.org/10.1109/SMARTCOMP55677.2022.00064

unspecified
acceptedVersion

Principal Component Analysis Visualizations in State Discovery by Animating Exploration Results

Miki Sirola
Department of Agricultural sciences
University of Helsinki
Helsinki, Finland
miki.sirola@helsinki.fi

Olli-Pekka Rinta-Koski
Department of Computer Science
Aalto University
Espoo, Finland

Le Ngu Nguyen
Department of Computer Science
Aalto University
Espoo, Finland

Jaakko Hollmen
Department of Computer Science
Stockholm University and Aalto University
Espoo, Finland
jaakko.hollmen@aalto.fi

***Abstract*—Visualization is a key point in data exploration. In this paper we have emphasis in adding dynamic features by constructing exploration animations. We use Principal Component Analysis (PCA) in dimensionality reduction and K-means clustering algorithm in defining states. In predicting state transitions, we use Hidden Markov Model (HMM). Analyzed physical data is got from self-healing autonomous data centers. Our research methodology is to animate state transitions for data exploration in modern computerized environment. We use Jupyter tool and Python 3 programming language in our experimental realization. As results we get PCA animations for exploration purposes. Our approach is based on state discovery, where it is possible to find some physical interpretations for the defined states and state transitions. State structure and behaviour depend strongly on analyzed data.**

Keywords—Principal component analysis, visualization, state discovery, animated data exploration, clustering, machine learning, data mining.

I. Introduction

Data explorations lean on visualization of data in various forms. Mostly the visualizations are static views of the data from some particular point of view. The most common way is to draw time-series curves of variables involved. Two or three variables can also be drawn in a scatter plot forgetting the time-dimension. For example, density functions and histograms can be used to complete certain visual view. With high-dimensional data having big number of variables, projection methods can be used to reduce the dimensionality, and many kinds of graphical representations are constructed. Most visualizations are static pictures. In this paper we concentrate on including dynamic features in data visualization with animations to help in data exploration.

We use Principal Component Analysis (PCA) methodology in dimensionality reduction, and we separate different operational states from the data with K-means clustering algorithm. For predicting the state transitions, we use Hidden Markov model. To get the best possible view for data exploration we utilize animations for instance to help in recognizing operational states and state transitions between them. Our example data is from autonomous data centers.

Our work is carried out in a project of exploring data from autonomous and self-healing datacenters in close co-operation with Swedish partners [1]. We have published another paper [2] on the same project about state discovery and prediction of multivariate sensor data.

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].

Our interest is also to find out how Principal Component Analysis (PCA) techniques have been utilized in animation research. Our approach 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].

The paper structure is organized as follows. In Section II the methods and tools that we use in this work are presented including our research methodology approach. In Section III available data is presented. In Section IV we present our visualization results in PCA animations. In Section V more exploration animation examples are presented and used in process state discovery. After that in discussion and conclusion sections (VI and VII) the results are discussed through thorough interpretations and conclusions are drawn.

II. Methods and Tools

Our research methodology is prototyping animated data exploration animations in a modern computerized environment. As a result, we present ideas how to represent data exploration results to the viewer so that both data structure and evolvement in time comes clear.

Principal Component Analysis (PCA) [10] is a dimensionality reduction method. PCA compresses N variables to defined number of projections, where first component contains most variance in data, next component contains most of remaining variance, etc. Two or three first PCA components are commonly used in visualizations.

PCA visualizations describe well data structure and properties, transients and state transitions are easy to detect. There is a possibility to look in detail portion or effectivity of each variable in the main PCA components and check the dominance of each of the first PCA components in the analysis.

As additional support methods we use K-means clustering [11] to illustrate local data accumulation and densities. Hidden-Markov Model [12] can be used in prediction of state transitions. In reference [2], we provide deeper analysis of the used methods in state discovery.

PCA animations illustrate states and state transitions in data. Visualizations including animations can be used also in anomaly detection recognizing failure states in data. We present some examples to demonstrate the potential in PCA visualizations.

For realizing our exploration animations, we use Python 3 programming language in Jupyter tool and programming environment.

III. Data

We use data from autonomous and self-healing datacenters in our analysis and exploration animations. The data includes such measurements as power, CPU temperatures, fan signal and fan power, etc., and chamber temperature and ambient temperature of a datacenter. We have several datasets including various combinations of measured variables including in addition such variables as humidity, utilization, fan speed and memory utilization. The largest dataset has over two million measurement samples and about 300 variables. The largest number of variables in our sets is almost 600 having a bit less measurement samples. We have both datasets where we have all these variable types in one set, and such where there are huge number of variables from one variable type only in one set, for instance only temperatures or only powers.

IV. Visualization Results

We explore datasets with PCA animations that show for instance the state behaviour in time and state transitions. We present here two first PCA components in a 20-time points time window in a time-series curve presentation and a scatter-plot presentation, and a cumulative three-dimensional PCA presentation including clustered state information.

In Figure 1 two first PCA components in a time-series presentation are gliding through a 20-time points time window over the whole dataset. The animation through the whole dataset takes about 1 minute and 40 seconds. The dataset includes 5072 measurement samples and 44 variables without timestamp. The whole time-period in this dataset is about 42 hours. The sample interval is 30 seconds, so 10 minutes is in one time-window. Variations in the component time series show state transitions in the data.

In Figure 2 two first PCA components are gliding through a 20-time points time window over the whole dataset in a scatter plot presentation. The location of the dots shows current state in the data and when they move into another location, a state transition occurs. Also, this exploration animation takes about 1 minute and 40 seconds. The same dataset as before is used in the exploration animations in Figure 2 and Figure 3.

In Figure 3 three first PCA components appear cumulatively to a three-dimensional scatter plot beginning from first time point and going through the whole PCA result of the dataset in time order. The PCA result is clustered into three clusters that we call states, and this state information is pointed out with three different colours in this exploration animation. The states and the state transitions are well illustrated in this animation, and the viewer gets a good view of the whole physical development path that the dataset contains. The length of this exploration animation is also about 1 minute 40 seconds.

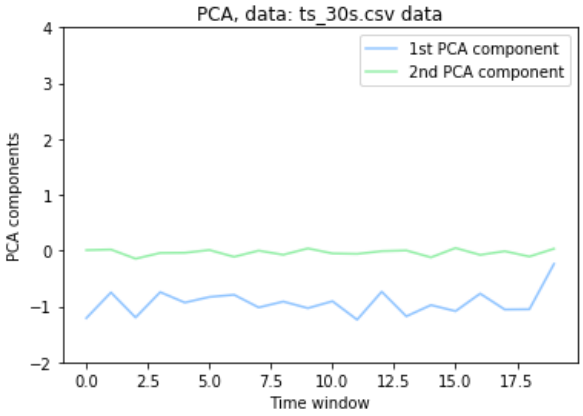

Fig. 1. Two first PCA components in 20-time points time window sliding through the whole dataset.

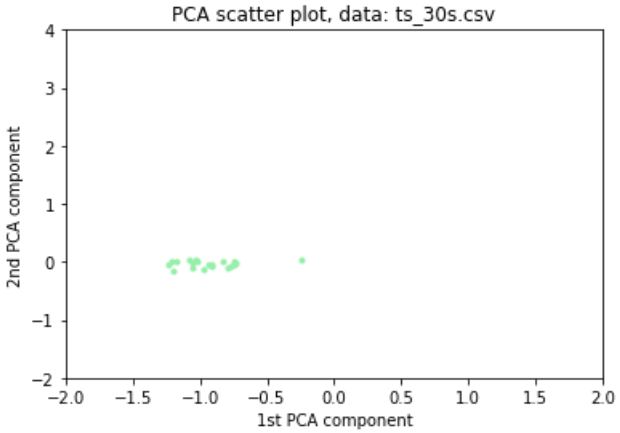

Fig. 2. Two first PCA components in a scatter plot sliding through the whole dataset in 20-time points time window.

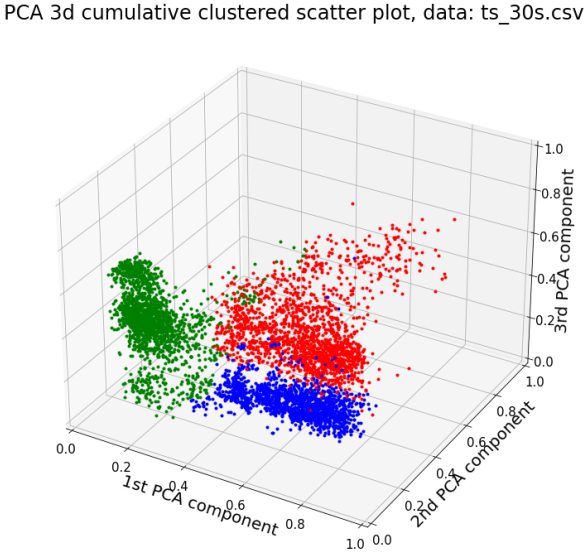

Fig. 3. Cumulative three-dimensional clustered scatter plot visualizing states and state transitions through the whole dataset in time order.

These three PCA exploration animation-types give a good view what it is possible to show for the viewer with this methodology. We have repeated similar animations with various other datasets from the autonomous data center, and with these exploration tools we are able to view and define different state structures in many different datasets including various time-periods and variable combinations.

We have made 'worm plot' versions of several scatter plot type exploration animations, both in time-window type animations and cumulative animations. The difference is that all following points are joint together with a line, when instead of 'fly swarm' a long worm appears to the screen. In Figure 4 in the following section there is one example of a worm plot, see second undermost plot. The animation length is about 1 minute and 40 seconds.

V. State Discovery with Example Explorations

We have very limited information available of the physical process and no datacenter details, which makes the interpretation of states difficult and challenging. Looking the PCA loadings in our analysis we get information about which variables are dominating in each PCA components though, and that opens some opportunities to interpret the connections between states and measured variables behaviour.

We notice some obvious correlations (and reverse correlations) between some variables and variable groups. Typically: powers, temperatures, utilization, and fan speed correlate more or less with each other, while humidity value has reverse correlation. Effect of some variables may be delayed. The room temperature control may also have affects to some variables in addition to the heat production of the data center equipment itself.

PCA loadings tell the variance stored in each PCA component. In our dataset presented in previous section the 1st PCA component has 54% of the total variance, 2nd PCA component has 14%, and the 3rd PCA component has 4% of variance. In Table 1 are listed the most effecting variables to each PCA component. The scale in the left column is from 0 to 1 (from no domination to full domination). The numbers in the variables refer to the six channels. In CPU temperatures the second number refer to two different cores in each channel.

From Table 1 and variable plots, we can make following observations considering states and state transitions. Seems that power and CPU load (with some correlating variables) are the dominating factors composing the 1st PCA component. The fan signal (affecting to fan power, fan speed, etc.) is the reason for the smaller and bigger variations in 2nd PCA component. The 3rd PCA component seems to follow some delayed temperature and possibly humidity changes.

The 1st PCA component defines the states 'idle' and 'CPU power', and the third state comes from the 2nd and 3rd PCA component variations. The vibration in fan signal in 2nd PCA component and delayed temperature changes in 3rd PCA component define the two states in not idle mode. These two states could be named e.g. 'cooler power on' and 'hotter power on' (and more stable). The stability of CPU load also has effect defining these two powered states, maybe because the average load is higher in stable load than in oscillating load.

Finally, we could name these three states as: idle, hot CPU load, cool CPU load (varying fan signal), see Figure 3 in previous section. With similar logic we can name seven states as: hot idle, cool idle, hot even CPU load, hot varying CPU load (varying fan signal), cooler even CPU load, cooler varying CPU load (varying fan signal), strongly varying fan signal (varying CPU load), see Figure 4 (uppermost plot)

This seventh state appears in the end part of the data where fan signal is strongly oscillating. This state differs most from all the other states and could even be a failure state. This kind of example study demonstrates that our approach constructing states according to the physical behaviour of the process may also reveal failure states, and therefore would fit well also to anomaly detection.

TABLE I. Dominating variables in each PCA component. Note that the higher dominance is down in the table.

pca.components_	1st PCA component	2nd PCA component	3rd PCA component
➢ 0.15	yS22, yS31, yS41, yS42, yS51, yS52, ys61, yS62	uF1, uF2, uF3, uF6, pF3	
➢ 0.20	pS1, pS2, yS21, yS32, pS4, pS5	T_c *, uF5, nF1, nF2, nF3, nF4, nF5, nF6	
➢ 0.30	pS3, pS6		
➢ 0.45			T_a *
➢ 0.75			T_c *

pS = power
yS = CPU temperature
uF = fan signal
nF = fan related unknown variable
pF = fan power
T_c = chamber temperature
T_a = ambient temperature
star (*) = - (reverse correlation)
xS = utilization (very small involvement and therefore not present in the table)

Described state behaviour is valid only to this type of data having certain set of measured variables. There seem to be different state characteristics in different types of data. We have experimented also e.g. large amounts of pure temperature data (or some other single measured variable types) from autonomous datacenters and noticed that it is possible to define similar states also there, but the characteristics are very different. Temperature data typically have very clear and separable clusters – states. There are also less such characteristics as variable delays, and the states form mostly from different power levels. Sometimes the vibration

effect form two alternating well-separated states in one power level in addition.

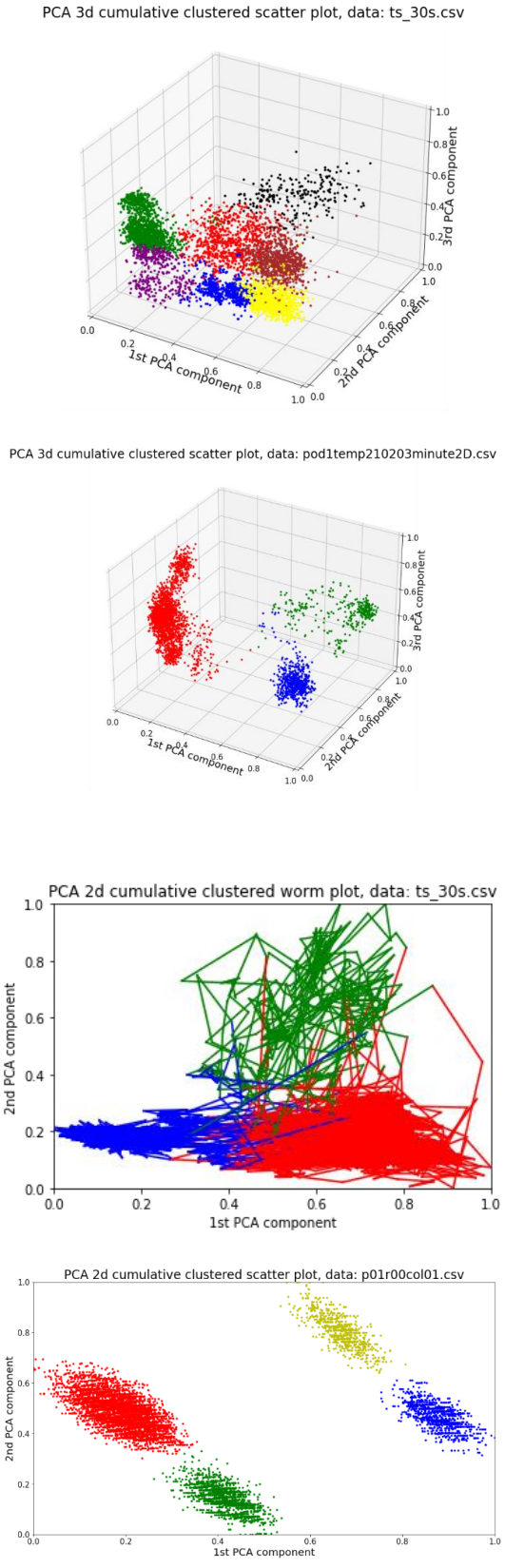

Fig. 4. More state discovery example exploration animations. The uppermost plot and the scond undermost plot are from the same dataset that was presented in previous section. In the second uppermost plot there is temperature dataset and in the undermost plot dataset including powers, temperatures, and fan signals.

TABLE II. DOMINATING VARIABLES IN EACH PCA COMPONENT IN ANOTHER DATASET. NOTE THAT THE HIGHER DOMINANCE IS DOWN IN THE TABLE.

pca.components_	1st PCA component	2nd PCA component	3rd PCA component
➢ 0.20	s(fd), T(r2), s(hd)	P(c)	s(fd) *, s(hd) *, P(f)
➢ 0.30	T(r1), P(c)	T(r2) *, T(dd) *	P(c) *
➢ 0.50		P(f)	
➢ 0.60		T(r1) *	
➢ 0.70	P(f)		
➢ 0.80			T (dd)
Relative variance	0.60	0.24	0.09

P(f) = fan.default.power
P(c) = cooling.power
T(r1) = return.1.default.temperature
T(r2) = return.2.default.temperature
T(dd) = discharge.default.temperature
s(fd) = fan.default.signal
s(hd) = hrk_s.default.signal
star (*) = - (reverse correlation)

In uppermost plot in Figure 4 the first presented dataset in the previous section is divided into seven states. In the first dataset there occurs most transitions between different states to many directions. The animation length is about 1 minute and 40 seconds.

In temperature dataset, see second uppermost plot in Figure 4, three-state division is used. First a red 'Norwegian map' appears in the left in the screen. Then a clear state transition occurs through Sweden to Southern Finland (thinking of Scandinavian map) where this new state is in green colour. Finally, towards the end another state transition occurs somewhere into the Baltic area where this last state is seen in blue colour. These states are rather well ordered, and the follow each other in strict time order. This dataset includes 2880 measurement samples and 498 variables. The animation length is a little under one minute.

In the dataset including powers, temperatures, and fan signals, see undermost plot in Figure 4, in the beginning two well-separated states in red and yellow colours in the figure are alternating due to a vibration effect. After one clear state transition similar vibration occurs between two other states (with green and blue colours in the figure) until the end of the dataset. These states are also well ordered. The state transition is due to changing power level and the vibration between two alternating states is due to fan signal. This conclusion can be interpreted from the PCA loadings tables of this dataset seen in Table 2. This dataset includes 10419 measurement samples

and 12 variables. The animation length is about 1 minute and 45 seconds.

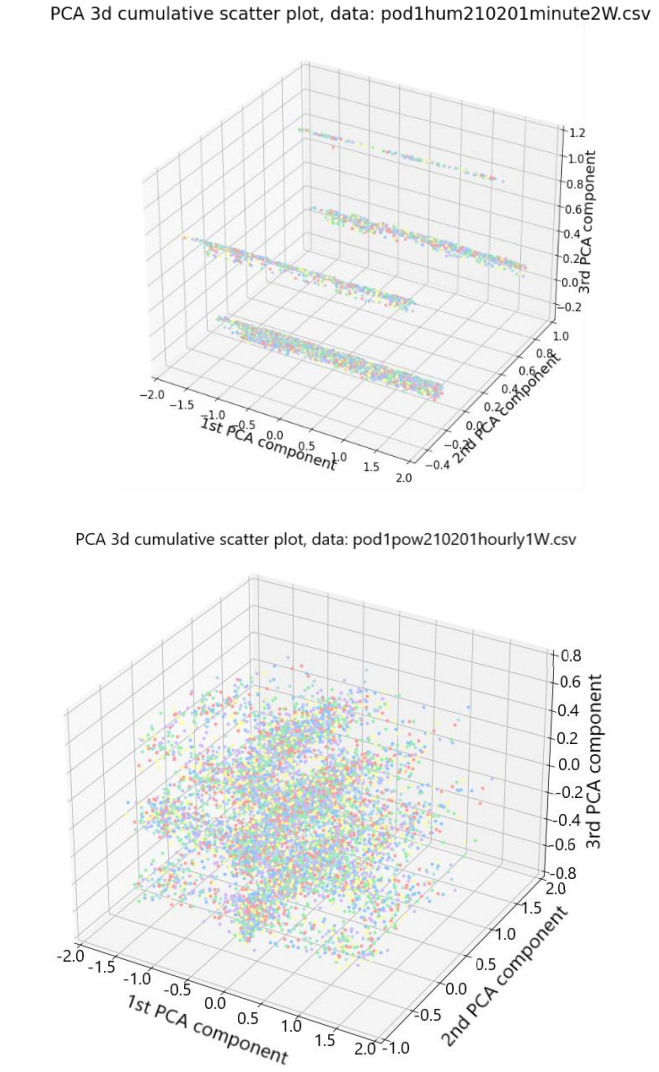

Fig. 5. Exploring humidity data (upparmost plot) and power data (middle plot and undermost plot).

We have explored large amounts of datasets, where variables are in separated variable groups such as temperatures, powers, humidity, etc. Each variable group have little different characteristics, which we can also point out clearly with our exploration animations. In Figure 5 is seen examples humidity and power data, and in Figure 4 examples of temperature data in addition to several variable groups in combination in the same dataset.

The humidity data, see Figure 5 uppermost plot, seems to alternate between four states having similar vibration behaviour effect as some temperatures showed out before. Also, the power data, see Figure 5 middle plot and undermost plot, has a lot of vibrations, but not so clearly separable vibration states. The states in Figure 5 undermost plot are clearly time separated. In the beginning of the dataset we are fully in red-marked state. The first state transition jumps in the blue-marked state. The later a rather shot jump to the green-marked state follows ending up back to the blue-marked state.

The humidity dataset in uppermost plot in Figure 5 includes 20160 measurement samples and 21 variables. The animation length is about 1 minute and 40 seconds. The larger power dataset in middle-plot in Figure 5 has 10021 measurement samples and 568 variables, and the other powerset in undermost plot in Figure 5 has 8640 measurement samples and 348 variables. The animation lengths are about 1 minute and 40 seconds and a little under one and a half minute.

VI. Discussion

We have used PCA result presentations in time series and scatter plots. We have presented both time-window and cumulative approach. It is of course possible to expand the used concepts to many directions, and animate for instance density functions, histograms, or state probabilities. Our examples show enough of the opening possibilities of dynamical exploration though to help in understanding the structure and physical behaviour in the data.

We have used PCA for reducing the dimensions, K-means clustering algorithm for defining the states and Hidden Markov Model in state predictions. We do not have any example of Hidden Markov Model in this paper, because the focus is in visualization, and HMM does not give more input in this respect.

We present basically three different dynamic concepts for PCA visualizations, which each have some additional variations. Two first concepts use a time window approach showing either selected number of PCA components as time flows, or a scatter plot of the two first PCA components. The latter one can be presented in a line form (‘worm plot’) instead of point form (‘fly-swarm plot’).

The third concept use cumulative approach, where two or three component PCA analysis result appears to the screen point by point in time order. Here, it is possible to add in the state information by colouring each cluster with different colours. Also, here point form and line form are possible options. In our opinion the three-dimensional scatter plot in point form including state information in different colours is the most promising concept for state discovery.

We are fully aware of the difficulty to describe dynamic figures and animations in a scientific paper with the only possible options such as written text and static figures. We have tried to do this as clear as possible by detailed explanations and many illustrative examples of practical exploration tasks.

In the section about state discovery, we show how the state structure including state transitions depend on the data type we are analyzing. Although all our data is from self-healing, autonomous data centers, the grouping of data vary in our

collection of datasets. Different variables and different combinations of variables formulates a little different characteristic in the PCA analysis. The state definitions on the top of the analysis are constructed with clustering methods, mainly using K-means clustering.

We have tested to analyze same datasets by using two different types of scalers. All examples in this paper are scaled with Min-Max scaler [13], but we have also tried out to use Standard scaler [13] in some cases and noticed some differences in the output. These two scalers have a little different approach in scaling and gives out little differently distorted data. In data manipulation it is very difficult to avoid any kind of distortion. It is important though to be aware of the possible distortions.

Min-Max scaler gives out rather nice from for visualization purposes, which is also good and clear form for defining physical states out of data. Scaling helps especially in getting better clustering results with our method. There are some weaknesses in every scaler though. Min-max scaler is not the best possible, if there are left outliers in the data. The outlier may define the scale so that transitions in the normal scale look rather small. In our data examples it seems that Min-Max scaler for instance filters out some sharp peaks out of data, which may have effect to the physical behaviour. If the whole range of some variable is not representative in the dataset, it causes problems with every scaler.

VII. Conclusion

We have explored PCA visualizations focusing on dynamic behaviour in state discovery to demonstrate how the time dimension helps the viewer to perceive the states and state transitions in the data and make a mental illustration of the data structure. To learn state and state transition behaviour we have done visually dynamic data exploration in addition to static one. We have constructed PCA animations that clearly demonstrate how states compose and when the state transitions occur in a cumulative process as seen in the previous sections of this paper.

The contribution of our paper is in the case study in this domain, where we show how our Principal Component Analysis animation constructions help in data exploration and state discovery on multivariate sensor data.

Acknowledgment

We acknowledge the computational resources by the Aalto Science-IT project. We thank Rickard Brännvall and Jonas Gustafsson of RISE ICE Datacenter for their help with the datasets.

References

[1] R. Brännvall, "Machine learning based control of small-scale autonomous data centers," Lic.Sc. thesis in Luleå University of Technology, 2020.

[2] 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.

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

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

[12] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE 77.2, 1989, pp. 257-286.

[13] J. Han, Jiawei. M. Kamber, J. Pei, Data Transformation and Data Discretization, Data Mining: Concepts and Techniques. Elsevier, 2011, pp. 111–118.