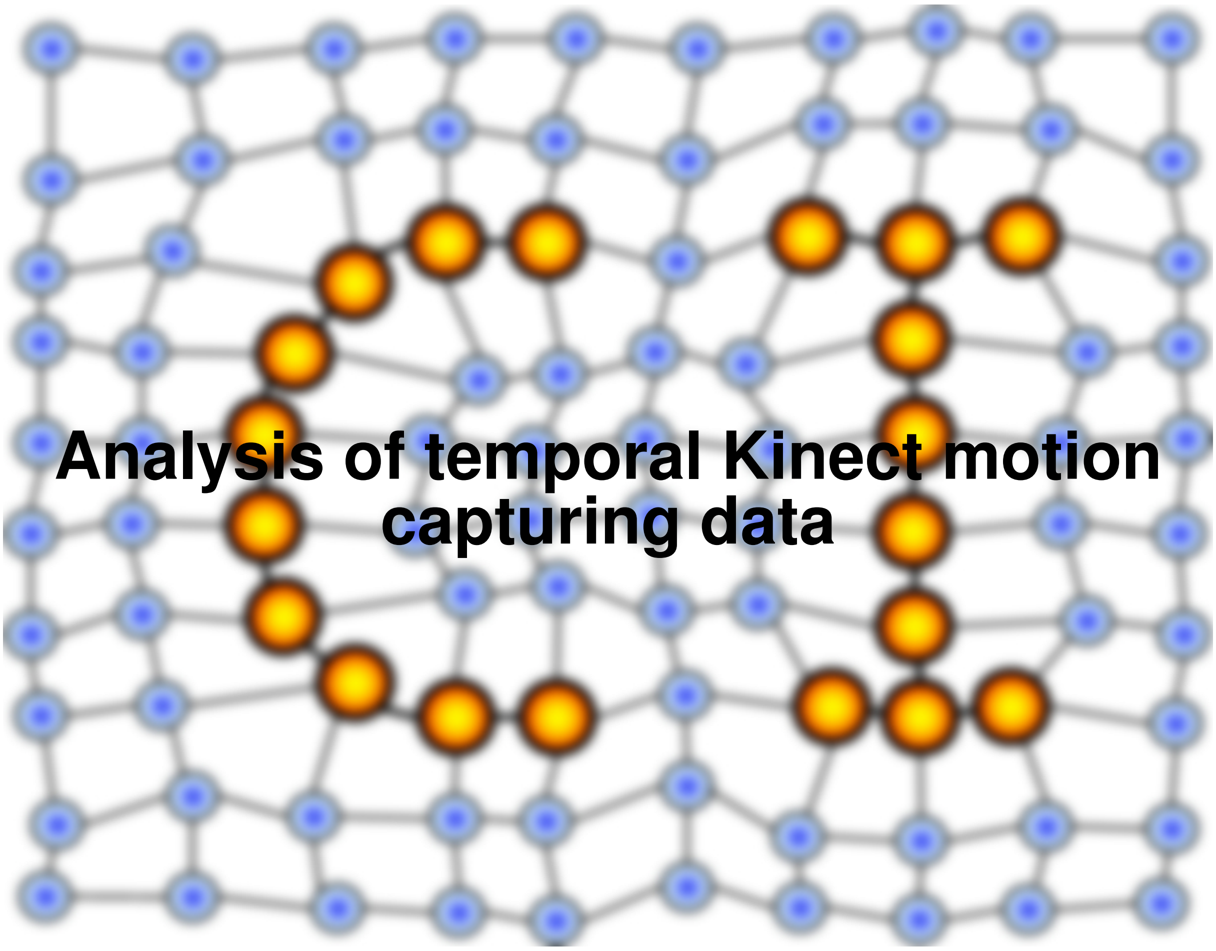
MACHINE LEARNING REPORTS



Analysis of temporal Kinect motion capturing data

Report 05/2013
Submitted: 13.10.2013
Published: 31.10.2013

Frank-Michael Schleif[1], Thomas Villmann[1]
(1) University of Applied Sciences Mittweida,Technikumplatz 17,09648 Mittweida, Germany

Abstract

Motion capturing is a popular technique in movie production, for animated cartoons, but also effectively used in computer games. The development of computer games and animated cartoons requires a lot of manual work getting even more complicated by the use of motion capturing data. Novel approaches try to employ machine learning algorithms to support the computer assisted development of games and cartoons. The major task we consider is to predict movements and gestures from motion capturing data. In a small study we compare different machine learning approaches for the analysis of static and dynamic motion capturing data and highlight current potentials and challenges.

1 Introduction

The virtual production of animations and effects for entertainment became more and more important in the last decade. With the raise of simple and effective motion capturing systems it became possible to transfer the movements of real actors to virtual avatars. This concept has now also moved to the field of game development and the production of animated cartoons. To simplify the on hand production process, improved computer assistive systems are desirable. One interesting task is to predict movements and gestures from motion capturing data, in general human motions. A recent introduction to the field can be found in [15].

These prediction can be used to provide pre-defined models and data for the game developer or the animator based on available databases. Here we address the objective to learn predictive models from a given set of static or dynamic motion capturing data. Our objective is to predict a class of similar motions, available in a training database. This predicted class could than be further analyzed to provide e.g. a ranking of similar motions. In the first part of the paper we briefly review related work, followed by a brief discussion of different standard and recently proposed methods which can be used to generate classification models based on static or dynamic motion capturing data. We will focus on dynamic motion capturing data, in the form of larger, multidimensional time series. Subsequently we show results for a static and a dynamic motion capturing study of human gestures and conclude with open issues and future research questions.

2 Related work

Motion capturing is a technology to record any kind of movements such that the movements can be analyzed by electronical systems. Here we consider the case of human motion capturing and accordingly we are interested on the specific type of movements, e.g. in relation to a specific body part. Motion capturing has attracted a lot of interest in recent years, mainly driven by technological improvements of the acquisition systems. While many approaches are still based on the professional Vicon system [2] also the simple Kinect [1] technology was found to be very effective. In our study we are interested on gesture recognition for a wider group of applicants such that the Kinect is an interesting option, given the results are reasonable accurate. Early approaches typically analyze video sequences or multi-views and try to extract the body movements from the consecutive image frames [24, 21]. In the case of cartoon animation also so called re-targeting was proposed, where the movements of cartoon figures are transformed to new characters using rigid transformations [17, 5]. These approaches obviously lack or do not need 3D information and can not be effectively used for the motion capturing data considered here. The more recent approaches use in general either the Vicon system or the Kinect, depending on the specific objective [3, 20, **?**]. The later one is favored in applications were high accurate position measurements and high frequency measurements are less relevant. Motion capturing has been applied in very different areas, one major objective is gait recognition. This is of interest e.g. in medicine, in sport research and bio-cybernetic research as well as robotics [20, 18]. Another field is the cartoon, movie and game production but also the field of surveillance systems [25, 13, 11]. Here the objective is in general to reconstruct the trace of the observed movement for further post-analysis steps. A related problem is the identification of the silhouette of a body in a complex scenery [7, 6]. All these methods try to obtain an in deep analysis of the data, reconstruct a complex body model or operate in a complicated scenery. In our work we will try to provide a classification model on rather simple features, directly obtained from the Kinect systems. The underlying, hidden process and the relation between the measured features shall be learning using concepts from machine learning. Specifically we focus on Hidden Markov Models (HMM) and Learning Vector Quantization (LVQ) originally proposed in the neural network domain. Both HMM and neural network concepts have been used already before in the context of motion capturing and human action recognition [12, 16, 14] but not yet in the line of supervised gesture recognition based on Kinect motion capturing data.

3 Methods

The motion capturing data are in general given as multiple measurements of references points in 3D, measured over a number of time steps, referred to as frames in the following. The measurement points are reference markers between rigid body parts e.g. bones of an assumed body model. This assumption is reasonable to simplify the measurement process. Instead to measure a large point cloud it is sufficient to focus on the reference points only. A detailed description of the process, focused on the Kinect system is given in [14]. In the following we consider single frames (static motion capturing data) and dynamic motion capturing data, based on multiple consecutive frames. The static data can be considered as standard datasets represented by a matrix of samples vs. features. Accordingly a multitude of methods can be used to define discrimination models. In the following we will consider a learning vector quantizer (LVQ) extended by relevance learning as defined in [19], referred to as Generalized Matrix LVQ (GMLVQ). GMLVQ defines a classification model based on so called prototypes (similar as cluster centers in k-means), which are prototypical representants of the dataset, maximizing class separation. The model is based on a cost function employing a parametrized Euclidean distance. The adaptation of the metric parameters is also focused on class separation and can be subsequently inspected to highlight most discriminating features in a univariate or class correlative way. In the experiments we will use one prototype per class and train the GMLVQ model until convergence of the cost function is achieved. A detailed mathematical derivation of GMLVQ including a discussion on relevance learning can be found in [19] and is skipped here for brevity. Additionally we analyze the static data using a linear Support Vector Machine (SVM) [23] and a SVM with an extreme learning (ELM) kernel [9].

For the dynamic data sets we first have to take into account that the number of frames differs between the samples. Each sample can be considered as a, in general, rather short multivariate times series. Classical dynamic time warping (DTW), see e.g. [8], could be used but is complicated for multi-dimensional data. An alternative is to employ Hidden Markov Models (HMM). We assume that an observed gesture, characterized by e.g. a different speed and person specific movements is caused by a simple underlying generation model. Further we assume a limited number of hidden states, defining the basic primitives of a movement. The structure of our HMM is depicted in Figure 1. Obviously, it would be possible to reduce the number of free parameters of the HMM by limiting the possible transitions to e.g loop forward transitions only or to define alternative HMM structures, but this will be addressed in future experiments. For dynamic data the HMM models the temporal dynamic (transition probabilities) in consecutive frames and the most likely frame (emission probabilities) within a temporal context. In our case we have a quite small database and the target classes contain similar movements, e.g. all focusing on the upper body defined by very few reference points in a $3D$ space. The pre-defined HMM structure has a direct impact on the representation accuracy of the data and more complex HMMs, by e.g. a larger number of states, could be expected to perform better. Interestingly a doubling of states from 4 to 8 had only a weak effect such that we kept the 4 state representation. This can partially be explained by the increased number of parameters which have to be estimated from the data and can also introduce extra noise in the representation model.

The number of states can be considered as a meta parameter and could be optimized by a double crossvalidation scheme, given a reasonable number of data points. Since the HMM is a representation model we combine it with different classifiers. Thereby we consider either a single HMM for the whole dataset and use a gradient representation per sample (see e.g. [22]) or one HMM per class, alternatively we also employ a type of fisher kernel learning (FKL) as suggested in [22]. We use the following classifiers as a on top method to the HMM model: a nearest neighbor classifier (NN), a soft-max classifier (SOFTMAX) a maximum aposterior classifier (MAP) and a support vector machine classifier (SVM) see e.g. [4]. Another alternative is to align the different sequences to obtain sample representations of equal length and to process the data again by some of the classifiers mentioned above. Here we use a fast global alignment kernel (GAK) [8] in favor of a classical dynamic time warping (DTW) or its extensions for multi-dimensional due to time constraints. All datasets are evaluated

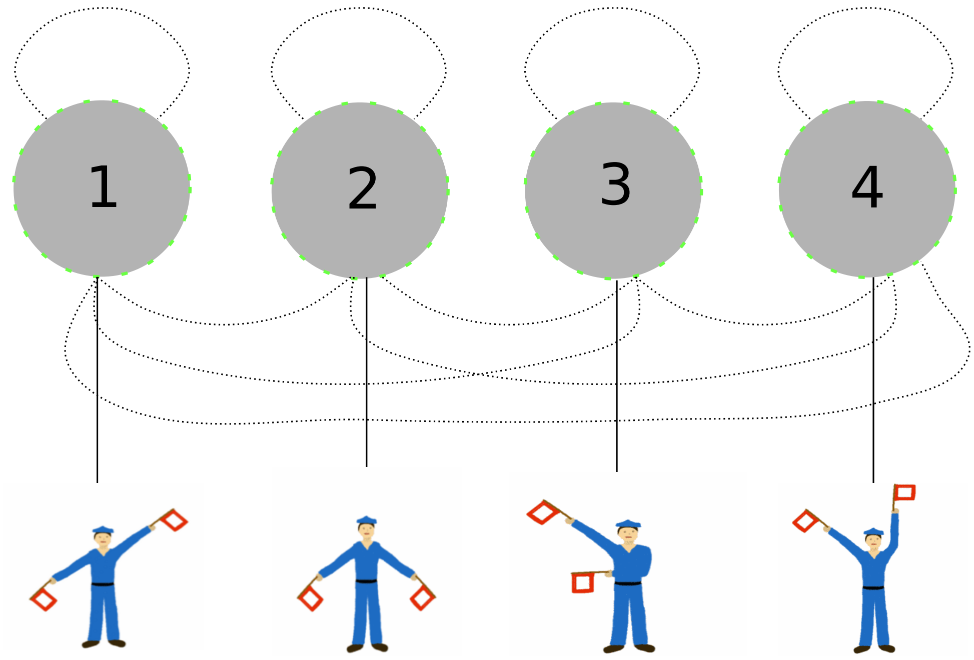


Figure 1: Sample HMM with 4 hidden states. The transitions are indicated by doted lines and the emissions by straight lines. The emissions are only exemplary and are taken from *wikimedia*.

Representer	DS1	DS2
GMLVQ	$98.14\% \pm 1.62$	$89.85\% \pm 12.59$
linear SVM	$99.2\% \pm 1.38$	$88.23\% \pm 10.71$
ELM SVM	$98.39\% \pm 2.50$	$79.43\% \pm 14.86$

Table 1: Prediction accuracy (mean/std) for the static motion capturing data (DS1) and (DS2) using different classifier approaches.

in a leave-one-out crossvalidation (LOO) such that all samples measured for a single persons are kept out. The objective is accordingly to predict the movements of this omitted person based on the data of the other persons.

4 Experiments

4.1 Static datasets

The static dataset consists of 2080 frames of 20 different gestures measured with multiple repeats for 10 persons. The gesture is captured by 20 body measurements in 3D leading to 60 features per frame. A view of typical poses and labels is shown in Figure 2. The static data have been analyzed in two settings: (1) considering a 5 class classification problem (DS1), combining the classes for each position and (2) considering all given 20 class labels (DS2).

The dataset has been normalized to $N(0, 1)$.

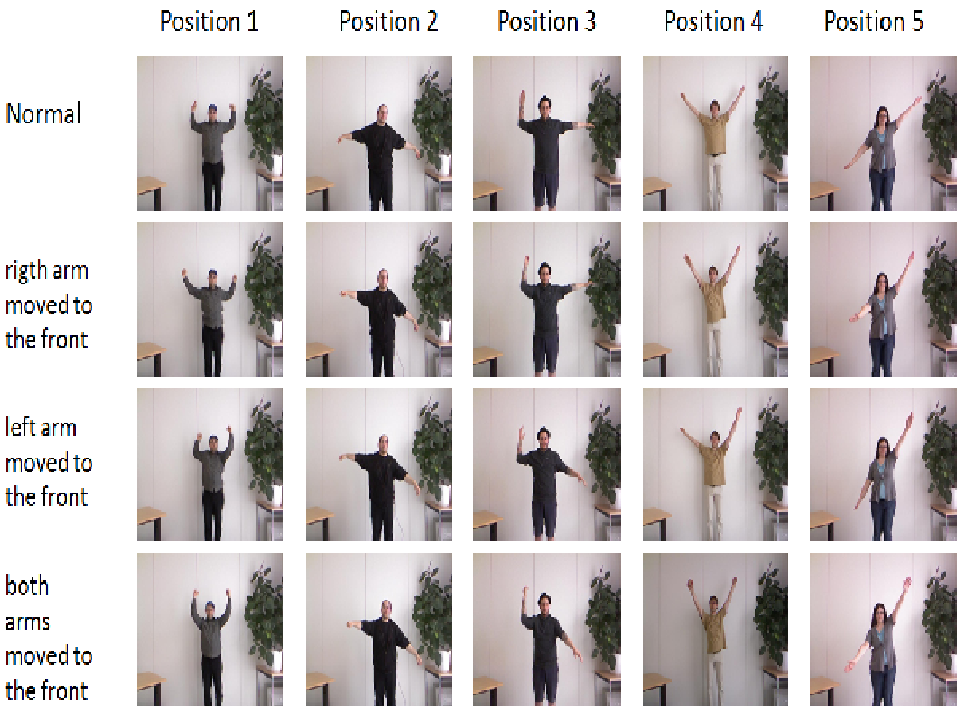

Position 1
Position 2
Position 3
Position 4
Position 5
Normal
rigth arm moved to the front
left arm moved to the front
both arms moved to the front

Figure 2: 5 classes of gestures measured as static single frames.

We observe that for DS1 all algorithm achieve almost perfect models whereas for DS2 with a larger number of classes the results is a bit worse but still at almost 90% for GMLVQ. An additional exploration of the learned parametric distance matrix (class correlation matrix) in Figure 3 can be used to identify the most relevant features and correlated sensors (bright points).

4.2 Dynamic dataset

The dynamic dataset consists of 766 frames of 13 different gestures measured with multiple repeats for 6 persons. The gesture is captured by multiple body measurements. The head rotation is measured by a signal in 3D. Further the movement and position of 10 body parts is captured by 11 measurements include 3D position and angle information. Accordingly, each frame is measured by a 113 dimensional feature vector. For each gesture multiple frames are captured in a range of 2 to 115 frames with a gap of 5 frames each. The gestures are similar as shown in Figure 2 but with an associated movement. The following classes are defined *flap,head nodding,wink left, wink right,cross-armed,shaking, left shoulder shaking, right shoulder shaking, bending, left bending, right bending, nodding to the left shoulder, nodding to the right shoulder*. The dataset has been normalized to $N(0,1)$. We consider this dataset as DS3 in the following.

The evaluation was done by a LOO over the probands which can be analyzed across the different models to identify prediction characteristics for the different probands.

We observe that all probands show a strong variance with respect to their prediction accuracy but for two probands we can correctly predict the 13 gestures in above 50% of the cases. If we focus on the best observed model from Table 2 (FKL with SVM-

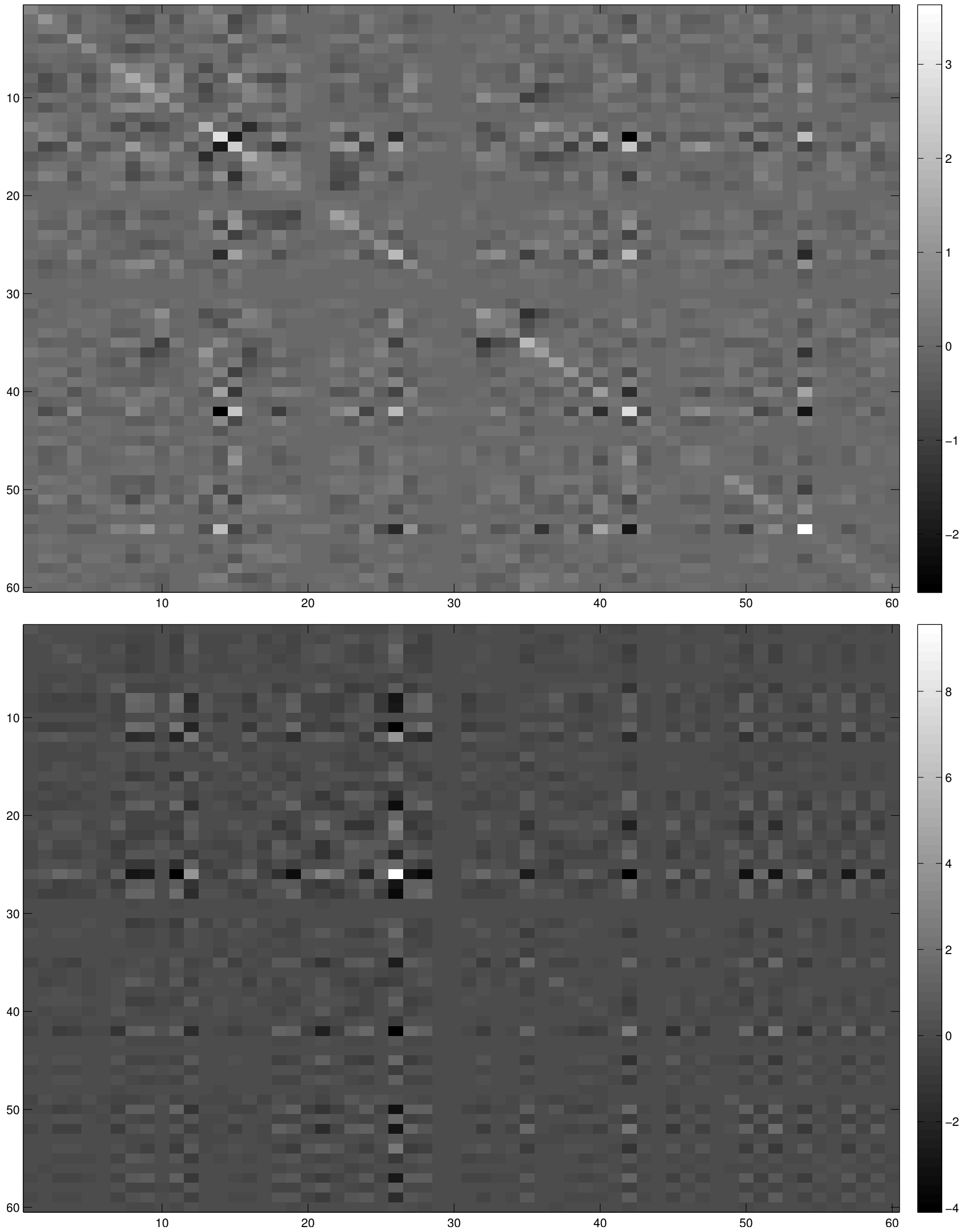


Figure 3: Class correlation matrix for DS1 and DS2 of GMLVQ.

Representer	Classifier	Prediction accuracy
HMM	NN	$47.05\% \pm 10.72$
HMM	MAP	$25.01\% \pm 7.28$
HMM	SOFTMAX	$53.27\% \pm 10.89$
HMM	SVM-Linear	$48.54\% \pm 8.12$
HMM	SVM-ELM	$55.14\% \pm 12.00$
∇ HMM	SVM-Linear	$20.92\% \pm 5.44$
FKL	SVM-Linear	$63.52\% \pm 15.40$
FKL	SVM-Elm	$61.32\% \pm 20.70$
GAK	SVM	$54.97\% \pm 9.87$

Table 2: Prediction accuracy (mean/std) for the dynamic motion capturing data (DS3) using different representation and classifier approaches.

1	2	3	4	5	6
$38.91\% \pm 13.65$	$56.59\% \pm 22.37$	$37.26\% \pm 14.35$	$33.64\% \pm 13.18$	$54.29\% \pm 24.00$	$49.90\% \pm 16.7$

Table 3: Prediction accuracy (mean/std) for the 6 probands.

Linear), we get the following confusion matrix shown in Figure 4. We observe that the class 1 (flap) is consistently predicted over all probands, also class 5 (cross-armed), 10 (left bending) and 13 (nodding to the right shoulder) show good results. A confusion between semantically related classes occurs e.g. for the class 3 (wink to left) and class 4 (wink to right). The worst classification accuracy can be observed for class 6, 7 (shaking and left shoulder shaking) and class 9 (bending).

5 Conclusion

We explored different algorithms to predict gestures of static and dynamic motion capturing data. The prediction of the intended gestures is most effective for static measurements with almost 90% also for a 20 class experiment. Using GMLVQ one can additionally identify the most relevant measurement channels, which can be used to simplify the measurement process. For the very challenging dynamic motion capturing data (DS3), focusing on the prediction of gestures with respect to 13 classes, the accuracy suffers, but is still surprisingly good with above 63% for the best algorithm. We found that the use of a non-standard type of metric, either by means of a parametrized metric like in GMLVQ or in form of a reliable kernel improved the overall results. For the dynamic data set we also observe that the used representation matters most and the fisher kernel description, which also takes label information into account, was most successful. The specific post-classifier is less relevant in general. In summary we found that machine learning approaches can be effectively used to predict human gestures with quite good accuracy. In future work improved pre-processing of the data, e.g. by invariant coding of the kinematic reference points and hierarchical classification models will be address.

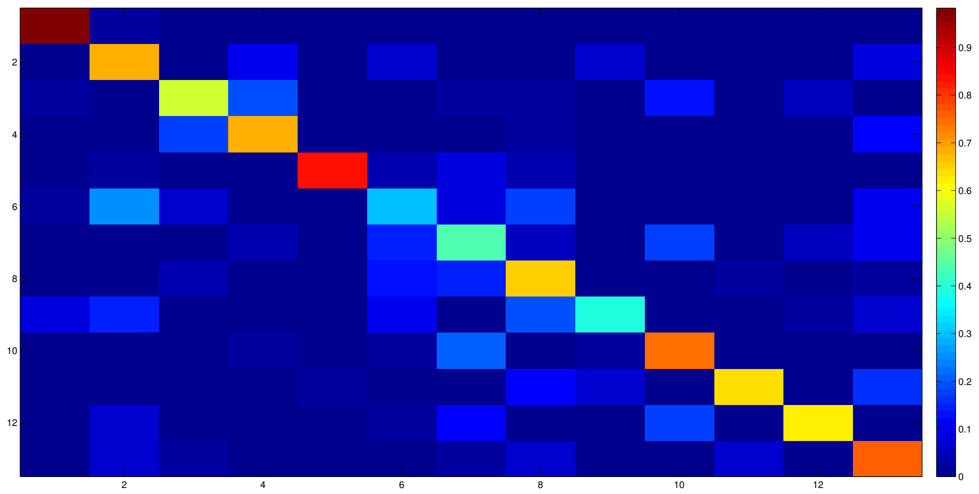


Figure 4: Confusion matrix with respect to the 13 classes in the dynamic motion capturing dataset.

Acknowledgments: We would like to thank the people from the *Gamecast* group, University of Applied Sciences Mittweida in Germany for providing the static and dynamic motion data sets. We also thank Marika Kaden, University of Applied Sciences Mittweida in Germany, for earlier work on the static dataset. Further we thank Bassam Mokbel, University of Bielefeld in Germany, for interesting discussions and suggestions with respect to the Kinect system and the data pre-processing.

References

[1] Microsoft kinect, 2013.

[2] Vicon motion systems ltd, 2013.

[3] S. Bailey and B. Bodenheimer. A comparison of motion capture data recorded from a vicon system and a microsoft kinect sensor. page 121, 2012.

[4] C. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006.

[5] C. Bregler, L. Loeb, E. Chuang, and H. Deshpande. Turning to the masters: motion capturing cartoons. *ACM Trans. Graph.*, 21(3):399–407, 2002.

[6] G. Burghouts and K. Schutte. Spatio-temporal layout of human actions for improved bag-of-words action detection. *Pattern Recognition Letters*, 34(15):1861 – 1869, 2013.

[7] A. A. Chaaraoui, P. Climent-Perez, and F. Florez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799 – 1807, 2013.

[8] F. Coleca, S. Klement, T. Martinetz, and E. Barth. Real-time skeleton tracking for embedded systems. In *Mobile Computational Photography*, volume 8667D. Proceedings of SPIE, 2013.

[9] M. Cuturi. Fast global alignment kernels. In Getoor and Scheffer [10], pages 929–936.

[10] B. Frénay and M. Verleysen. Parameter-insensitive kernel in extreme learning for non-linear support vector regression. *Neurocomputing*, 74(16):2526–2531, 2011.

[11] L. Getoor and T. Scheffer, editors. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. Omnipress, 2011.

[12] G. Goudelis, K. Karpouzis, and S. D. Kollias. Exploring trace transform for robust human action recognition. *Pattern Recognition*, 46(12):3238–3248, 2013.

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[14] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4), 2013.

[15] B. Mokbel, M. Heinz, and G. Zentgraf. Analyzing motion data by clustering with metric adaptation. In T. Villmann and F.-M. Schleif, editors, *Proceedings of the German-Polish Workshop on Computational Biology, Scheduling and Machine Learning (ICOLE'2011)*, volume 1 of *Machine Learning Reports*, pages 70–79, 2012. ISSN:1865-3960 http://www.techfak.uni-bielefeld.de/f̃schleif/mlr/mlr_01_2012.pdf.

[16] M. Mueller. *Information retrieval for music and motion*. Springer, 2007.

[17] F. Ofli, E. Erzin, Y. Yemez, and A. Tekalp. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia*, 14(3 PART 2):747–759, 2012.

[18] M. Rastegari, M. Rouhani, N. Gheissari, and M. Pedram. Cartoon motion capturing and retargeting by rigid shape manipulation. In *Digital Image Computing: Techniques and Applications, 2009. DICTA '09.*, pages 498–504, 2009.

[19] F.-M. Schleif, B. Mokbel, A. Gisbrecht, L. Theunissen, V. Dürr, and B. Hammer. Learning relevant time points for time-series data in the life sciences. In *Proceedings of ICANN 2012*, pages 531–539, 2012.

[20] P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[21] E. Stone and M. Skubic. Evaluation of an inexpensive depth camera for in-home gait assessment. *Journal of Ambient Intelligence and Smart Environments*, 3(4):349–361, 2011.

[22] L. F. Teixeira and L. Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 30(2):157–167, 2009.

[23] L. van der Maaten. Learning discriminative fisher kernels. In Getoor and Scheffer [10], pages 217–224.

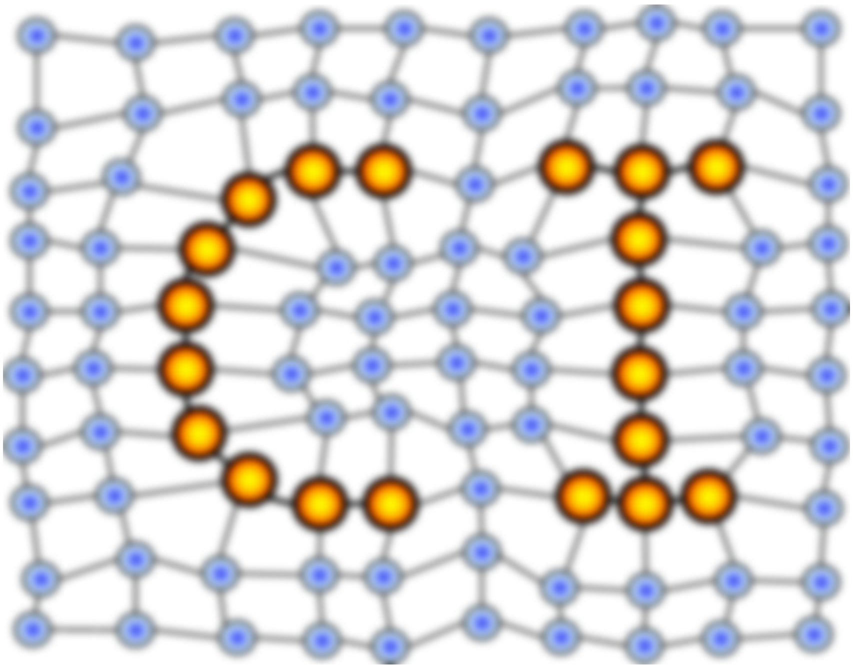
[24] V. Vapnik. *The nature of statistical learning theory*. Statistics for engineering and information science. Springer, 2000.

[25] Y. Wu, J. Lin, and T. Huang. Analyzing and capturing ariculated hand motion in image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1910–1922, 2005.

[26] J. Xu, X. Li, Y. Ren, and W. Geng. Performance-driven animation of hand-drawn cartoon faces. *Computer Animation and Virtual Worlds*, 22(5):471–483, 2011.

MACHINE LEARNING REPORTS

Report 05/2013



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.