

# System and Analysis used for a Dynamic Facial Speech Deformation Model

Jürgen Rurainsky

Fraunhofer Heinrich-Hertz-Institute, Einsteinufer 37, 10587 Berlin, Germany  
[rurainsky@hhi.fraunhofer.de](mailto:rurainsky@hhi.fraunhofer.de)  
<http://www.hhi.fraunhofer.de>

**Abstract.** While facial expressions and phoneme states are analyzed and published very well, the dynamic deformation of a face is rarely described or modeled. Recently dynamic facial expressions are analyzed. We describe a capture system, processing steps and analysis results useful for modeling facial deformations while speaking. The capture system consists of a double mirror construction and a high speed camera, in order to get fluid motion. Not only major face features as well as a high accuracy of the tracked facial points, are required for such analysis. The dynamic analysis results demonstrate the potential of a reduced phoneme alphabet, because of similar 3D shape deformations. The separation of asymmetric facial motion allows to setup a personalized deformation model, besides the common symmetric deformation.

**Key words:** motion, facial deformation, personalized

## 1 Introduction

Different capture systems for the analysis of facial motions have been presented in recent years. Although single or multi camera approaches have been addressed with different configurations, mirrors are rarely used. One reason for this could be the resolution of the capture unit, which is shared with all virtual views.

Many different techniques for the classification of facial expressions in still images has been published. Recently also the methods for dynamic analysis of facial expressions pushing forward and described by Cohen *et al.*[1], Hu *et al.*[2] and Zhang *et al.*[3]. The dynamic deformation produced while speaking has not been analyzed so far.

Since we target for the analysis of the dynamic behavior of facial motion, the sampling rate, in which the motion states are recorded, is an important issue. Important transitions from one state to another maybe get lost if only a video frame rate of 25 fps is used and these details are not available for the natural animation of 3D models.

We present our capture system based on a high speed camera and two surface mirror. These three views are used to reconstruct a 3D model sequence of a talking person's face. This model sequence is used to compensate the rigid motion and analyzing the facial deformation to a reference model. The analysis of the dynamical performance of face while speaking phonemes is described as well.

## 2 Capture Device

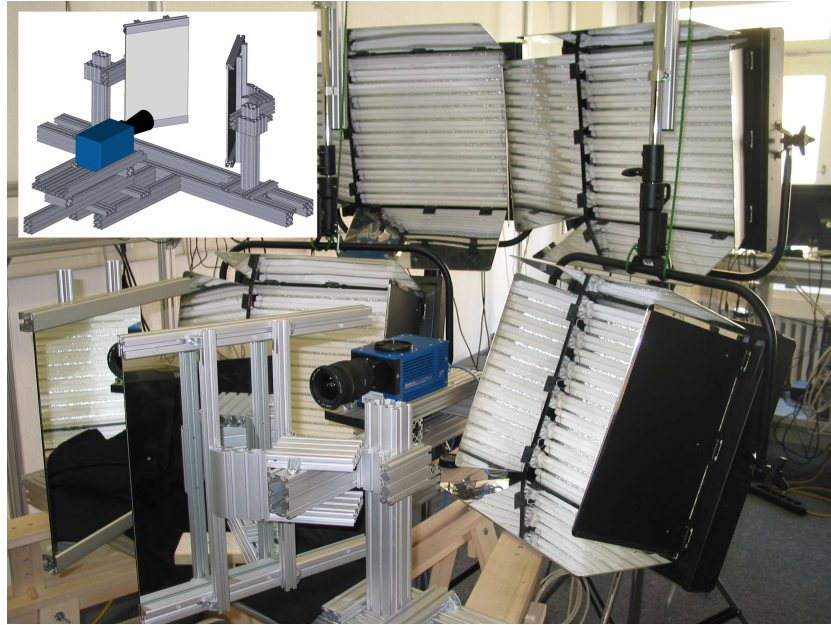


Fig. 1: High speed camera and mirror construction with two surface mirrors. Top Left Corner: Schematic scheme of the construction.

We have constructed a system with two mirrors which is shown in **Fig. 1**. Mirrors are widely used within capture devices and described in several publications [4–6], but no high speed camera was used so far as capture unit and not always a 3D reconstruction was placed in the processing steps. The four flat lights are used to illuminate the scene uniformly.

By using mirrors additional points have to be considered. The surface of the used mirror is one important parameter, because these surface properties will be incorporated into the view calibration parameter. Other parameters are the reflection properties in the sense of color, magnification and multiple reflections. Multiple reflections appear because of the reflection on the glass and coating boundary. It is very common to use not surface mirrors based on a glass body, but to use polished metal mirrors instead. The stiffness of such metal mirror is much lower than a glass body based mirror. Therefore, an additional fixation has to be considered in case of metal based mirrors. We have used surface mirrors with a glass body.

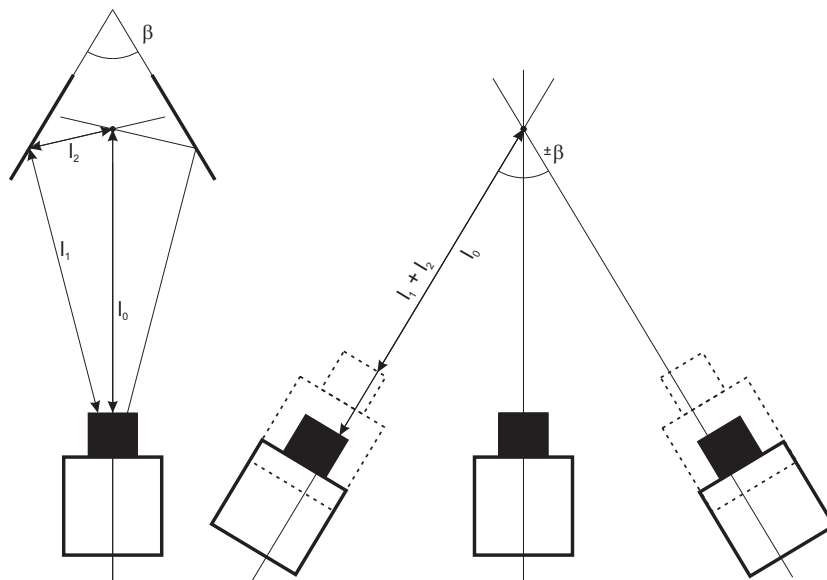


Fig. 2: (left) Capture device with two mirrors and one camera. (right) Setup interpretation used for calibration and 3D reconstruction.

## 2.1 Calibration

In order to get full benefit out of such system, the system has to be calibrated. Taking one point in 3D world as reference and measuring the light ray distance from this point to each view (two mirror views and one direct view), leads us to a system with two virtual cameras and one real camera. **Fig. 2** shows this representation, where the same point viewed by the virtual cameras appears to be more far away than in reality. The distance between several 3D world points should be therefore also closer, but there are not. They have the same distance as seen by the real camera. The light ray distance can not be neglected and therefore the virtual cameras has to be adjusted, in order to magnify these views. Therefore, each view (camera) has is own position in 3D world.

Using point correspondences captured from a calibration cube and a non-linear equation solver gives as the required intrinsic and extrinsic camera parameters. The back projection of the manual selected corner points used for the calibration can be used to determine the calibration error, which is below  $0.5\text{pixel}$  and therefore sufficient for the following analysis. The accuracy of the calibration also allows the usage for depth map determination approaches, like described in [7] *et al.* for a high density of feature points.

### 3 Principal Analysis

The performance of a face while speaking a specific set of words was captured with 200 fps using blue tape markers and for each frame a 3D shape was reconstructed by triangulation of the tracked marker points. The total amount of 43 words has been captured with an average duration of 217 frames. Each word represents a specific British and American phoneme [8].

#### 3.1 Motion Model

For the analysis of facial motion, we separate rigid body motion from deformations using a 3D sequence of facial points. The 3D model sequence is generated by triangulation of markers, which are placed on a human face and captured by a double mirror construction and a high speed camera. Rotation and translation for all axes (6 DOFs) of the associated 3D model describe the rigid-body motion and all other changes are regarded as deformation and noise.

## 4 Dynamic Comparison of Spoken Phonemes

While facial expressions and phoneme states are analyzed and published very well, the dynamic deformation of a face is rarely described or modeled. The amount of sampling points and the sampling frequency are the interesting measurements for such analysis and therefore define the value of an appropriate deformation model.

#### 4.1 Dynamic Time Warping (DTW)

Is a very known method for the comparison and motion model description used for dynamic data. Early works of Rabiner *et al.*[9] as well as Sakoe and Chiba [10] use DTW for the comparison of audio signals of spoken words. Gestures comparison as well as recognition are well described in representative publications of Corradini [11] or Li and Greenspan [12]. The recognition of hand writing using DTW is described by Niels [13]. There are many more publications dealing with DTW in various scenarios.

The main idea behind Dynamic Time Warping (DTW) is to map two sampling sets independently from the sampling rate as well as sampling period. Ratanamahatana and Keogh analyzed the behavior of DTW with different preprocessing steps and constraints [14]. One suggestion is not equalize different datasets before mapping, which was used for our analysis.

We use DTW to analyze dynamic 3D shape deformation. Actually, we just use the weights for a specific set of eigenshapes in contrast to Angeles-Yreta and Figueroa-Nazuno, who describe a measurement method for similarities of 3D objects in [15] by using distances within the 3D shapes. The eigenshapes are the same for all phonemes and only the frame based weights are describing the

difference from one phoneme data set to another phoneme data set. With other words, this could be used to compare data sets from different speakers while the sampling time do not have to be the same through all data sets. In addition facial deformations for phonemes can be correlated, in order to find an appropriate subset and therefore to reduce the amount of phonemes for the same performed deformation.

The **Eqn. 1** shows the reconstruction of a specific shape  $F$  for a defined range of eigenshapes  $R$  after adding the average  $\bar{\mathbf{B}}$  as well as the offset  $\mathbf{A}^{(3N,0)}$ .

$$\mathbf{A}_{recon.}^{(3N,F)} = \mathbf{A}^{(3N,0)} + \bar{\mathbf{B}} + \sum_{i=0}^R w_{i,F} \cdot \mathbf{Eig}_i \quad (1)$$

While the main method for DTW is well described, the used type of mapping function is based on a best performance result and therefore no specific formulation is defined. In the book of Rabiner and Juang [16] 7+1 types of path specifications are described, but mostly type I is used and shown in **Fig. 3**. Besides

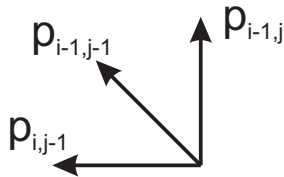


Fig. 3: Type I mapping function described by Rabiner and Juang [16].

the type of path specification the incorporation of multidimensional data can be used to change the performance of the mapping. Holt *et al.*[17] extending the DTW algorithm to the multidimensional approach MD-DTW, where different dimensions could be connected by a simple Euclidean distance for instance. For the analysis of dynamic facial deformation based on weights, the weights of several eigenshapes are included to the distance matrix by calculating the Euclidean distance. The distance matrix of two different datasets using 8 eigenshapes and therefore also 8 weights as well as path type I is shown in **Fig. 4**. The determined path is visualized with the white colored line. The mapping of these two different datasets (different length) is used to specify the correlation of the words *away* and *arm*, which represent two different visual dynamic deformations of a face during the pronunciation of these phonemes. **Fig. 5** and **Fig. 6** show the mapping results by incorporating the weights for one and eight eigenshapes to the distance matrix. Around frame 150 are difference between these both mappings can be seen, which also lead to different results. Both data sets have different length, but the mapping shows useful results, which supports the suggestion of Ratanamahatana and Keogh [14].

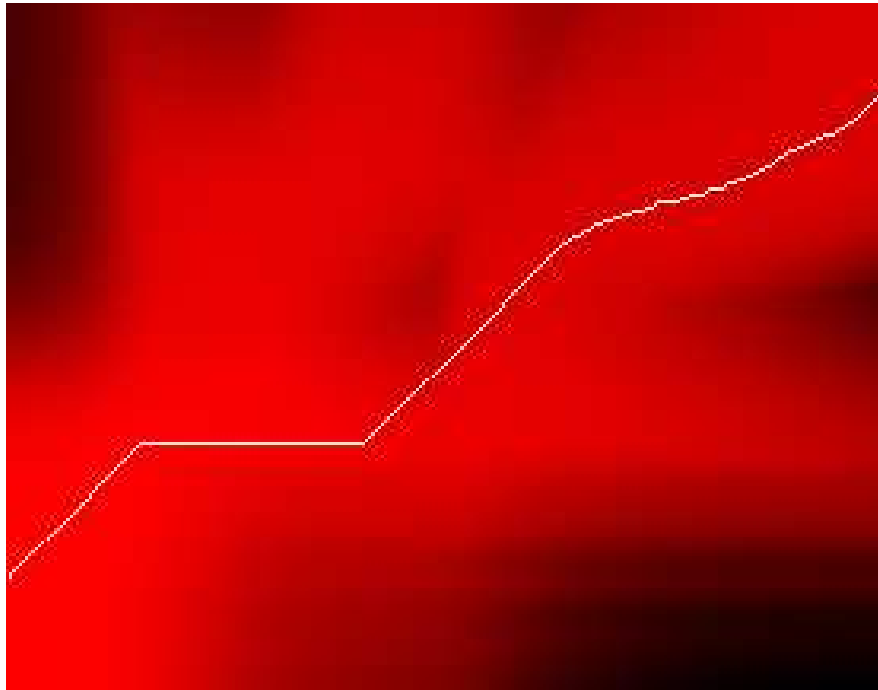


Fig. 4: Distance matrix for the DTW mapping of two phonemes represented by the words *away* and *arm*. The white colored line shows the mapping path with smallest energy. The weights for the first eight eigenshapes are incorporated into the distance matrix.

## 4.2 Experimental Results

The idea is to analyze the dynamic behavior of facial deformation while speaking. In order to find the right representation or subset of dynamic shape deformations, the smallest elements have to be compared. Phonemes already used to model static facial deformation and therefore we compare the dynamic shape of a phoneme alphabet. Each phoneme was compared to all other phonemes and the result will be a distance matrix showing the Euclidean distances after the DTW mapping of the to be compared data sets. We compared 52 data sets including phonemes and facial deformations, which are done by unattended motions, like getting the lips wet. We incorporated only the first eight eigenshapes for the common facial deformation representation and left the higher order eigenshapes for asymmetric and therefore personalized deformation out. **Fig. 7** shows the result of this experiment, which leads to the awareness that the recorded phoneme examples show a high correlation and therefore this data set can be reduced.

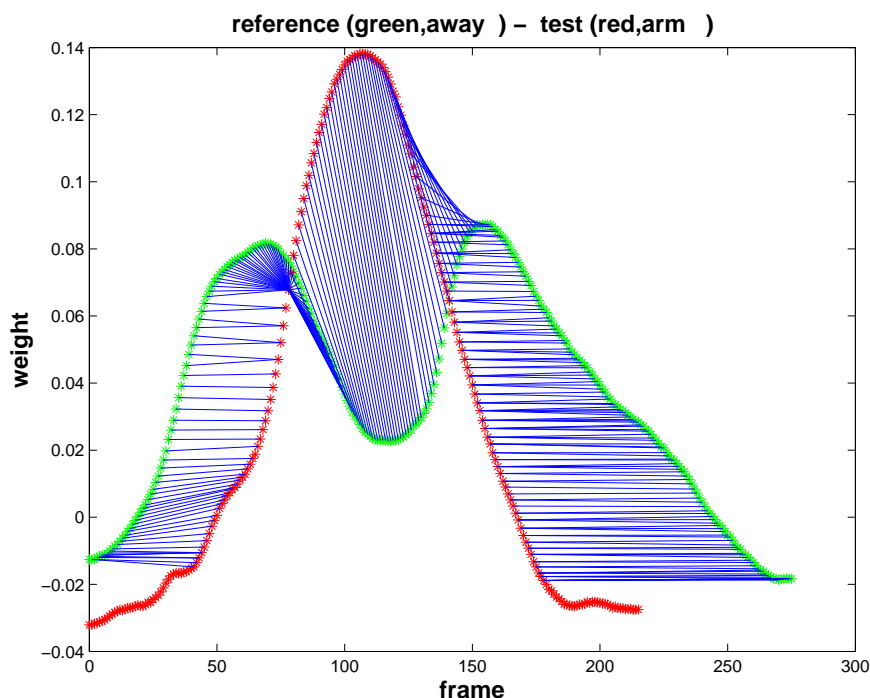


Fig. 5: Mapping between the frame based weights with the first eigenshape used to represent the words *away* and *arm*. Only the weights for the first eigenshape was incorporated into the distance matrix.

## 5 Conclusion

We have shown and described a system as well as analysis methods and results for dynamic facial deformations, which can be observed during the pronunciation of words. The system allows to capture the dynamic shape motions with up to 200 fps and the double mirrors provide us with the desired 3D shapes. Extracting the rigid motion and the eigenshape representation of these observed deformations are described as well. Dynamic Time warping (DTW) is used to compare different data sets, where multidimensional data in the form of eigenshape weights are incorporated into the data set mapping. The direct comparison of a set of 52 phoneme and unattended motions is provided in the form of a distance matrix. This matrix allows the assumption, that further reduction can be applied without losing major deformations.

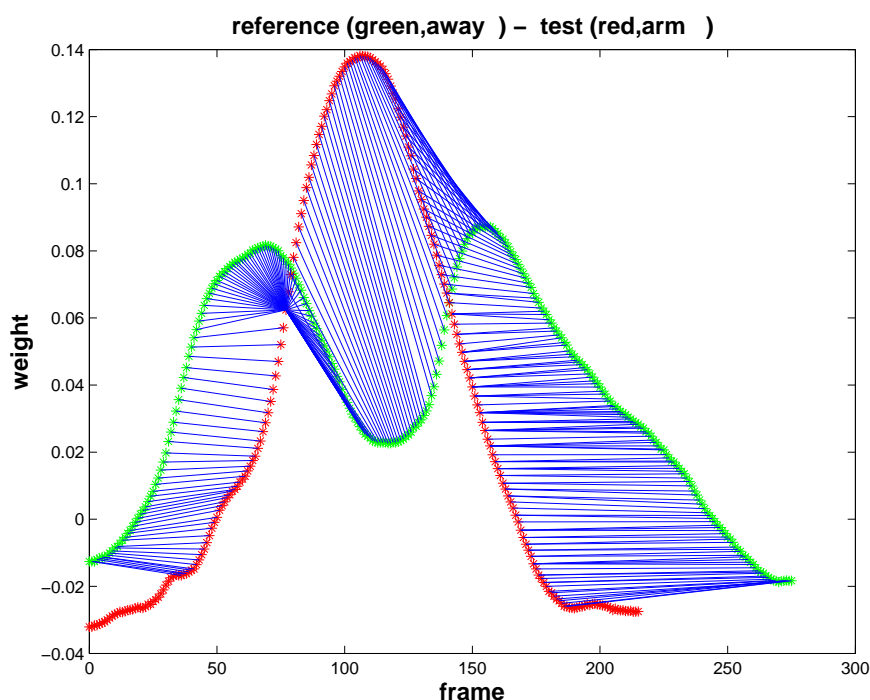


Fig. 6: Mapping between the frame based weights with the first eigenshape used to represent the words *away* and *arm*. The weights for the first eight eigenshapes were incorporated into the distance matrix.

## References

1. Cohen, I., Sebe, N., Garg, A., Lew, M.S., Huang, T.S.: Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. In: Computer Vision and Image Understanding. (September 2003) 160–187
2. Hu, C., Chang, Y., Feris, R., Turk, M.: Manifold based analysis of facial expression. In: CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5, Washington, DC, USA, IEEE Computer Society (2004) 81
3. Zhang, Y., Ji, Q.: Facial expression understanding in image sequences using dynamic and active visual information fusion. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 1297
4. Basu, S.: A Three-Dimensional Model of Human Lip Motions. Master's thesis, Massachusetts Institute of Technology Department of EECS, Cambridge, MA, USA (February 1997)
5. Odisio, M., Bailly, G., Eliseia, F.: Tracking talking faces with shape and appearance models. *Journal of Speech Communication* **44** (October 2004) 63–82

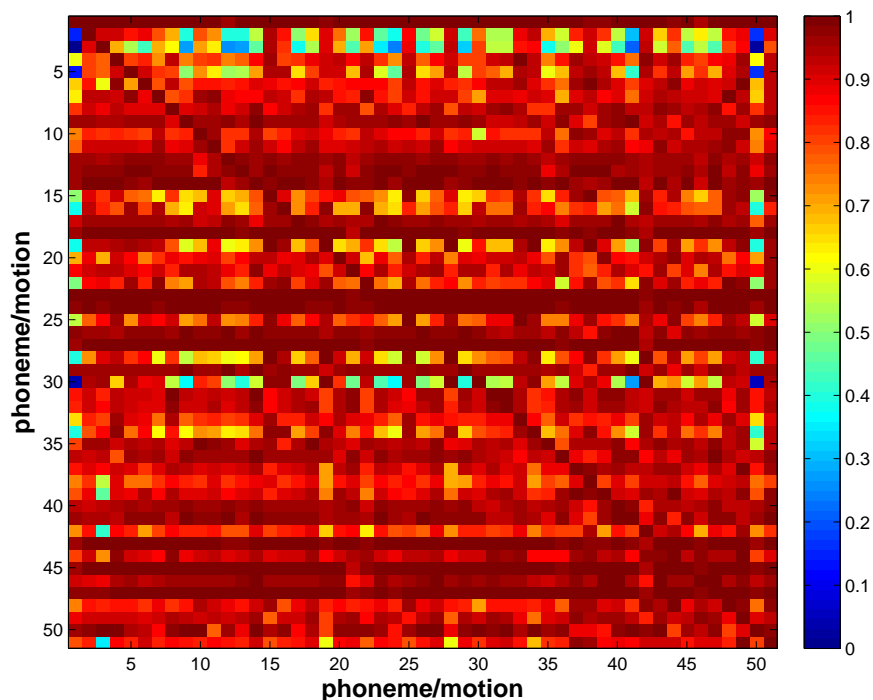


Fig. 7: Cross comparison of data sets including the dynamic facial deformation of phonemes and unattended motions. The dialog line shows the highest comparison result, because of the comparison of identical data sets.

6. Badin, P., Bailly, G., Revéret, L., Baciú, M., Segebarth, C., Savariaux, C.: Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics* **30**(3) (2002)
7. Dainese, G., Marcon, M., Sarti, A., Tubaro, S.: Accurate Depth-map estimation for 3D face modelling. In: 13th European Signal Processing Conference. (September 2005) 1883–1886
8. : International Phonetic Alphabet (IPA) for British and American phonemes. World Wide Web electronic publication
9. Rabiner, L., Rosenberg, A., Levinson, S.: Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Transaction on Acoustics, Speech and Signal Processing* **26** (December 1978) 575–582
10. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transaction on Acoustics, Speech and Signal Processing* **26** (February 1978) 43–49
11. Corradini, A.: Dynamic Time Warping for Off-Line Recognition of a Small Gesture Vocabulary. In: IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01), Washington, DC, USA, IEEE Computer Society (2001) 82

12. Li, H., Greenspan, M.: Segmentation and Recognition of Continuous Gestures. In: Proc. of the IEEE International Conference on Image Processing ICIP 2007, San Antonio, TX, USA (September 2007) 365–368
13. Niels, R.: Dynamic Time Warping: An Intuitive Way of Handwriting Recognition? Master's thesis, Radboud University Nijmegen, Nijmegen, The Netherlands (2004)
14. Ratanamahatana, C.A., Keogh, E.: Three Myths about Dynamic Time Warping. In: Proc. of the SIAM International Conference on Data Mining (SDM '05), Newport Beach, CA, USA (April 2005) 506–510
15. Angeles-Yreta, A., Figueroa-Nazuno, J.: Computing Similarity Among 3D Objects Using Dynamic Time Warping. Springer Berlin / Heidelberg (2005)
16. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1993)
17. ten Holt, G., Reinders, M., Hendriks, E.: Multi-dimensional dynamic time warping for gesture recognition. In: Proc. of the 13th conference of the Advanced School of Computing and Imaging, Delft, The Netherlands (2007) 158–165