Correlation analysis of facial features and sign gestures

Zdeněk Krňoul, Marek Hrúz, Pavel Campr
Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Plzeň, Czech Republic
Email: {zdkrnoul, mhruz, campr}@kky.zcu.cz

***Abstract*—In this paper we focus on the potential correlation of the manual and the non-manual component of sign language. This information is useful for sign language analysis, recognition and synthesis. We are mainly concerned with the application for sign synthesis. First we extracted features that represent the manual and non-manual component. We present a simple but robust method for the hand tracking to obtain a 2D trajectory representing a portion of the manual component. The head is tracked via Active Appearance Model. We introduce initial experiments to reveal the relationship between these features. The procedure is verified on the corpus of isolated signs from Czech Sign Language. The results imply that the components of sign language are correlated. The most correlated signals are the vertical movement of head and hands.**

I. Introduction

Sign language (SL) is the main communication form for hearing impaired people. To enable the communication between deaf and hearing people systems of SL recognition and synthesis are developed [1], [2], [3]. SL is composed of manual and non-manual component [4]. The manual component is represented by the handshape, palm orientation and the arm movement. The non-manual component consists of face expression, mouth movement and pose. In SL recognition they are combined to achieve better recognition rates [1]. In SL synthesis an avatar performs the signs by synthesizing both components [3]. In this paper we analyze the correlation between the components of SL for the purpose of SL synthesis.

SL synthesis adopts methods of machine translation, 3D modeling and animation. The idea is to render animation of signs representing the translated text (or possibly speech). Current systems create the animation of mouth movements which correspond to the phonemes of the spoken language [3], [5]. It is known that native signers do not produce the mouthing in such way. To control the animation we use a notation system based on hamnosys. This includes the non-manual component [6] which has not been addressed much before. However, the notation does not take into account the interaction between the different components of SL. In this paper we analyse the correlations of both SL components. The results will be used to model the dependencies of the different SL components to get the animation more authentic and more intelligible.

The outline of this work is as follows. Section 2 describes SL corpus used for the experiments, Section 3 summarizes techniques used for feature extraction for both the manual and non-manual component. Initial experiments with the features are described in Section 4. Section 5 concludes our work.

II. Data

For consequent experiments we use data from UWB-07-SLR-P sign language corpus[1] [7]. The corpus was recorded in laboratory conditions (black background, uniform and static illumination, static cameras) and allows easier usage of computer vision methods than other corpora recorded primarily for linguistic needs.

Fig. 1. Sample frame from the corpus: front, face and top perspective

It contains video data of four signers recorded from three different perspectives (see fig. 1). Each signer performed 378 signs from Czech SL with five and more repetitions. The corpus consists of several types of signs: numbers (35 signs), one and two-handed finger alphabet (64), town names (35) and other signs (244). Further detail can be found in [7].

III. Methods of Feature Extraction

In the following sections we will describe the methods used to extract features representing the components of SL. First we present an object detection based tracking suitable for precise representation of the manual component. To represent the non-manual component we use the shape parameters of Active Appearance Model (AAM). The appearance parameters of the model are used to robustify the tracking.

A. Tracking of Hands

The tracking process is based on object detection and consecutive tracking of scalar description of the objects. The process is highly dependent on image segmentation and should not be used in cluttered environment. In our experiments we use skin-color segmentation. Samples of manually selected skin-colors are processed by Expectation Maximization (EM)

[1]available at European Language Resources Association (ELRA), www.elra.info

978-1-4244-5900-1/10/$26.00 ©2010 IEEE

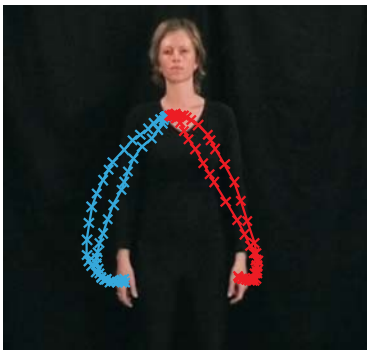

Fig. 2. Example of detected hand trajectory.

algorithm to train a Gaussian Mixture Model (GMM). The mixture is thresholded so that low probabilities are zeroed and a look-up table is created so that for every RGB value there is a probability of belonging to a skin segment. This allows us to segment an image very quickly. Next, we find objects (connected components) in the segmented image. In the first frame we initialize the trackers using the knowledge of approximate starting locations of hands and head. Each tracker tracks one object. We create a scalar description of the objects consisting of the position, velocity, perimeter of the contour, area of the bounding box, area of the object and seven Hu moments. Furthermore, we store the image of the object for template matching. In consecutive frames we detect a new set of objects and describe them with the above mentioned features. Each tracker then computes distances of the tracked features and the new features. Template matching is used to compute the distance of the appearance of the objects. According to a trained probability model (GMM) of the distances the tracker decides which of the new objects is the tracked one. The tracker has special models for occlusions and for the first frame after occlusion. The occlusion can be detected by the tracker since the area of the bounding box of occluded objects increases rapidly when they touch. Since the tracking process is more complex and to fully describe it here would be extensive we kindly refer the reader to [2]. The tracking process provides us with many features. For our experiments we chose the position of hands as the main feature describing the manual component of SL. An example of the tracking result can be seen in figure 2.

B. Tracking of Head

We consider face features corresponding to position and shape of face. However for more robust processing we propose to use the multi-resolution combined AAM [8]. The AAM essentially combines two types of linear models. One for the shape and one for the appearance, but we can extract the position and shape of face.

The shape of AAM is given by image points $s = (x_1, y_1 ... x_N, y_N)$ defining contours around face, nose, mouth, and eyes. We experimentally designed a set of 53 image points including 19 points for both inner and outer lip contour,

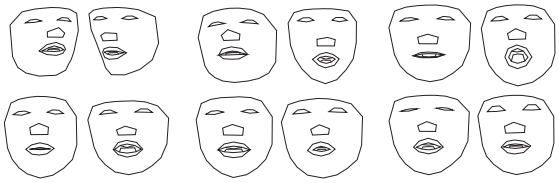

Fig. 3. Shapes of the AAM for $\pm 2.5SD$, where SD is standard deviation of the corresponding parameter, upper line $F1..F3$ parameters and lower line $F4..F6$.

upper teeth and tongue (see figure 4). In accordance with the Principal Components Analysis (PCA) the shape of AAM is expressed as the basic shape s_0 plus a linear combination of N shape vectors s_i:

$$s = s_0 + \sum_{i=1}^{N} Fis_i, \tag{1}$$

where shape parameters Fi describe important variations of the face shape. We have selected the first six principal components as the base of the shape space ($N = 6$). Fi, $i = 1..6$ parameters are summarized in table I.

We have manually identified all considered points in 61 training images. We have experimentally chosen subset of 46 training images in order to get such parameters that are suitable for the planned experiments. Illustration of the shape AAM parameters is shown in figure 3. To allow the searching, we have to enter additional pose parameters describing the affine transformation of the shape in the image plane, see table I.

TABLE I
The parameters expressing the non-manual component.

Parameter	*Meaning*	*Group*
FR	face rotation around optic axis	face pose
FX	face horizontal displacement	face pose
FY	face vertical displacement	face pose
$F1$	face looking right-left	face pose
$F2$	face looking up-down	face pose
$F3$	mouth opening 1	mouthing
$F4$	mouth opening 2	mouthing
$F5$	mouth rounding (protrusion)	mouthing
$F6$	eyes open-close	upper face

Next the shape of AAM is processed by Delaunay triangulation to get pixels covering the face (an appearance). The appearance of AAM is standardly defined in the base mesh s_0. Each training image is backward warped by to the base shape s_0. $W(x, F)$ is a piecewise affine transformation defined from the corresponding triangles and directly determines coordinates of corresponding pixels in the training image. The appearance of AAM is an image $A(x)$ defined as the intensity of pixels at these positions. The appearance A_0 and A_i are the eigenvectors obtained by PCA on a set of warped training images.

$$A(x) = A_0 + \sum_{i=1}^{N} \lambda_i A_i \tag{2}$$

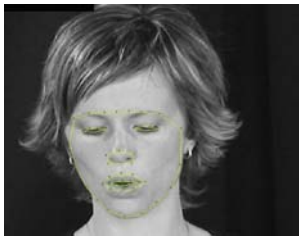

Fig. 4. Final fitting of a face shape in the input frame.

where λ_i are appearance parameters. Finally combined AAM operates with a single set of parameters $c = (c_1, c_2, ..., c_L)$ to search the best fit of the AAM in a input video frame. Vector c is obtained by another PCA collected from the appropriately weighted parameters Fi and λ_i. For searching a scale and three positional parameters (FX,FY and FR) of the AAM shape in image plane have to be considered as well. Standard gradient descent optimization algorithm is used to minimize the sum of squares of the difference between intensity of input image $I(W(x;F))$ and the appearance $A(x)$. The illustration of fitting shape is in figure 4.

Face detection has to precede processing of frames by the AAM. We employed a well-known face detector [9]. The most likely area showing the speaker's face is detected from one or more initial frames of the processed video record. Tracking of head is implemented by OpenCV library, AAM-library (C++ implementation by Yao Wei) and the scripts for batch processing.

IV. Relationship between Manual and Non-manual Component

The aim of the experiments is an initial scope of the relationship between the non-manual component and the manual component. We have selected 200 entries from the corpus that capture 44 signs interpreted by one deaf speaker. The entries repeatedly captured each sign 1–7 times. We used the method described in section III-A to obtain the trajectories of hand movements. Furthermore we obtain trajectories of the movement of the face by the method described in section III-B. Face parameters are divided into two groups: face pose and mouthing, see Table I. The hand and head features are concatenated yielding a vector $T(t)$. When a correlation is computed note that we have computed Pearson's linear correlation coefficient.

A. Experiment 1

The experiment attempts to compare the non-manual and manual component globally for all the observed data. The first step of the experiment is to calculate both cross and auto-correlation of the components. For this purpose the trajectories are transformed using the relationship:

$$D_j(t) = T_j(t) - T_j(t-1). \tag{3}$$

Relationship 3 indicates the time difference of the $j-th$ parameter. The signal $D(t)$ is filtered with 100 ms window to reduce the noise level. Furthermore, we define the energy:

$$E_P(t) = \sum_{j \in P} D_j(t)^2 \tag{4}$$

where P is a subset of vector T. We repeatedly used the relationship 4 to calculate three energies. For the manual component the energy $EDH(t)$ is derived from the positional parameters of the dominant hand only. The first energy of the non-manual component parameters $EFP(t)$ is collected from positional parameters $P \in \{FR, FX, FY, F1, F2\}$. Energy $EFM(t)$ is computed for articulatory parameters $P \in \{F3, F5, F5\}$. Parameter F6 (eyes closing/opening) is not considered. Furthermore we determined the cross-correlation and auto-correlation coefficients of the energy of the non-manual component compared to the energy of the manual component shifted by 2 seconds forward and backward. We note that the average duration of one sign is 1.8 sec. The resulting relationship for all considered combinations is shown in the graphs in figure 5.

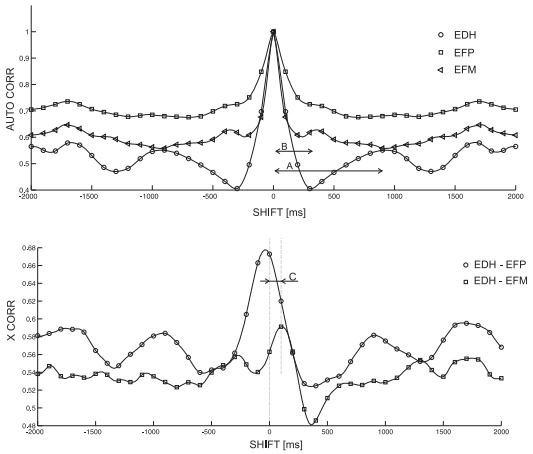

Fig. 5. The upper figure depicts auto-correlation separately for energy dominant hand: EDH, face pose: EFP and mouthing: EFM, the lower figure shows cross-correlation between energy of manual and non-manual component only.

The auto-correlation of energy of the manual component indicates that the positional parameters of the dominant hand have a frequency with period of $A = 800 - 900msec$. In contrast, approximately 2.5 times smaller ($B \cong 360msec$) frequency is observed on the auto-correlation of energy for mouthing. This suggests that the signals will not be globally correlated and each sign should be processed separately.

B. Experiment 2

In this experiment we compute the cross-correlation between individual features of manual and non-manual component for individual signs. A peak in the cross-correlation signal indicates that the two signals are well correlated when one signal is shifted to this position (figure 5). We find the maximal absolute value in each cross-correlation signal to determine the shift. The values of the shift were centered

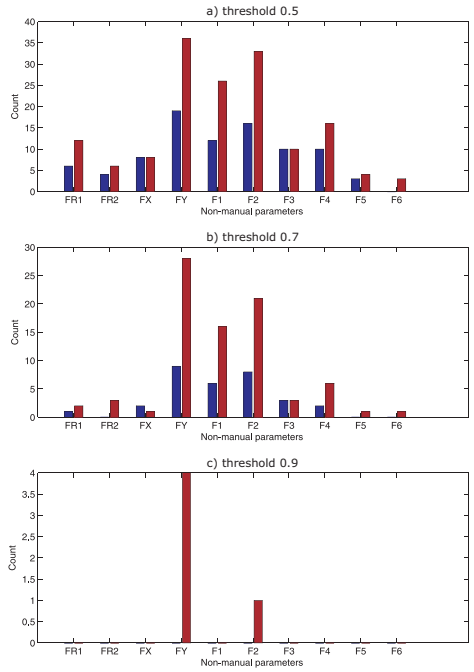

Fig. 6. Histograms of best correlations for the threshold a) 0.5 b) 0.7 c) 0.9. First bar is X position of the dominant hand and the second bar is the Y position of the dominant hand.

around zero with deviation smaller than $20msec$. This initial part of the experiment rejects the assumption that the signals of the manual and non-manual components are shifted (shift C in the figure 5 is not significant).

In the next part of the experiment we compute correlations between the zero-shifted signals of manual and non-manual component. The signals of manual component were the x and y positions of the dominant hand and the signals of the non-manual component were all parameters in table I. The parameter FR is decomposed into two components corresponding to the cosine and sine of the in plain rotation (FR1, FR2). For each recording we determined 20 correlations. We clustered the correlations according the signs and computed an average correlation for each sign. In the end there were 44 times 20 correlations (44 signs, 20 combinations of manual/non-manual signals). For each sign we selected 10 best correlations that were above a given threshold. We selected three types of thresholds. Threshold "better than chance" (0.5), correlated signals (0.7) and highly correlated signals. (0.9).

We created histograms from the 10 best correlations for the different thresholds. The results can be seen in figure 6.

It can be seen from the results that only few signs from observed data have very correlated non-manual and manual signals. This is due to relatively small shifts and different frequency of the signals. When we observe the histogram with threshold 0.5 we can see that the greatest dependency is between the y movement of the head and the y movement of the dominant hand. The second best correlation is between the pitch of the head and the y position of hand. Also, we found dependencies between the articulation parameters and hand movement. When we increase the threshold we observe similar dependencies but in lower counts. Also it seems the non-manual component is more dependent on the y movement of the hand than the x movement. This is also given by the signs themselves where the movement in y direction is more frequent than in x direction.

V. Conclusion

We have extracted features from the SLR-P database in order to analyze the dependencies between the manual and non-manual component of sign language. The features were the 2D trajectory of both hands and nine features describing the non-manual component. The tracking is based on image segmentation and object detection. The non-manual component is tracked by the AAM. In the first experiment we tried to find global correlations of energies of dominant hand and face parameters. We have shown that the frequencies of the energies are different over the whole considered data. This required the individual approach. Thus, in the second experiment, we compute the correlations between the manual and non-manual parameters for each sign and show the resulting correlations. The results imply that most correlated signals are the vertical position of dominant hand and head. The mouthing is also correlated with the hand movement but it needs next experiments. These conclusions provide initial information on how to properly animate a signing avatar, so that we will respect the correlations.

Acknowledgment

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/P609,the Ministry of Education of the Czech Republic, project No. ME08106 and by the grant of The University of West Bohemia, project No. SGS-2010-054.

References

[1] Oya Aran, Thomas Burger, Alice Caplier and Lale Akarun "Sequential Belief-Based Fusion of Manual and Non-manual Information for Recognizing Isolated Signs", GW 2009, Series LNCS, 2009.

[2] Trmal Jan, Hrúz Marek et al., "Feature Space Transforms for Czech Sign-Language Recognition", In Proceedings of Interspeech 2008, p. 2036-2039, Causal Production Pty ltd., 2008.

[3] Zdeněk Krňoul, Jakub Kanis, Miloš Železný and Luděk Müller "Czech text-to-sign speech synthesizer", MLMI 2007, Series LNCS, 2008.

[4] Byron Bridges and Melanie Metzger, "Deaf Tend Your: Non-Manual Signals in ASL", Silver Spring, MD:Calliope Press, 1996.

[5] R. Elliott, J. R. W. Glauert and J. R. Kennaway, "A framework for non-manual gestures in a synthetic signing system", In CWUAAT, pages 127-136, 2004.

[6] Zdeněk Krňoul, "New features in synthesis of sign language addressing non-manual component", 4th Workshop on Representation and Processing of Sign Languages, ELRA, 2010.

[7] Pavel Campr, Marek Hrúz and Jana Trojanová, "Collection and Preprocessing of Czech Sign Language Corpus for Sign Language Recognition", Proceedings of LREC 2008, ELRA, Marrakech, Morocco, 2008.

[8] T.F. Cootes, G.J. Edwards and C.J. Taylor, "Active Appearance Models", In Proceedings of the European Conference on Computer Vision, volume 2, pages 484-498, 1998.

[9] Paul Viola and Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", in Computer Vision and Pattern Recognition, IEEE Computer Society Conference, 2001.