Towards Automatic Annotation of Sign Language Dictionary Corpora*

Marek Hrúz, Zdeněk Krňoul, Pavel Campr, and Luděk Müller

Department of Cybernetics
University of West Bohemia
306 14, Plzen, Czech Republic
{mhruz,zdkrnoul,campr,muller}@kky.zcu.cz

Abstract. This paper deals with novel automatic categorization of signs used in sign language dictionaries. The categorization provides additional information about lexical signs interpreted in the form of video files. We design a new method for automatic parameterization of these video files and categorization of the signs from extracted information. The method incorporates advanced image processing for detection and tracking of hands and head of signing character in the input image sequences. For tracking of hands we developed an algorithm based on object detection and discriminative probability models. For the tracking of head we use active appearance model. This method is a very powerful for detection and tracking of human face. We specify feasible conditions of the model enabling to use the extracted parameters for basic categorization of the non-manual component. We introduce an experiment with the automatic categorization determining symmetry, location and contact of hands, shape of mouth, close eyes and others. The result of experiment is primary the categorization of more than 200 signs and discussion of problems and next extension.

1 Introduction

Sign language (SL) is a communication form mainly used by deaf or hearing impaired people. In this language visually transmitted manual (MC) and non-manual (NMC) components are used to convey meaning. The MC consists of hand shape, palm orientation and the arm movement. The NMC component consists of face expression, body pose and lip movement. Because the majority language (usually the language used by the hearing) is the secondary language of the Deaf a bi-directional translation is highly important for better Deaf orientation in our day-to-day shared social environment. Currently, human interpreters provide this translation but their service can be expensive and not always available. Therefor systems of SL recognition and synthesis are being developed [1,2,3]. The results from these fields of research can be used in various ways.

* This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416 and project No. GAČR 102/09/P609, by 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.

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 331–339, 2011.

Our goal is to use the recognized features for an automatic categorization of signs for the use in a SL dictionary. The proposed categorization algorithm considers sign categories corresponding to the entries in the symbolic notation HamNoSys (HNS) and SignWriting (SW)[1]. Symbolic notations are used to describe the sign. Usually these notations are created manually which is very time consuming. This process is influenced by the skills and experience of the human annotators. On the other hand automatic categorization of video files is deterministic provided the same input parameters. Also it fastens the work of human annotators who will only need to correct the mistakes of the automatic annotation. This annotation allows to search among the signs and enables the translation from SL to spoken language.

2 Related Work

MC recognition is closely related to tracking. There are many methods that vary depending on the scenario. Sometimes markers or color cues are used to help the process. A good survey can be found in [4]. We also refer to [5]. Our approach is based on color segmentation and object detection and description. Similar approach can be found in [6]. In our work we experiment with linear dimension reduction methods to obtain better tracking results.

For NMC signal, there are generative parametric models commonly used to track and synthesize faces in images and video sequences. We can distinguish two types of automatic face tracking algorithms. The first type is feature-based, matching the local interest points between subsequent frames, such as a 3D pose tracker [7] and 3D deformable face tracking [8]. The second type is appearance-based, using generative linear models of face appearance. There are Active Appearance Models (AAM) [9] and 3D Morphable Models [10].

AAM is a combined model of shape and texture. It ensures precise alignment, is very powerful and efficient to describe the movements in the face. The original proposal is used for identification of different faces as well as tracking [9]. Further improvements of AAM for local inter-frame appearance constraint optimization are integrated [11]. There are 3D AAM including 3D models to cover non-linear changes in the observed data. In most cases, the condition is pre-aligned data points arranged in the training images.

3 Data

Data for our experiment are selected signs from the on-line dictionary [12]. The dataset consists of pairs of synchronized video files capturing one speaker from two different views in the same lighting conditions. The recordings of the first and second view are RGB color images in HD resolution, 25 frames per second with high-quality compression. The first view captures the entire body of the speaker and the second one is a detail of the face, see Fig. 1. Audio track is not included. Totally 213 signs (video files) are considered.

[1] www.sign-lang.uni-hamburg.de/projects/hamnosys.html, www.signwriting.org/

Fig. 1. Example of considered data. Left - manual component, right - non-manual component.

4 Hand Tracking

The hand tracking is based on skin color segmentation and object description. Because of the nature of our data we can assume constant lighting and environment conditions. This makes the problem of tracking much easier but one has to still account for the occlusions and self-occlusions occurring in SL (for example see [13,14]).

4.1 Skin Color Segmentation

Because we are working with data of a SL dictionary we can assume that there will be not many performers and the characteristics of the video data will be constant or at least piecewise constants. That is why we use a constant skin color model in a form of a look-up-table. There have been a lot of papers published in this field. The approaches usually differ in the color models used for skin color representation and a parametric or non-parametric description of the model. We work with the native RGB color space and a hybrid model which in the end yields a non-parametric model in the form of a look-up-table. We train a Gaussian Mixture Model (GMM) from examples of skin color manually selected from our database. We threshold and scale the probability of the model to obtain a $256 \times 256 \times 256$ look-up-table with values from 0 to 255. We were inspired by the work [15] which we refer to for further details.

4.2 Tracking

In the scenario of SL movements tracking there are several objects of interest. The head that is usually static but changes a lot in the appearance. The hands that move rapidly, change the shape and orientation to the camera. We assume that the changes of the appearance occur slowly relatively to the camera frame rate. This should enable us to track the objects in a discriminative manner.

We define a tracker for each object we want to track. In our case there are three trackers. The tracker contains several discriminative models for object tracking. The number of models depends on the number of events we want to take into account. In our case there are 4 models. Model of non-occluded object, model of the change of state from non-occluded to occluded, model of occlusion and model of the change of state from occluded to non-occluded. This is due to big changes in the appearance of the

objects when they travel from one state to the other. Each model is a 4 mixture GMM in a 5D space that is defined by the properties of the objects. In the first frame the trackers are initialized by the object lying in a predefined region based on the knowledge of the starting pose of a performer. In the next frame a new set of objects is detected. Each tracker compares every object with the identified object from the last frame via the discriminative probability model. The input vector for the model is a 5D vector of relative differences. Then the probability of this vector is the probability of the unknown object to be the tracked one. This probability will be noted as $p_m(t_i|o_j)$, where m is a specific model, t_i is the i^{th} tracker and o_j is the j^{th} object.

Object Description. In the last paragraph we mentioned a 5D vector of relative differences that is used for object comparison. The vector is obtained as follows:

1. Use skin color segmentation to obtain all possible body parts.
2. Eliminate all segments that do not fulfill the defined conditions (size, width/ height ratio)
3. For each object compute - bounding box, Hu moments of the contour, area of the object and perimeter of the contour
4. From the information in point 3. compute for every tracker/object pair - normalized correlation between object image and tracked object image, normalized distance between their contours (computed from Hu moments), relative difference between their bounding box areas, relative difference between their perimeters, relative difference between their areas, relative difference between their velocity and location

This procedure yields a 7D vector for each tracker/object pair. Next, we want to find a transformation that reduces the correlation between the features and possibly reduces the dimensionality of the model. We experimented with Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA) and Heteroscedastic LDA (HLDA). These experiments are out of the scope of this paper but note that HLDA transformation into a 5D feature space resulted in the best performance.

Configuration Determination. Next, we want to determine which object belongs to which tracker. Let $\mathcal{O} = \{o_j\}, j = 1..N$ where N is the number of objects be a set of detected body parts. Let $\mathcal{T} = \{t_i\}, i = 1..3$ be a set of trackers. A configuration $\mathcal{C}$ is a mapping $\mathcal{T} \rightarrow \mathcal{O}$ that fully describes which tracker tracks which object. In SL scenario there exist 5 cases depicted in Figure 2 that describe the mutual relation between body parts. Each case is conditioned by the number of body parts detected. This enables us to hypothesize only about the plausible configurations. The algorithm for configuration determination is as follows:

1. Based on the number of detected body parts select a $C_k \in \mathcal{C}$ that fulfills the condition
2. Compute the log-likelihood of the selected configuration

$$L_k = \sum_{i=1}^{3} \log p_m(t_i|C_k(t_i)) \tag{1}$$

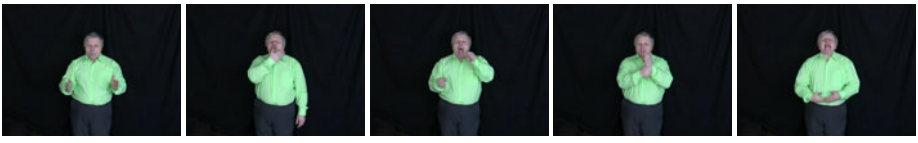

Fig. 2. Five possible cases of hand/head mutual relation. Note that several configurations may represent each case. This is due to the fact, that the case does not tell us which object is which.

3. If this is the maximum likelihood seen so far, store it as a new maximum
4. If there are no more configurations to test, select the C_k with maximum L_k as the recognized configuration, else go to point 1

In Eq. 1 we have to select a proper model m to evaluate the probability. The model is determined by the configuration C_k. Based on the last known configuration we are able to tell from which state the individual objects travel to the hypothesized state defined by C_k. The model can be different for each tracker i.

5 Head Tracking

The NMC integrates several face expressions, mouthing, 3D position of head. To ensure robust processing, we assume the multi-resolution combined active appearance model (AAM) [9]. The AAM traces the position and local shape of the face by a combination of two linear models for shape and texture.

The training set consists of 51 images selected from the dataset. The training set includes NMC of complex face gestures and head position. Naturally NMC involves 3 DOF for rotations (the x, y image axes and the z optical axis). Nevertheless, rotation around the image axes (pitch and yaw) has to be incorporated to the shape model because we consider 2D AAM. Rotation around the optical axis (roll) is described by the pose parameters and is not incorporated in the shape model. This is done by manual normalization of the training frames to get the outer lip and the eye corners horizontal. Thereafter the PCA produces the basic shape s_0 plus linear combination of N shape eigenvectors s_i:

$$s = s_0 + \sum_{i=1}^{N} F_i s_i. \tag{2}$$

The first 9 principal components preserve 97.5% of variance. Illustration of the shape parameters $F_1..F_4$ is in Fig. 3.

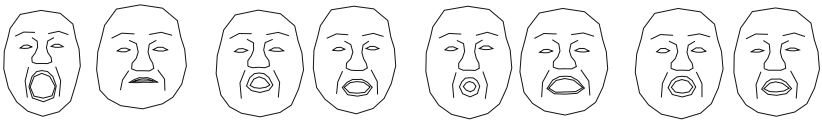

Fig. 3. First four modes of the shape model for $\pm 150\%$ of standard deviation

Fig. 4. Final fitting of AAM, from the left: fitted appearance of three consecutive input frames and incorrect tracking caused the occlusion

The appearance of AAM is an image defined as the RGB intensity. The eigenvectors are obtained by second PCA on warped training images. 41 texture parameters describe 97.5% of variance. Finally, combined AAM operates with a single set of parameters c to get the best fit of the AAM in an input video frame. Vector c is obtained by another PCA computed from the appropriately weighted shape a texture parameters.

The illustration of head tracking is in Fig. 4. AAM is sensitive to the initial shape and can end in local minima. Therefore the searching algorithm requires initial localization of face in the first frame of each processed video file. The most likely area showing the speaker's face is detected via a tree-based 20x20 gentle adaboost frontal face detector [16].

6 Experiment

The aim of the experiment is to prove the potential of automatic categorization of lexical signs. In the experiments we make use of parameters from tracking. From the tracking of MC we have obtained a contour of hands and head. The values of the contour are in absolute image coordinates. That means that position is also encoded into the contours. From the tracking of NMC we have obtained the shape and texture parameters, Sec. 5. The categories for MC and NMC were chosen similar to the linguistic categories of signs. The linguists have not yet established a universal categorization of signs so we tried to choose more abstract categories. This approach will allow us to describe more detailed categories by combination.

Categorization of Manual Component. For this experiment we have chosen categories summarized in Table 1.

To determine the category of a sign we need to compute 2D trajectories of the centroids of the contours. Then the sum of variance of x and y components of the trajectory

Table 1. The sign categories chosen for the experiment

Hand movement	*Body contact*	*Hand location*	*Head*
one handed	no contact	at waist	mouth wide open
two handed	contact of head and right hand	at chest	mouth wide closed
symmetric	contact of head and left hand	at head	lip pressed together
non-symmetric	contact of hands	above head	lip pucker
	contact of everything		closed eyes

determines whether the sign is one handed or two handed. If the variance is sufficient enough it means the hand has moved. To determine the symmetry of the trajectory we compute the sum of absolute values of Pearsons correlation coefficients for x and y positions of both hands. If the trajectories are correlated enough (better than 0.89 each dimension) we claim the sign trajectories are symmetric. The absolute value of the correlation coefficient reflexes the anti-symmetry that occurs in symmetric signs. The location of hand symbolizes what space relative to the location of the head has the hand occupied the most. We compute a histogram of relative y positions of hands consisting of 5 bins. The bins are chosen so that they correlate with the categories. Then the category connected to the most occupied bin is chosen. This approach can fail if the sign duration is relatively small to the video duration. That is why we consider only the segment of the video where the hands are moving and are out of starting position. The last category is body segment contact. For now we can only tell whether the objects occlude each other relative to the camera or touch each other. This is a necessary condition for the body parts contact, but not sufficient. Further experiments are needed. This condition is met when two trackers report the same object as the tracked one. This can be recognized easily since both (or all three) body parts will be represented by the same contour.

Categorization of Non-manual Component. In this experiment, we focus on shape information. The information includes the position of face in an image extracted from positional parameters and geometric information extracted from the shape model.

We consider categories of NMC uniquely described by a predetermined subset of all parameters. Few of the categories are mentioned in Table 1. The parameters X describing one category tend to cluster around their single mean value. We considers a simple Gaussian model as the univariate normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$.

The minimum variance unbiased estimator provides us with the sample mean ($\hat{\mu}$) and variance ($\hat{\sigma}$) for randomly selected signs that are manually labeled to the categories. We consider the likelihood of parameterized frame and the category $\mathcal{C}$ as:

$$L(\mathbf{x}, \mathcal{C}) \approx \prod_{i \in \mathcal{C}} f_{\mathcal{C}}(x_i | \hat{\mu}, \hat{\sigma}^2). \tag{3}$$

The algorithm determines (3) for all categories and all frames of the video file. The likelihood of the category $\mathcal{C}$ given the sign $\mathcal{S}$ (the video file) can be derived from the maxima over all frames:

$$L(\mathcal{C}|\mathcal{S}) = \max_{\forall \mathbf{x}} L(\mathbf{x}, \mathcal{C}) \tag{4}$$

The positional parameters describe the x, y translations and rotation in the optical axis independently. However, the contribution of parameters of the shape model into the particular categories in consequence of the used PCA is not evident. Albeit, for example, the first shape parameter describes the opening of the mouth very well , see Fig. 3, however the remaining shape parameters incorporate the partial opening of the mouth as well.

For robust fit of (3), we use the shape parameters only for back projection to the shape s (2). Normalization of the training set of AAM ensures that s is always horizontally

aligned. This condition enables a definition of a new set of derived parameters: height, width of lips, closed eyes and raised eyebrows. A category ”small lip rounding”, for example, incorporates two derived parameters related to width and height of the lip.

7 Conclusion

The MC tracking algorithm has a 94.45% success rate. This rate was computed against manually annotated video files. A sign was tracked successfully if in every frame the configuration was determined according to the annotation.

The proposed face tracking algorithm fails if the hands occlude significant parts of the face (eyes, nose or mouth), see Fig. 4 on right. The tracking was successful approximately in 95% of signs.

In general, an automatic categorization provides additional information about lexical signs and extends the potential of searching. In the experiment, we consider only a subset of sign categories that can be automatically derived from the features from tracking. These categories can be expressed in a writing form as well, for example by the symbolic notations HNS and SW [17]. The user of the on-line dictionary can search for signs using one of the notation systems and form new search request entering relevant symbols. For every sign we are able to determine the confidence factor for every defined category.

Tracking results provide additional information about the sign. However, for example, repetitive movements of head or hand shape categorization require more complex models. Other categories such as nose folding, forehead wrinkles, cheeks inflate, presence of tongue and teeth require further research in particular with the texture parameters.

References

1. Aran, O., Burger, T., Caplier, A., Akarun, L.: Sequential belief-based fusion of manual and non-manual information for recognizing isolated signs. In: Sales Dias, M., Gibet, S., Wanderley, M.M., Bastos, R. (eds.) GW 2007. LNCS (LNAI), vol. 5085, pp. 134–144. Springer, Heidelberg (2009)
2. Trmal, J., Hrúz, M., Zelinka, J., Campr, P., Müller, L.: Feature space transforms for czech sign-language recognition. In: Interspeech 2008, pp. 2036–2039 (2008)
3. Krňoul, Z., Kanis, J., Železný, M., Müller, L.: Czech text-to-sign speech synthesizer. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 180–191. Springer, Heidelberg (2008)
4. Ong, S., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning, pp. 873–891 (2005)
5. Zieren, J., Canzler, U., Bauer, B., Kraiss, K.: Sign Language Recognition, Advanced Man-Machine Interaction - Fundamentals and Implementation, pp. 95–139 (2006)
6. Hrúz, M., Campr, P., Železný, M.: Semi-automatic annotation of sign language corpora (2008)
7. Wang, Q., Zhang, W., Tang, X., Shum, H.Y.: Real-time bayesian 3-d pose tracking. IEEE Transactions on Circuits and Systems for Video Technology 16(12), 1533–1541 (2006)

8. Zhang, W., Wang, Q., Tang, X.: Real time feature based 3-d deformable face tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 720–732. Springer, Heidelberg (2008)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685 (2001)
10. Volker, B.: Face recognition based on a 3d morphable model. In: Proceedings of FGR 2006, pp. 617–624. IEEE Computer Society, Washington, DC, USA (2006)
11. Zhou, M., Liang, L., Sun, J., Wang, Y.: Aam based face tracking with temporal matching and face segmentation, pp. 701–708 (2010)
12. Campr, P., Hrúz, M., Langer, J., Kanis, J., Železný, M., Müller, L.: Towards czech online sign language dictionary - technological overview and data collection, Valletta, Malta, pp. 41–44 (2010)
13. Piater, J., Hoyouyx, T., Du, W.: Video analysis for continuous sign language recognition. In: LREC 2010, 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (2010)
14. Buehler, P., Everingham, M., Zisserman, A.: Employing signed tv broadcasts for automated learning of british sign language. In: LREC 2010, 4th Workshop on the Representation and Processing of Sign Languages (2010)
15. Aran, O., Ari, I., Campr, P., Hrúz, M., Kahramaner, D., Parlak, S.: Speech and sliding text aided sign retrieval from hearing impaired sign news videos, Louvain-la-Neuve, TELE, Universite catholique de Louvain, pp. 37–49 (2007)
16. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001, 4th Workshop on the Representation and Processing of Sign Languages, IEEE Computer Society Conference (2001)
17. Krňoul, Z.: New features in synthesis of sign language addressing non-manual component. In: LREC 2010, 4th Workshop on the Representation and Processing of Sign Languages, ELRA (2010)