Local Binary Pattern based features for Sign Language Recognition

Marek Hrúz, Jana Trojanová, Miloš Železný

University of West Bohemia, Faculty of Applied Sciences, Pilsen, Czech Republic
{mhruz, trojana, zelzeny}@kky.zcu.cz

Abstract. In this paper we focus on appearance features particularly the Local Binary Patterns describing the manual component of Sign Language. We compare the performance of these features with geometric moments describing the trajectory and shape of hands. Since the non-manual component is also very important for sign recognition we localize facial landmarks via Active Shape Model combined with Landmark detector that increases the robustness of model fitting. We test the recognition performance of individual features and their combinations on a database consisting of 11 signers and 23 signs with several repetitions. Local Binary Patterns outperform the geometric moments. When the features are combined we achieve a recognition rate up to 99.75% for signer dependent tests and 57.54% for signer independent tests.

1 Introduction

Sign language (SL) recognition is from computer vision view closely related to the field of human activity recognition. As such it uses the same methodologies. From the perspective of data acquisition SL came a long way during the past 20 years. At the beginning data were acquired via cyber gloves (data gloves with accelerometers for 3D localization) [19]. Then regular gloves with different colors [2] and even with different color for each finger were used [20]. Nowadays the data acquisition is close to real environment (no gloves and uncontrolled background). Yet and effective algorithm which can recognize sign language in a real environment is still to come. This paper focuses on hand/head representation, especially on feature extraction.

The Local Binary Pattern (LBP), introduced by Ojala [15], serves for texture representation. The LBP is used across various computer vision fields (e.g. image synthesis, light normalization, face detection, face/expression recognition). In this work we experiment with LBP descriptors, to show its discriminative power for handshape classification (i.e. capability to capture enough information in order to distinguish among handshapes in the classification phase). In SL the hand appearance varies extremely, and existing works about SL recognition are mainly concentrated on binary image of the hand followed by geometric moments [17], [20], [1], [7].

The non-manual features (head and body pose, head movement, facial expression, lip movement) have lately got high interest in the sign language recognition field [17], [1]. They help to distinguish between signs that are very similar in the manual component. Usually such signs have opposite meaning like "tasty" vs. "disgusting" in Czech

SL. They have the same hand movement and handshape and the only differentiator is the facial expression (eyebrows lowering respectively, head nod).

The outline of the paper is as follows. Section 2 deals with the hand and the head localization. Section 3 describes the features used for manual and non-manual component of SL. Section 4 introduces the system of recognition. In section 5 we present our experiments and section 6 concludes our work.

2   Hand/Head Localization

In this section we describe how we process the images and localize the hands and the head. We work on a database with laboratory-like conditions (static background, dark clothes). This helps to speed up the process of localization so that we can focus on feature extraction. Also we want the localization to be robust and precise. We decided to follow the work [8]. We detect objects in the image using skincolor (Section 2.1). We initialize the trackers based on the approximately known locations of hands and head (Section 2.2) and finally track the objects using a distant measure and a probability model (Section 2.3).

2.1   Image Segmentation

First, we apply a skincolor segmentation in the processed image. For this purpose we have trained a look up table in which all RGB values are assigned a likelihood of being a skincolor pixel. The look up table has been trained on example colors manually selected from the database. These examples form a distribution in RGB space which we model as a gaussian mixture (GMM). This mixture is normalized so that for every RGB value we obtain likelihood values in range from 0 to 255. This way we are able to process images quickly only by looking for the likelihood of the given color in the look up table. We define a threshold to segment only the most probable colors. In our case the threshold was set to 127. After this processing we obtain objects in the image that represent the head and the hands.

2.2   Initialization

We assume that the signing person is facing the camera up front. The person should have his or her arms alongside body. The left hand in the right part of the image and the right hand in the left part of the image. We simply divide the image into parts where we expect a certain body part and look for the segmented objects. Sometimes fragments from background or signer's clothing can mimic a body part. We filter these objects out using predefined rules (size and size ratio). This should leave us with three identified objects.

2.3   Tracking

For each object we have a single tracker. The tracker is used to store information about the tracked object and to perform object comparison. In each frame, objects are detected

and filtered as mentioned in previous sections. The objects are compared with the objects from the last frame in a distance measure. Each object is described by several features. These include:

– 7 Hu moments of the contour: translation, scale & rotation invariant features
– a gray scale template: hands and head appearance representation
– scalar features - position, velocity, perimeter of the contour, area of the bounding box, area of the contour.

For each feature a different distance function is defined. For the scalar features it is an absolute difference. For Hu moments we define the distance function as

$$D_{Hu} = \sum_{i=1}^{7} \frac{1}{m_i^A} - \frac{1}{m_i^B} \tag{1}$$

where A denotes the first shape (tracked object in the last frame), B denotes the second shape (object in the actual frame),

$$m_i^A = sign(h_i^A) \cdot log(h_i^A) \tag{2}$$
$$m_i^B = sign(h_i^B) \cdot log(h_i^B) \tag{3}$$

where h_i^A is the i-th Hu moment of the shape A and accordingly for h_i^B. D_{Hu} then denotes the shape (contour) distance. And finally for the template we apply a normalized template matching method based on correlation. This way we obtain a 7D vector with distances. Normal approach would be to choose the nearest object (nearest in the feature space) as the tracked object. But in SL the objects are frequently occluded and self-occluded. This means that the nearest object is often not the tracked one. That is why we trained a GMM for each object (left hand, right hand, head) using an Expectation Maximization (EM) algorithm. The training samples are obtained from the database using bootstrap approach. The mixture defines the probability of the distance measurement. High probability means that the evaluated object is the tracked object and vice versa. The probability itself is considered to be the result of the distance measurement. All the results are stored in a database. For all trackers the best result is chosen and verified. If it is accepted as a plausible configuration of the hands and head the objects are stored in the trackers accordingly. If it is refused then the next best result is chosen.

2.4 Result verification

From the distance measurement each tracker has a best candidate among the detected objects to be the tracked one. But it does not mean that it is always the true solution. Based on the size of the bounding box the tracker is able to determine whether the tracked object is in occlusion with another one [8]. This information must be in compliance with the resulting configuration of hands. That is if a tracker detects an occlusion, the final configuration must report that at least two trackers track the same occluded object. In ideal case the trackers would report the occluded object as the best candidate from the distance measurement. On other hand if a tracker does not detect an occlusion the final configuration must respect this fact.

3   Hand/Head Shape Description

Previous section explained objects localization in a video sequence. This section defines the manual features used for the handshape description and the non-manual features used for the facial shape description.

The handshape description comprises two main steps: hand segmentation and feature computation. Both steps are important for proper and discriminative handshape representation. The output of the first step is either segmented hand image or binary version of segmented hand. Three types of features can be computed on top of the segmented hand image.

- Appearance (e.g. Gabor filter followed by PCA [9])
- Contour (e.g. parametric representation of boundary contour [3],[11])
- Geometric moments (e.g. area size, angle of object major axis [1],[7],[20])

This paper focuses on the appearance features (LBP) described in section 3.2. Their performance is benchmarked agaits the geometric moments described in section 3.1. For other feature extraction examples please see survey [17].
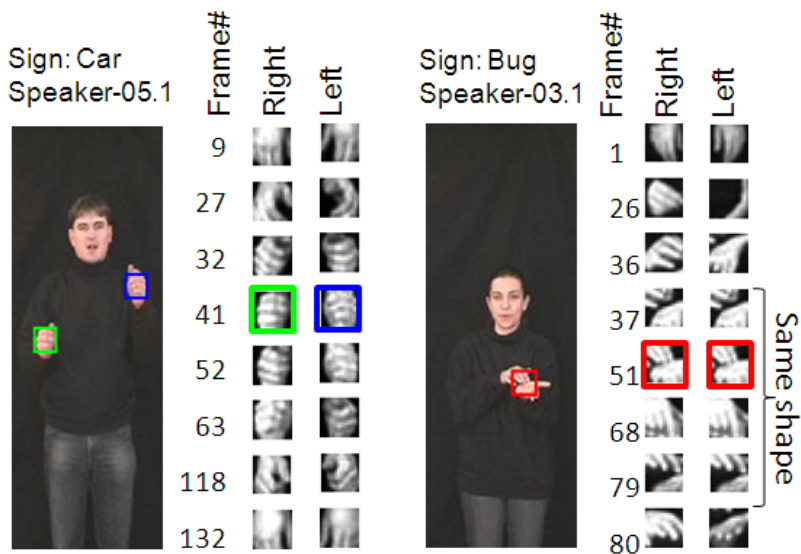

Fig. 1. Examples of segmented hands for LBP extraction

The important role of non-manual features has been pointed out in the sign language survey [17]. Several papers then concentrate on a subset of non-manual features. First comprehensive study experimenting with all non-manual features was done in [1]. Their

system uses active appearance model requiring training for each signer in database. The computed features consist of various distances among facial points (e.g. mouth width and orientation, distance between eyes and eyebrows). The section 3.3 describes facial feature localization system not requiring training on any signers in database.

3.1 Manual feature Extraction - Geometric moments

The geometric moments along with other information about the manual component can be derived directly from the tracking process. During the tracking we segment objects representing the hands and the head. We describe the known object with 10 features. It is the position of the center of the object, 7 Hu moments of the contour and the orientation of the object relative to the x-axis. Furthermore we compute the width of the object. This information is used in the post-processing phase to set a metric of the position space as mentioned in Section 4. The features are concatenated into a single vector of size 30. The selected features should describe the manual component of the sign as defined by sign linguists [17]. The position of the objects represent the location and movement of the hands, orientation of the objects describes the hand orientation and the Hu moments correspond to the handshape.

3.2 Manual Feature Extraction - Appearance

Appearance features do not require the hand segmentation to be as perfect (i.e. exact determination of pixels belonging to hand and pixel belonging to background) as when using geometric moments. A simple cropping around the hand center is satisfactory. Figure 1 illustrates the segmented hands. The Local Binary Patterns (LBP) serving for texture representation can be seen as appearance features. Here we want to investigate LBP feature vector, to show its discriminative power for hand shapes classification.

Figure 2 illustrates the forming of the LBP feature vector. Having a segmented hand image an LBP operator (non-parametric kernel) is applied to each pixel in the image. Created LBP image is then divided into sub-blocks. From each sub-block local LBP histogram is computed. Finally histograms are concatenated to form the LBP feature vector. Description of each step follows.

The LBP captures local texture information. We use multi-scale LBP operator with circular neighborhood of two pixel radius size [16] depicted in Figure 2. At given position, LBP is defined as an ordered set of binary comparison among center pixel intensity and surrounding pixels. The LBP contains two types of values - 0 for neighborhood intensities below the center pixel intensity and 1 vice-versa. The LBP is of length 8, therefore 256 possible combinations exist (0,1 are two values $2^8 = 256$). To visualize LBP image each pattern is assigned a decimal label from 0 to 255, thus a gray-scale mapping can be used.

The LBP histogram represents region texture information. Histogram measures the quantity of LBP in the sub-block of the image. Each bin in histogram stands for one LBP label. The regular size of the histogram is 256 bins. A shorter version of histogram exists when only uniform LBP is used. The uniform LBP is a pattern which has maximally two pattern changes among 0 and 1 value (e.g. 00011100). This condition is fulfilled
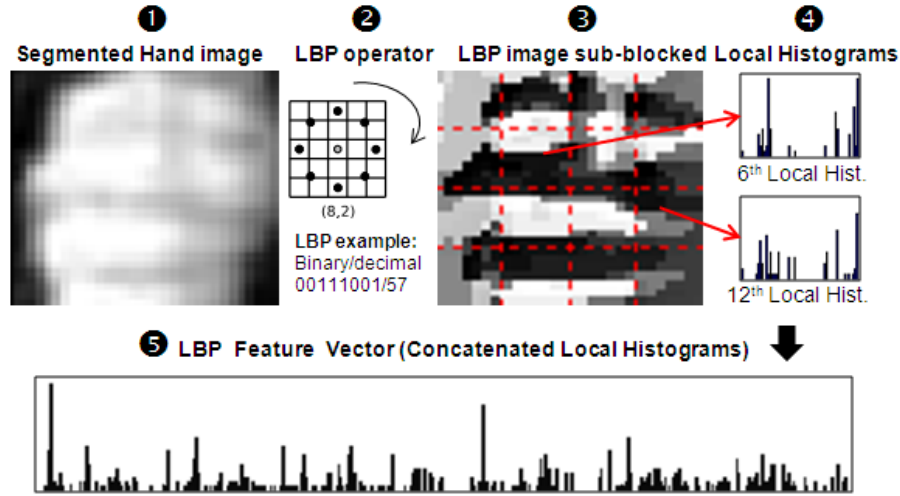
Fig. 2. LBP feature vector forming

with 58 LBPs the rest of them is assigned to an extra bin. Thus the histogram size for uniform LBP has 59bins.

The LBP feature vector captures the global description of a handshape, due to concatenation of local histograms. The concatenation of histogram is up to the user, but it should be the same across all images. The LBP feature vector has huge size. For instance having an image of size 32x32 pixel, size of sub-block 8x8 and uniform LBP results in 944 dimensional feature space. Such size is quite big for recognition via HMM, see Section 4.

One way to reduce the size of the LBP feature vector is to increase the size of sub-blocks over which LBP histogram is computed. Another way is to lower the image resolution. However the size of the image has to be sufficient to capture enough information about the texture so that we are able to distinguish among handshapes.

There are other forms of LBP, for instance multi-scale block LBP or volume LBP. Due to their high dimension we do not investigate them.

3.3   Non-Manual Feature extraction

To represent non-manual features (head and body pose, head movement, facial expression, lip movement) system localizing facial landmarks is used. The system combines active shape model (ASM) with landmark detector (LD). The code for ASM is inspired by [14]. To increase robustness of ASM fit a landmark detector inspired by [6] is used.

The localized facial landmarks are used to describe the non-manual features. The pose information is captured by all landmarks (change in relative position among landmarks represent the pose, no extra parameters are computed). The face expression is

captured by eyebrows and lip movement via landmarks on outer lip. Short description of system follows.

The Active Shape Mode is composed from point distribution model (PDM - representing the shape) and local appearance model (LAM - representing the landmark intensities around it). The ASM heavily depends on data introduced during the training phase. As pointed out in [14] more landmarks give you a better fit. To form a PDM model we use manual landmarks annotation for XM2VTS database available at [13]. This database does not contain facial expression. Thus we use CMU database [10] and interpolate the missing landmarks via mean shape generated from XM2VTS. To form a LAM at each landmark CMU database was used. For more details on ASM see [14].

The landmark detector was trained for localization of four landmarks (eyes' centers, nose tip, and a point on the center of upper lip). Training data were extracted from BioID [4] and CMU database [10]. To read more about the landmark detector, reader is directed to [6].
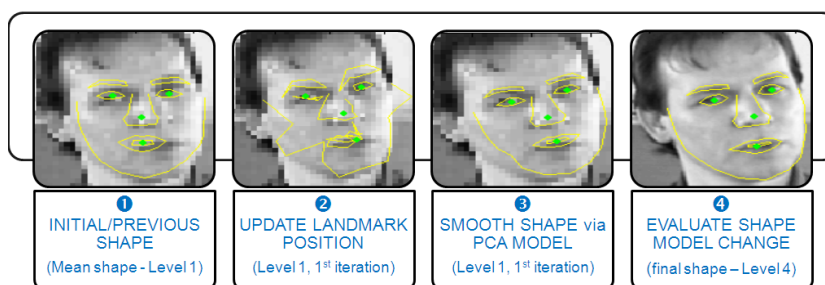

Fig. 3. Active Shape Model combined with Landmark Detector

Figure 3 illustrates the process of model fitting. First the mean shape is placed on the image, then the landmark detector is applied to search four landmarks in close neighborhood around the mean shape. In the second step mean shape is aligned with detected landmarks and each model point is updated based on its LAM. The third steps smooths the shape using the PDM model. In fourth steps the change between previous and actual shape is evaluated. If 80 % of landmarks did not change the model has converged. These four steps are repeated over four image resolutions. The landmark detector is used only at lowest resolution.

4   Sign Recognition

While building a sign recognition system following factors must be taken into account: intra/inter-signer variation, pose, lighthing, and occlusion. Each part of the system described in section 2 and 3 reduces these factors.

The occlusion is handled by the tracker. When occlusion of two hands appears, both hands are represented by same object (example shown in Figure 1). In case of head occlusion the same principle is used.

The variability of the object size (whether a different speaker is captured, various distance to camera or different resolutions) is addressed via normalization to a reference object.When geometric moments are used the head position and width is used as reference. Thus hand coordinates are specified relatively to the head center and moments are normalized by head width. For the manual appearance features the size of ROI around the hand is relative to the head width. The facial features are scaled to have a face width equal to 1.

The lighting independent component are used. The geometric moments are computed on binary image. The only hook is proper skin color segmentation. The LBP features have a capability to reduce the illumination effect since they work with relative brightness levels.

For sign recognition we use a HTK toolkit capable of learning the intra/inter-signer variation. Following the work [18] the left-right Hidden Markov Model (HMM) is used. In every experiment the first and the last state is non-emitting. To perform optimal recognition each feature set requires a different number of HMM states and different number of gaussians in a mixture for state modeling (in experiment denoted as *# of state mixtures*). To obtain the optimal number of these two parameters a brute force test was applied with no initial guess.

When training HMM we have to take into account the dimension of the data versus the number of data available. The problematic part is when GMMs are trained using the EM algorithm. When we have high dimensional data we need more samples to robustly train the GMM.

5   Experiments

We have evaluated the performance of sign recognition for isolated signs with respect to features as follows:

  – Comparison among individual manual features (Table 2)
  – Selection of best non-manual features (Table 3)
  – Combination of manual and non-manual features (Table 4)

Table 1 lists all features used for experiments. The experiments were realized for signer-semi-dependent (SSD) and signer-independent(SI) recognition test. Data used for experiments consist of 23 signs performed by 11 signers with 3-6 repetitions, in total 1065 video files. Training data for the SSD test contains 60% of repetitions for each signer and every sign, remaining data is used for testing. Training data for the SI test contains all signs from 60% of signers, testing is done for unseen signers.

At first we tested the performance of individual manual features. The results are shown in Table 2. The LBP features have better performance then geometric moments for SSD and SI tests. The LBP_2 computed over whole segment of hand image increases the SI test by 10% in comparison to LBP computed over segmented hand image. This is caused due to variation in hand center localization. The LBP computed over 4 subblocks

Table 1. List of features used for experiments. First part lists the manual features, second part lists the non-manual features, and third part lists their combination. Abbreviation: Dim - dimension of feature vector, LBP - Local Binary pattern, GM - geometric moments, ASM - active shape model, PCA - principal component analysis

Features Abbreviation	Features Description	Dim
GM	Hu moments, location, orientation	30
$GM + \Delta GM$	GM + delta coefficients between 2 frames	60
LBP	Cropped hand 16x16, hist. from 4 subblocks	472
LBP_2	Cropped hand 16x16, hist. from whole image	118
ASM_1	Inner + outer lip, eyes, eyebrows	96
ASM_2	Inner + outer lip, eyebrows	62
ASM_3	Outer lip, eyebrows (best non-manual)	48
ASM_{3_norm}	ASM_3 centered by head in first frame	48
ΔASM_3	Delta coefficients between 2 frames	48
$ASM_3 + \Delta ASM_3$	ASM_3 + its delta coefficients	96
ASM_4	Outer lip	24
ASM_5	Inner lip	14
ASM_6	Eyebrow	24
$GM + LBP_2$	Manual-geometric + Manual-appearance	148
$GM + ASM_3$	Manual-geometric + best non-manual	84
$LBP_2 + ASM_3$	Manual-appearance + best non-manual	166
$Comb - 1$	$GM + LBP_2 + ASM_3$	196
$Comb - 2$	$GM + LBP_2 + ASM_3 + \Delta GM$	226
$Comb - 2_{PCA}$	$(GM + LBP_2 + ASM_3 + \Delta GM)_{PCA}$	40

Table 2. Recognition rate for individual manual features.

Features	# of HMM states/ # of state mixtures	Signer-Semi Dependent	# of HMM states/ # of state mixture	Signer Independent
GM	9 / 10	83.09%	10 / 10	27.90%
$GM + \Delta GM$	9 / 30	87.58%	-	-
LBP	6 / 10	94.36%	6 / 10	58.18%
LBP_2	8 / 10	**94.61%**	5 / 10	**68.42%**

is more sensitive to proper hand localization. Even slight shift of hand center causes the same shape look differently, especially for unseen signer.

In the next set of experiments we search for the best set of the non-manual features. All experiments were done for SSD test only, see Table 3. We have omitted the SI test since the non-manual features alone do not have the capability to differentiate all signs [1]. The best results were obtained for ASM_3 (outer lip contour, eyebrows). We thus experiment with the best set of non-manual features. First we normalize head location in scene (as reference point the head center in the first image was used). The normalization

Table 3. Recognition rate for individual non-manual features.

Features	# of HMM states/ # of state mixtures	Signer-Semi Dependent
ASM_1	8 / 20	41.94%
ASM_2	8 / 20	41.94%
ASM_3	8 / 20	43.75%
ASM_3	7 / 40	48.99%
ASM_{3_norm}	10 / 30	31.85%
ΔASM_3	9 / 20	30.85%
$ASM_3 + \Delta ASM_3$	9 / 30	**51.01%**
ASM_4	8 / 20	35.89%
ASM_5	8 / 20	33.47%
ASM_6	8 / 20	41.33%

reduces the recognition by 17%. Then we computed delta coefficients (computed as a difference in the signal along the time axis). The results indicate that delta coefficients improve the recognition rate, but alone they give lower results.

Table 4. Recognition rate for combined features.

Features	# of HMM states/ # of state mix.	Signer-Semi Dependent	# of HMM states/ # of state mix.	Signer Independent
$GM + LBP_2$	8 / 20	92.72%	5 / 10	57.54%
$GM + ASM_3$	8 / 40	83.09%	-	-
$LBP_2 + ASM_3$	9 / 10	92.89%	9 / 30	20.10%
$Comb - 1$	9 / 10	90.69%	-	-
$Comb - 2$	10 / 5	91.67%	10 / 10	35.17%
$(Comb - 2)_{PCA}$	10 / 5	99.75%	9 / 10	37.93%

The last set of experiments investigates the combination of individual features. For the SSD test the combination $comb - 2$ successfully recognizes all realizations of signs except one. However for the SI test the recognition significantly dropped. The best performance for SI test was obtained with LBP_2 (individual manual features). Also the combination with non-manual features results in lower recognition rate. The reason might be in HMM recognizer (high dimension of the combined feature vector versus the small number of samples in training data).

6 Conclusion

The experiment results showed that feature representing hand shape via appearance are more discriminative than geometric moment containing information about the trajectory

and parametric representation of hand (orientation, area size, etc.). On the other hand the appearance of the hand is strongly dependent on the position and orientation relative to the camera. To reduce the effect of different point of view the camera placement should be front looking.

Using the non-manual features more than half of the signs were recognized correctly. Anyway these results depend on the type of signs in database. Having a large set of data might lead to lower recognition rate. The main message of our experiments with non-manual features is that eyebrows and outer lip contour are discriminative enough to capture the difference among the signs. The future plan in short period time is benchmarking of our non-manual features versus the one described in [1].

The combination of features $comb - 2$ has for the SSD test the best recognition rate among all experiments. This information is valuable when we want to recognize a database with low number of known signers (for example in a dictionary). The SI test achieved relatively low recognition rate. This can be due to relatively high dimension of the data versus small number of samples.

In future we intent to use the LBP features for sign sub-unit recognition. The sub-units can be used directly for sign recognition or can be used for other purposes like automatic sign annotation (from linguistic point of view). Another interesting task is to use the sub-units to represent the sign in a symbolic notation like hamnosys. The notation can be then used for sign searching with sign dictionary applications [5] or as an input for hamnosys driven signing avatar [12]. The most challenging will be the definition of these sub-units.

7 Acknowledgments

References

1. Agris, U. von, Knorr, M., Kraiss, K.-F.: The significance of facial features for automatic sign language recognition, *The 8th IEEE International Conference on Automatic Face & Gesture Recognition* pp. 1-6, (2008).
2. Aran, O., Ari, I., Benoit, A., Carrillo, A. H.., Fanard, F. X., Campr, P., Akarun, L., Caplier, A., Rombaut, M., Sankur, B.: Sign Language Tutoring Tool. *eNTERFACE 2006, The Summer Workshop on Multimodal Interfaces, Istanbul, Turkey, p. 23-33, HAL - CCSD*, (2006).
3. Bader, T., Rfipple, R., Beyerer, J.: Fast Invariant Contour-Based Classification of Hand Symbols for HCI, *The 13th international Conference on Computer Analysis of Images and Patterns*, vol. 5702, pp 689 - 696. (2009).
4. http://www.bioid.com/support/downloads/software/bioid-face-database.html
5. Campr P., Hrúz M., Langer J., Kanis J., Železný M., Müller L., Towards Czech on-line sign language dictionary - technological overview and data collection, *LREC 2010, 7th international conference on language resources and evaluation; 4th workshop on the representation and processing of sign languages: corpora and sign language technologies*, pp. 41-44, (2010)

6. Cristinacce, D., Cootes, T.: Boosted Regression Active Shape Models, textitProc. 18th British Machine Vision Conf., vol. 2, pp. 880-889, (2007).
7. Holt, G. A., Reinders, M. J. T., Hendriks, E. A., de Ridder, H., van Doorn, A. J., Influence of Handshape Information on Automatic Sign Language Recognition, *8th International Workshop on Gesture in Human-Computer Interaction and Simulation* (2010).
8. Hrúz, M., Campr, P., Železný, M.: Semi-automatic annotation of sign language corpora, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, (2008).
9. Huang, D.-Y., Hu, W.-Ch., Chang, S.-H.: Vision-Based Hand Gesture Recognition Using PCA+Gabor Filters and SVM, *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, iih-msp, pp.1-4, (2009).
10. Kanade, T., Cohn, J. F., Tian, Y.: Comprehensive database for facial expression analysis, textitAutomatic Face and Gesture Recognition, pp. 46-53, (2000).
11. Kelly, D., McDonald, J., Markham, Ch.: A person independent system for recognition of hand postures used in sign language, *Pattern Recognition Letters*, vol. 31 Pages 1359-1368, (2010).
12. Krňoul, Z., Kanis, J., Železný, M., Müller, L.: Czech Text-to-Sign Speech Synthesizer, *Lecture Notes in Computer Science*, pp. 180-191, (2008).
13. http://www.milbo.users.sonic.net/stasm/
14. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model, textitProceedings of the 10th European Conference on Computer Vision: Part IV, vol. 5305, pp 504-513, (2008).
15. Ojala, T.,Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions , *Pattern Recognition*, vol. 29, 51-59, (1996).
16. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with loval binary patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 24, 971-987, (2002).
17. Ong, S.C.W., Ranganath, S., Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873-891, (2005).
18. Trmal, J., Hrúz, M., Zelinka, J., Campr, P., Müller, L.: Feature space transforms for czech sign-language recognition, *In ICSLP*, (2008).
19. Wang C. G., Gao W., Ma J.: A Real-Time Large Vocabulary Recognition System For Chinese Sign Language, *Advances in Multimedia Information Processing, PCM 2001*, pp. 150-157, (2001).
20. Zieren, J., Canzler, U., Bauer, B., Kraiss., K.-F.: Sign Language Recognition, *Advanced Man-Machine Interaction - Fundamentals and Implementation*, Book pp 95-139, (2006).