Evaluation of Feature Space Transforms for Czech Sign-Language Recognition

Jan Trmal and Marek Hrúz
{jtrmal, mhruz}@kky.zcu.cz

Department of Cybernetics
University of West Bohemia
306 14, Plzen, Czech Republic *

Abstract. In the paper we give a brief introduction into sign language recognition and present a particular research task, where the access to MetaCentrum computing facilities was highly beneficial. Although the problem of signed speech recognition is currently being researched into by many research institutions all around the world, it lacks of a generally accepted baseline parametrization method. Our team introduced a parametrization method based on skin-color detection and object tracking. Because of the relatively high amount of information that is produced during the parametrization process, a method that reduces the unnecessary information while keeping the necessary information is required. Such methods are called *feature space dimension reduction* methods. We used the MetaCentrum facilities to evaluate several methods used widely in the field of acoustic speech recognition and their influence on recognition score of a Czech Sign-Language recognizer.

1 Introduction

For deaf people, sign language is the basic mean of communication – just as speech is for hearing people. Inspired by speech recognition, where the rate of progress has been enormous in the past decade, new ways of communication between deaf people and computers or hearing people are being developed. The main task of automatic sign language recognition is to recognize a sign performed by a signer.

In speech recognition it is obvious that microphone is used as the input device. In sign language recognition more kinds of input devices can be used. Mechanical devices, which measure location of various parts of body, such as data gloves and haptic devices have the advantage in accuracy of measurements. But there is a serious setback – the signer is forced to wear a cumbersome device. Another approach to this problem is to use a camera or even multiple cameras as input devices.

* This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, under project No. 1ET101470416, the Ministry of Education of the Czech Republic, project No. MŠMT LC536 and project No. ME08106 and by the Grant Agency of the Czech Republic, project No. GAČR 102/08/0707.

Lets consider the output of the camera or multiple cameras. Given the resolution of one picture and the frame frequency, the amount of data produced is huge. This is because the visual stream bears not only the bare information relevant for sign language recognition (SLR) but also additional information about the speaker, clothing, environment, etc. This additional information does not improve the recognition score – quite contrary, the information increase complexity of the task, dimensionality of the feature vectors and slow down the overall research process. The first thing we have to do is to reduce the amount of unnecessary information, while keeping the information useful for the SLR problem. This is the task of visual feature extraction (VFE) system. The VFE system employs computer vision algorithms to isolate (extract) the relevant information.

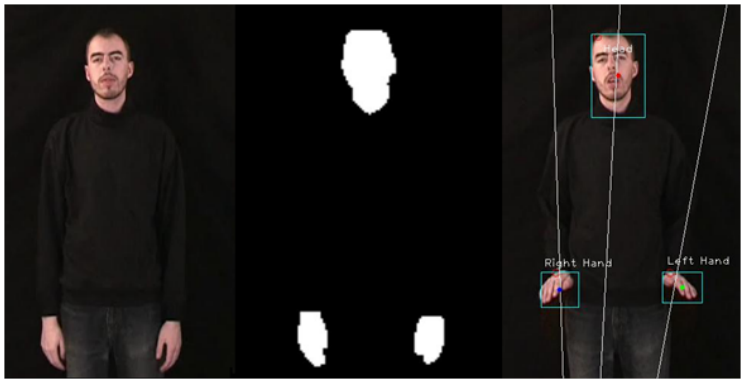

Fig. 1. a) source frame from the camera, b) segmented image c) head and hands tracking

1.1 Visual Feature Extraction System

The sign language consists of manual (i.e. employing hands) and non-manual (i.e. mainly facial expressions) component. We aimed our work against the manual component. The sign linguists distinguish several basic subcomponents of the manual component of the sign – hand shape, palm and hand orientation, location and movement.

To enable successful recognition of a sign, the VFE system must be not only able to find and isolate the hands, but also provide information about location in space and, more importantly, about the progressive change of the hand location. The process of finding the objects of interest in the image is called *object detection* and the process of monitoring of movements is called *object tracking*. In our case, the *objects* are left and right hand and the head of the speaker.

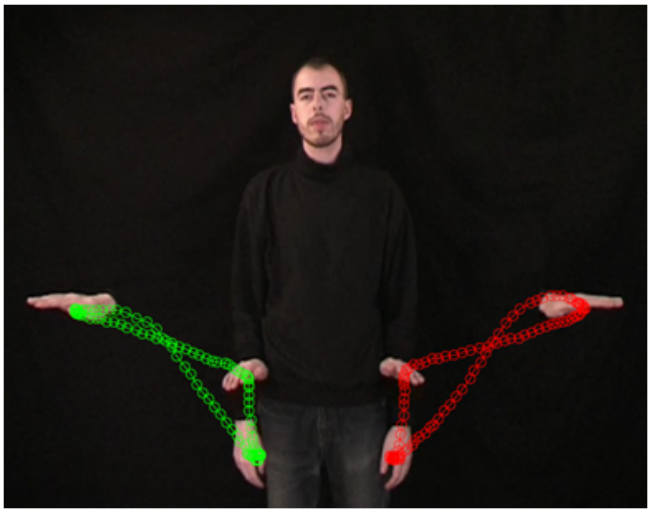

Fig. 2. Hand tracking for one of the signs

Object Detection A common method for detecting parts of a human body is the skin-color detection ([4]). Skin-color detection can be combined with motion cues ([2]) or edge cues ([5]). Although the method is widespread, it has several disadvantages. For example it is illumination dependent and there is a large variety in color of human skin. Therefore, an adaptation should be applied to the universal skin-color model ([7]). The skin-color detection is the first phase of VFE in our SLR system. Using the adapted, speaker-dependent model, we perform thresholding of the image – for every pixel in the image we determine its likelihood of being a skin-color pixel. If the likelihood is higher than a specific threshold, the pixel is considered to be a part of a skin-color object.

Object Description Every object isolated in the previous phase must be described by some features that correspond to the aforementioned manual subcomponents of the signs. For location and movement description we simply compute the 2D coordinates of the center of mass of the isolated object, for the shape we use the Hu's moments associated to the object.

Object Tracking After the object detection, we ideally obtain the objects representing the left hand, right hand and head. Our VFE system uses separate object tracker for each of these objects. The tracking process itself is based on measure of the distance of the old and new locations of the objects enhanced with occlusion detection system.

Feature Sets In the previous sections we described the VFE system. Using this system, we are able to reduce the amount of data in each input frame to 33 values, 11 for each of the tracked objects:

- x, y - the center of mass of the object
- 7 Hu's moments describing the shape of the object
- the angle of the object relative to the x-axis of the image
- a Boolean value representing whether the object is in occlusion

The occlusion flag is used in post-processing. If an occlusion is detected, the Hu's moments and the angle are linear interpolated between the last values before occlusion and the first values after occlusion. The final step of the post-processing is normalization in the spatial domain. The mean position of the head is considered as the origin and the mean width of the head is considered as one unit. The normalized features (excluding the occlusion flag) are concatenated in the following order: left hand, right hand, and head. For every object, 10 features are obtained, which makes a total count of 30 features for every frame.

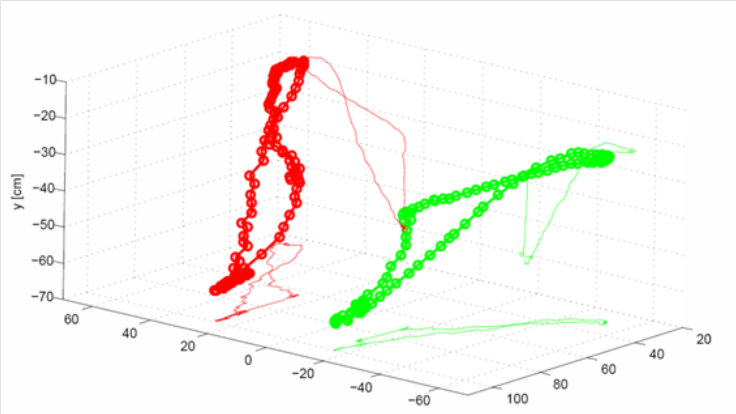

Fig. 3. Trajectory tracking in 3D space. The coordinates origin is located in the mean position of signers head

2 Feature Space Reduction Methods

Features obtained so far by the method described in the previous section are highly correlated and statistically dependent on each other. Also, the number of correlated features can be higher than the size of independent feature set – it can be proved ([1]) that the Hu system is dependent and incomplete. For our purpose, the Hu's moments are sufficient to describe the contours of the objects, but the dependence points at the possible use of a dimension reduction method.

In addition, it can be shown ([3]) that the number of the basic sign units can be interpreted as the Cartesian product of 4 sets (corresponding to the basic manual components) with cardinalities of 30, 8, 20, 40 - even when no context like "tri-signs" (as an analogy to tri-phones used in speech recognition) is considered. For this reason, the total number of model parameters to be estimated would become extremely large, particularly when one considers the limited size of a training corpus. Even if we had a good model, the total number of parameters is a limiting factor, when recognition in real-time is needed. For this reason, the choice of a suitable projection scheme method is a very important subtask of SL recognition. The projections schemes are usually $m \times n$ matrices, $m \geq n$, where m is the original dimension of feature vectors (30 in our case) and n the new feature vector dimension. The cause of the necessity to choose the right projection matrix only by experiments lies in the problematic definition of *relevant information*. We have investigated and compared 5 popular projection schemes: PCA as the baseline, ICA, LDA, HLDA, and rHLDA (HLDA with more robustly estimated covariance matrices). We will not discuss the methods properties or implementation, for further info, see paper ([6]).

3 Experiments and MetaCentrum Involvement

We used the MetaCentrum computation facilities to experiment with various possible choices of the constant n and, since our recognizer is based on Hidden Markov Models (HMM), with topology of the HMMs. To perform the experiments, we used the HTK toolkit for HMM modeling and proprietary C++ routines that we used through the Matlab MEX interface. The Matlab was directly available as a user module, but the HTK toolkit was not. After consultation with the MetaCentrum staff, they volunteered to prepare the HTK toolkit to be available in the same way as the Matlab is.

4 Publications Prepared Using MetaCentrum Resources

1. Jan Trmal, Marek Hrúz, Jan Zelinka, Pavel Campr, and Luděk Müller. Feature space transforms for czech sign-language recognition. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, pages 2036–2039. Causal Production Pty ltd., 2008.

References

1. J. Flusser. Moment invariants in image analysis. *Proc. of World Academy of Science, Engineering and Technology*, 11:196–201, 2006.
2. K. Imagawa, S. Lu, and S. Igi. Color-based hand tracking system for sign language recognition. *Proc. Int. Conf. Automatic Face and Gesture Recognition.*, pages 462–467, 1998.

3. S.C.W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27:873–891, 2005.
4. J. Sherrah and S. Gong. Resolving visual uncertainity and occlusion through probabilistic reasoning. *Proc. British Machine Vision Conf.*, pages 252–261, 2000.
5. J.-C. Terrillon, A. Piplr, Y. Niwa, and K. Yamamoto. Robust face detection and japanese sign language hand posture recognition for human-computer interaction in an "intelligent" room. *Proc. Int. Conf. Vision Interface*, pages 369–376, 2002.
6. Jan Trmal, Marek Hrúz, Jan Zelinka, Pavel Campr, and Luděk Müller. Feature space transforms for czech sign-language recognition. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, pages 2036–2039. Causal Production Pty ltd., 2008.
7. Y. Xiong, B. Fang, and F. Quek. Extraction of hand gestures with adaptive skin color model and its applications to meeting analysis. *Proc. IEEE Int. Symposium on Multimedia 2006*, pages 647–651, 2006.