Towards automatic annotation of Sign Language dictionary corpora

Marek Hrúz Zdeněk Krňoul Pavel Campr Luděk Müller

Univ. of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic

September 4, 2011

FACULTY
OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

1. Introduction and Motivation
2. Data
3. Hand/head tracking
4. Face tracking
5. Experiments with categorization
6. Conclusion

Introduction

Sign Language

Form of communication, mainly used by deaf or hearing impaired people. Consists of non-manual and manual component.

Manual component (MC)

hand shape, palm orientation, arm movement

Non-manual component (NMC)

face expression, body pose, lip movement

Motivation

The need of bi-directional translation is very important to enable communication between users of different languages

It is hard to make a good bi-directional dictionary from SL to spoken/written form

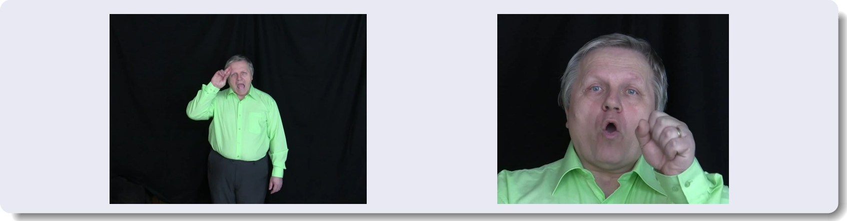
Currently the translation is provided by human interpreters

Our goal is automatic categorization of videos in the dictionary

That should enable a better search of signs in the dictionary

Data

- Data for our experiment are selected signs from on-line dictionary (http://signs.zcu.cz/)
- They consist of pairs of synchronized video files capturing one speaker from two different views
- The first view captures the signer's entire body, the second view captures a detail of the face



The conditions are: constant lightning, uniform background and clothing, long sleeves

In total we processed 213 video files

Hand/head tracking - MC

- Is based on skin-color segmentation (it is ok, because of the character of the data)
- The resulting objects (blobs) are filtered - only probable blobs remain (1-3 blobs: left, right hand and head)
- We track the blobs using **discriminative measures** and **probability models**
- We employ 3 trackers each with 4 different models ($m_1 \dots m_4$), each model is modeling different situation

Last state / new state	*Not occluded*	*occluded*
Not occluded	m_1	m_2
Occluded	m_3	m_4

The models are GMMs, trained on annotated data

In every frame the detected blobs are compared with the last known blobs - all combinations

The comparison yields from 7D vectors of differences

Feature vector

1. normalized correlation between gray scale intensities of blob and tracked blob
2. normalized distance between their contours (computed from Hu moments)
3. relative difference between their bounding box areas
4. relative difference between their perimeters
5. relative difference between their areas
6. relative difference between their velocity
7. relative difference between their location

Using *Heteroscedastic Linear Discriminant Analysis* (HLDA) we reduce the dimension to 5D - the models ($m_1 - m_4$) are then 5D GMMs with 4 components each

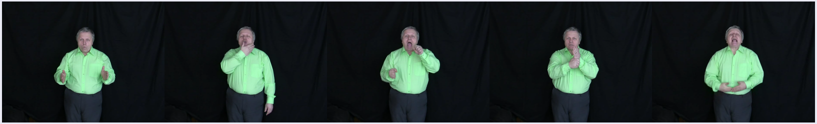
This enables us to compute the probability between the new objects and the last known objects

We can just choose the most probable solution for each tracker independently, but often the alone-standing models are not powerful enough

The same blob can be identified as both hands (in occlusion) and the other blob is left unlabeled

Therefor we take into account the configuration of the head and the hands - we test the number of detected blobs

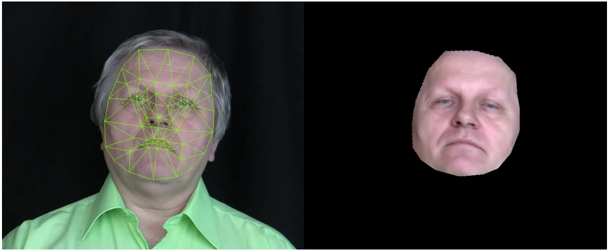
5 configuration of head/hand configuration



In the end we choose the most probable configuration based on the number of detected body parts. The result of tracking is the trajectory of all body parts and the contours of the blobs.

Face expression tracking - NMC

- Generative parametric models are commonly used to track human faces in images
- 2D multi-resolution combined active appearance model (AAM)
- AAM is based in PCA and combines model of shape and texture (appearance)
- **Shape model** provides geometric features very useful for categorization
- **Texture model** ensures robust tracking of face expression



Training of AAM

- Training set consists of 51 random selected images of the face to cover maximum variation in the face
- We manually identify the mesh in each training image
- 9 shape and 41 texture principal components preserve 97.5% of total variance

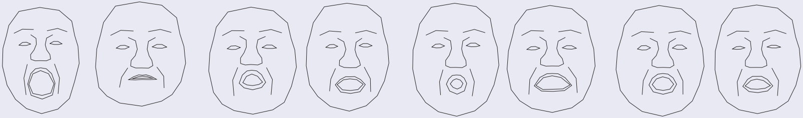
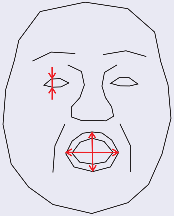


Figure: Example of first four shape principal components (±150% of standard deviation)

- Contribution of the shape principal components into acceptable categories in consequence of the used PCA is not evident
- For example, the first shape parameter describes the opening of the mouth, however the remaining shape parameters incorporate the partial opening of the mouth as well
- For each frame of the input video, we use the AAM only for identification of the face shape
- We extract following three face measures:



Processing of video files:

AAM is sensitive to the initial shape and can end in local minima

We have to use initial localization of face in the first frame of each processed video file

For this purpose, the most likely area showing the speaker's face is detected by Viola-Jones adaboost face detector

Finally, we have trajectories of the face measures for all frames and lexical signs

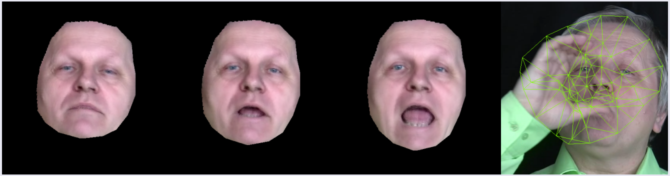


Figure: Final fitting of AAM, from the left: fitted appearance of three consecutive input frames and incorrect tracking with occlusion.

Categorization of lexical sign - Experiments

Linguists have not yet established an universal categorization

We chose more abstract categories so that we can build on them in the future

For experiment, we consider following categories:

Table: The sign categories chosen for the experiment.

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

For MC, we have 2D trajectories of the contours and centroids of the blobs

1. The sum of variance of centroids determines confidence factor for one and two handed variant
2. For symmetry, we consider a sum of absolute values of correlation coefficients for the left and right hand - if the trajectories are correlated enough the sign is symmetric
3. The confidence factors of body contact categories are directly extracted from GMM models of hand/head trackers
4. 5 bins histogram from relative position hands and head determines category for location of the sign

The NMC categories are derived from face measures

- For each category, we considers one simple Gaussian model defined for relevant face measure only
- These models are trained on extracted frames from manually labeled video files
- Trained models provide likelihood of the categories for all frames of the input video
- We propose the final NMC confidence factor of one category is maxima over likelihoods of all categorized frames

Conclusion

- We proposed new framework for automatic tracking and categorization of video files capturing lexical signs interpreted one signing human and with predefined conditions
- Proposed tracking method for MC has a 94.45% success rate against manually annotated video files
- For NMC, tracking was successful approximately in 95% of signs, the algorithm fails if the hands occlude significant parts of the face
- In the experiment, we consider only such categories that can be automatically extracted from video frames
- In conclusion, the proposed automatic categorization provides additional annotation about lexical signs and extends the potential of searching and translation in the SL dictionaries

Thank you!