Lips tracking using AAM

Miroslav Hlaváč[1]

1 Introduction

The task of computer speech recognition is dependent on acoustic background. In an environment with noise or for people with voice disorders the audio speech recognition could fail. We are looking for ways of providing additional information to the audio signal to increase the recognition rate. Such information may be the shape of the lips [2].

The shape can be obtained by different methods from a video record of the speaker. I have used statistical models of appearance to represent the shape of the lips. These methods were developed by Tim Cootes [1]. Cootes introduced two methods: Active Shape Model (ASM) and Active Appearance Model (AAM). ASM is using only shape information. AAM is adding texture to the shape. AAM is used for the purpose of this paper.

2 Implementation

AMM is composed of a shape model and a texture model. Both models are represented by a linear equation

$$x = \bar{x} + \Phi\, b \tag{1}$$

where $\bar{x}$ is a mean value of the shape(texture), Φ is a matrix containing eigenvectors of the shape(texture) model and b is set of parameters controlling the model. The model of lips' shape is created by selecting significant points along the lips' inner and outer boundary line. The area between these points is then divided into triangles and the texture is sampled from them. The triangles are important to distinguish which pixels belong to the shape. Principal Component Analysis is then separately applied on the sampled shape and texture data to reduce the model dimension. Both models are then put together to create a combined model which represents both shape and texture by one set of parameters b.

The search is started in the first frame of a video by putting the mean shape of the combined model on the expected position of speaker's lips. The texture under the mean shape is sampled and an error vector is computed by subtracting the sampled data and the mean texture of the combined model. New set of parameters b is then generated from the error vector. More information about this can be found in Cootes' paper [1]. New parameters b are used to create a new instance of the combined model and the process is repeated until the error vector converges to its minimal value. Parameters b are then used as starting parameters in the next frame and the process continues.

[1] student of postgraduate study programme Applied science and Informatics, field Cybernetics,
e-mail: mhlavac@kky.zcu.cz

3 Training data

I have created a model with 19 significant points. Training data for this model were manually selected from 190 pictures. Matlab was used to process the data and to track the shape of the lips. Images for training were selected from 10 hours of video recordings that were made specifically for the purpose of this project.

4 Results

The recognition rate of this algorithm depends on the global illumination in the video. Darker videos have worse recognition rate then brighter videos. The result is also dependent on the starting shape. I have find out that in videos with 25 frames per second the difference between shapes in two subsequent frames can be so big that the search algorithm will fail. This can be addressed by starting from different shapes and then selecting the result for which the norm of the error vector was the smallest.

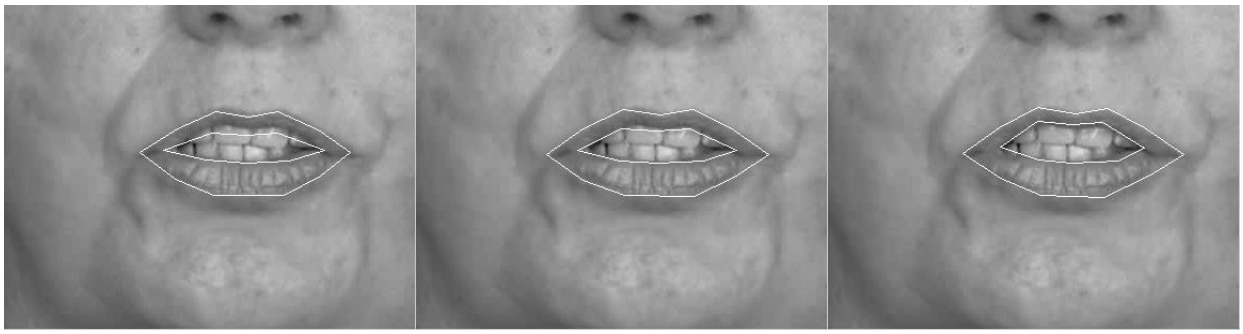

Figure 1: The shape after first, third and fifth iteration

5 Conclusion

Active Appearance Model was introduced in this paper. The algorithm was implemented and tested in Matlab. AAM can be used to track the lips' shape with an error lesser than 6 pixels.

Acknowledgement

This paper was supported by grant SGS-2013-032. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

References

[1] Cootes, T. F. and Taylor, C. J.. Statistical Models of Appearance for Computer Vision, Manchester, 2004.

[2] Císař, P.. Using of Lipreading as a Supplement of Speech Recognitioni, Plzeň, 2006