



# Viseme Recognition Using Multiple Feature Matching

*I. Shdaifat and R. Grigat*

*Stefan Lütgert*

TU Hamburg Harburg, Vision Systems  
D-21071 Hamburg, Germany  
{shdaifat,grigat}@tu-harburg.de

Philips Semiconductors, Systems Laboratory Hamburg  
D-22529 Hamburg, Germany  
stefan.luetgert@philips.com

## Abstract

In this paper, we present a technique for the extraction of the five main visemes produced in natural speech for German. The method belongs to the LDA (Linear Discriminant Analysis) family. The intensity, the edges, and the line segments are used to locate the lips automatically and for viseme classification. Using many features in the recognition maximizes the probability of recognition rate. The corners of the mouth are used in case of small rotation and scale.

An experiment has been carried out on different people, to understand the part of the speech that the human being use. The people grouped the phonemes into five different visemes. The number of distinguished visemes is not the same for each speaker. Everyone express the speech in a different visemes. Good recognition rate has been achieved on different speaker.

## 1. Introduction

Automated speech perception systems are very sensitive to background noise. They fail totally when multiple speakers are talking simultaneously (cocktail party effect). Humans deal with these distortions in considering additional signal sources. Besides the acoustic signals from both ears, visual information, mostly lip movements, are subconsciously involved in the recognition process [1].

Understanding how the human being reads information from the lips is a challenging aspect in audiovisual speech processing. This is because of the visemes and part of the speech for the same phoneme are widely varying from one person to another. Nevertheless, the human being can surprisingly differentiate between different shapes of the lips, and relate them to a group of phonemes from still images. Dodd [6] grouped the sounds or phonemes to 14 groups of visemes for English. A reasonable grouping was supposed by Abry [5]. The phonemes in French were subdivided into six groups according to the lip shapes. To understand how human beings do this, a group of people has observed sequences without sound. The results shows that the people have grouped the phonemes mostly into five common groups of visemes.

Tracking of the lips in image sequences and relating the features of the lips to the speech are also two challenging problems. This is due to the lack of dominant image features defining the lips [7]. For lip location and visual features extraction, several method were proposed in the literature. Deformable templates [4] [8] are effective in lip tracking, though quite computationally expensive [7]. Principal component analysis (PCA) was used [2]. But not each form of lips can be described by PCA, and a large data base of objects is needed [3]. A 3D model of the lips using two calibrated cameras was proposed [10]. However, the human being can read the speech just from still 2D images.

Here, we present a method for lip tracking and viseme classification using feature templates matching. Our method belongs to the Linear Discriminant Analysis (LDA). Given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean difference between the desired classes [3]. We try to use all the available features of the lips to discriminate between the various viseme classes. Because the human being uses all the available features in the classification. For a certain speaker, five images of the five different visemes are acquired. The different features of the visemes are stored in a template. These templates are used to track the lips and to recognize the viseme. The recognition rate of the visemes is the sum of the contributions of the various features. The contribution of each feature to the recognition depends on the dominance of the feature and on the lip and the mouth shape.

The method was tested for different people and has resulted in a good recognition rate. The test sequences can be sent via email.

The paper is organized as follows: in sec. 2 we describe how the phonemes are related to mouth shape, and how they are divided into visemes similar to the way that the human being does. In sec. 3 the use of the feature templates in locating the lips from the face of the speaker is illustrated. In sec. 4, the classification of the features into the viseme classes is illustrated. In sec. 5 a comparison between different sequences and the results are presented.

## 2. Grouping Phonemes into Visemes

The human being can relate the lip shape to groups of phonemes. This can be noticed especially when the lips are mainly used to produce a certain phoneme. For example, in the phonemes **m p b** the mouth must be closed.

To find the key shapes for visemes, image sequences of different speakers without sound were displayed image by image to different people. The observing people were asked to organize the phonemes they have seen in categories. The phonemes were subdivided mainly into five groups depending on the lip shape.

1. **C**: closed lips *m, b, p, ...*
2. **O**: open lips *a, h, ...*
3. **H**: half open lips *t, d, k, g, s, ...*
4. **L**: long lips *i, e, ...*
5. **R**: rounded lips *o, u, ü, ö ...*

The viseme classes were grouped into 14 groups in [6] and 6 groups in [5]. But this is not always the case for all the speakers. In some cases, the boundaries between the visemes are not



clear. This reduces the number of the viseme classes to even less than 5, which was the case in our experiment.

Fig. 1 shows the images of the five viseme classes. For each viseme class the features of the lips are extracted and stored in templates. Each template contains the two images of the left and right corners of the mouth, the edge image of the viseme image, line segments of 10 pixel length fitted to the edges.



Figure 1: An example of images of five viseme classes.

### 3. Tracking of the Lips

We assume that the image, in which we want to locate the lips, includes mainly the face of the speaker. An average image representing the lips is correlated with the speaker's image pixel by pixel, see Fig. 2. The average image of the five viseme images is a reasonable representation of these images. The speaker's image and the pattern image are scaled down to speed up the search. We use the maximum intensity correlation  $\rho$  to locate the lips.

The position of absolute maximum correlation is not always the correct position of the lips. The speaker can move towards and away from the camera. Consequently, the size of the lips image could be smaller or larger than that of the visemes in the templates. The head can be rotated, so the lips will not always be in the horizontal standard position given in Fig. 1. The eyes are two objects where the correlation coefficient can be larger than for of the lips location.

To avoid sticking with pseudo locations, we inspect other features of the lips. The edge image is found and fitted with 10 pixel line segments. Distinguished features of the lips are the corners of the mouth. We consider those positions, where the correlation coefficient  $\rho$  is greater than a certain threshold value. Along the line segments, we search for the corner of the mouth pattern the same way that was used for the lip pattern. Using these two corners, we rotate and scale the image of the lips to be in the standard position for further viseme recognition. Fig. 3 illustrates this. In sec. 4 further refinement will be done to reject non lips objects.

The region of interest, ROI, is set around 150% of the lip size at the current lips' location. The process is restricted to the ROI if  $\rho$  is less than a certain value minimum every 20 images.

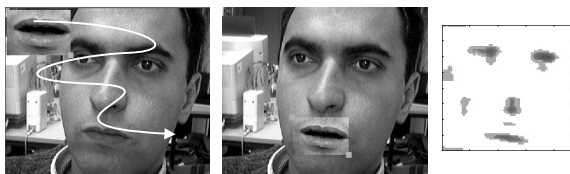


Figure 2: Search for the lip as a single feature, left the searched pattern, middle the pattern matched, right regions where  $\rho >$  threshold.

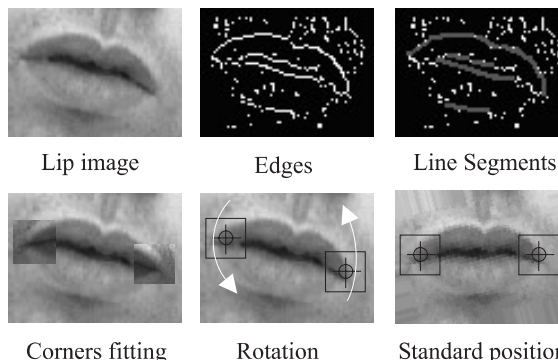


Figure 3: Searching for the corners of the mouth in ROI, and putting the lip in standard position

### 4. Viseme Recognition

Different visemes are very close to each other and only very small variations exist between them. This makes the recognition a non-trivial task. Fig. 4 shows a diagram of the five viseme classes and the large overlap between them. Nevertheless, human beings can notice these small differences, because they use many small details as features of the speaker lips and mouth.

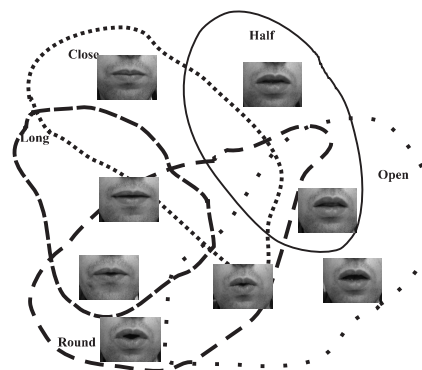


Figure 4: Overlap between the different viseme classes.

In sec 3, the image of the lips was detected, and was put in the standard position. Now, we have an image which should be similar to one of the viseme templates. How do you as a human being recognize the lips and the visemes? To understand this take the example in Fig 5. There is no problem for anyone to recognize that these are closed lips. This means we use intensity information, the boundary of the object. Another feature, that can be use, is the boundary segments. These segments are robust and contain more information than the boundary itself, for example the direction. A typical example of the human being's ability to recognize an object segmented into lines is the sleeping cat in [9].

The resulting image of the lip is compared with the viseme template. Each template contains the intensity image, the edge image, the line segments. Because we have no idea about the current spoken viseme, we have to compare it with the five templates.



Figure 5: Features of the one viseme.

For intensity comparison the intensity correlation  $\rho_i$  is used. If we have two images  $I_1$  and  $I_2$ , with  $N$  pixels each, and let  $\bar{I}_1$  and  $\bar{I}_2$  be the averages of  $I_1$  and  $I_2$ , then

$$\rho_i = \frac{\sum (I_1 - \bar{I}_1)(I_2 - \bar{I}_2)}{\sigma_1 \sigma_2 N^2}$$

For the edges  $\rho_e$

**for each edge point in image  $I_1$  do**  
**if edge point exists in  $w_e \times w_e$  window in  $I_2$  then**  
 counter = counter + 1  
**end if**  
**end for**

$$\rho_e = \frac{\text{counter}}{N \text{ edge points of } I_1}$$

For the lines  $\rho_l$

**for each line segment in image  $I_1$  do**  
**if line segment exists in  $w_l \times w_l$  window in  $I_2$  then**  
**if angle of line segment in  $I_1$  and in  $I_2$  are equal then**  
 counter = counter + 1  
**end if**  
**end if**  
**end for**

$$\rho_l = \frac{\text{counter}}{N \text{ line segments of } I_1}$$

The total recognition rate is  $\rho_t = \rho_i + \rho_e + \rho_l$ .  $\rho_t$  is calculated for each viseme template. The output viseme is that with maximum  $\rho_t$ . If  $\rho_t$  is less than a certain threshold value the search for lips is repeated again for the whole image. If  $\rho_t$  is again below the threshold the output viseme is "don't know".

## 5. Results

We tested our method for different speakers. Every speaker read the following passage: "Leider ist unbekannt, ob blonde oder ...", see the sequences. The duration of each sequence is 20 sec at 10 frames per sec (*fps*). For image size  $352 \times 288$ , the output frame rate was about 5 *fps* on Pentium III 700 MHz. The normalized recognition rate is shown in the output sequences. The location of the lips was incorrect only for few frames.

The way how people express the visemes varies significantly. For some of the speakers, it was difficult to distinguish between very close visemes e.g. **L** and **H**, or between **C** and **L**. This could be observed in e.g. sequence 3, where the **L** viseme was difficult to be detected.

Luetin [11] has observed that an accurate measurement of the quality of a lip tracking algorithm is indicative of how well it performs in recognition of visemes or words. Our quality measurement is the ratio of the number of correct estimated visemes to the number of the real visemes.

Table 1 shows the percentage of the correct estimated visemes for each sequence. For sequence 2, Table 2 shows the recognition rate and the error in recognition of each viseme separately.

The speaker in sequence 4 has beard and moustache, the viseme **H** was always classified as **R** or **L**. The reason is that

Table 1: Recognition rate of visemes.

| Sequence        | 1    | 2    | 3    | 4    |
|-----------------|------|------|------|------|
| Correct visemes | 61 % | 73 % | 59 % | 43 % |

Table 2: Viseme recognition for sequence 2. The visemes horizontally are the expected, vertically are the recognized

| Visemes | C    | O    | R    | L    | H    |
|---------|------|------|------|------|------|
| C       | 55 % | 0 %  | 0 %  | 0 %  | 0 %  |
| O       | 5 %  | 84 % | 0 %  | 7 %  | 35 % |
| R       | 15 % | 0 %  | 84 % | 4 %  | 4 %  |
| L       | 15 % | 12 % | 5 %  | 60 % | 22 % |
| H       | 10 % | 4 %  | 16 % | 30 % | 37 % |

the person himself does not express the visemes clearly. On the other hand, because of the moustache the edges and the line segments did not give sufficient discrimination between the visemes. Consequently, the recognition rate was low in this sequence. Table 3 shows the recognition rate after reducing the image size to 75%. This simulates the case when the speaker is moving away from the camera. The recognition rate was reduced due to the insufficiency of lips features.

Table 3: Recognition rate for 75% zoomed image.

| Sequence        | 1    | 2    | 3    | 4    |
|-----------------|------|------|------|------|
| Correct visemes | 41 % | 49 % | 34 % | 29 % |

## 6. Conclusion & Future Work

We have presented a method for lip tracking and recognition of the five main visemes in German. This can also be applied to other languages. The visemes were divided into five groups according to the shape of the lips in natural speaking. All the features of the lips were used for the recognition. This makes the locating of the lips and viseme recognition robust.

The method has been tested for different speakers, and has achieved a good recognition rate of the visemes about 60%. The frame rate was 5 *fps*.

Currently, visemes are extracted only from one image at a time. We did not take into account the sequence of visemes. But the speech is determined by the absolute shape and the transition between the visemes. This will be taken into account in our future work.

The rotation and the scaling are estimated using the two corners of the mouth. The corners themselves are variable with the visemes and not a fixed feature. A suitable reference for scaling and rotation of the mouth are the eyes.

## 7. Acknowledgment

This work is supported by Philips Semiconductors, Systems Laboratory Hamburg Germany.

## 8. References

- [1] C. Bregler, H. Hild, S. Manke, and W. A. "Improving connected letter recognition by lipreading". In International



- Joint Conference of Speech and Signal Processing, volume 1, pages 557–560, Mineapolis, MN, 1993.
- [2] C. Bregler and Y. Konig “‘Eigenlips’ for robust speech recognition”, Proc. Int. Conf. Acoust. Speech Signal Process. Vol. 75, 1994, pp 669-672.
  - [3] Aleix M. and Avinash K. “ PCA versus LDA”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 23, No 2, Feb. 2001, pp 228-233.
  - [4] Marcus E. Hennecke, David G. Stork, and K. Venkatesh Prasad. “Visionary speech: Looking ahead to practical speechreading systems”. Speechreading by Humans and Machines, volume 150 of NATO ASI Series, Series F: Computer and Systems Sciences, Berlin, 1995. Springer Verlag.
  - [5] Abry, C. and Boe L. -J “Laws for Lips”, J. Speech Communication, Vol. 5, 1986, p 97-104
  - [6] B. Dodd and R. Cambell, Eds, “Hearing by Eye: The Psychology of Lipreading”. London: Lawrence Erlbaum, 1987.
  - [7] Lucey, S., Srindharan, S. and Chandran, V., “Initialized Eigenlip Estimator for Fast Lip Tracking Using Linear Regression”, ICPR’2000, Barcelona, Spain.
  - [8] M. U. Ramos Sanchez, J. Matas, and J. Kittler. “Statistical chromaticity modles for lip tracking B-splines”. In Int. Con. Audio and Video based Biometric Person Authentication, Crans Montana, Switzerland.
  - [9] R. Duda and P. Hart., “Pattern Classification and Scene Analysis”, John Wiley and Sons, 1973, New York.
  - [10] L. Revéret, C. Benoît. “A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production”. Proc. 2nd ESCA Workshop on Audio-Visual Speech Processing, Terrigal, Australia, Dec. 4-6, 1998.
  - [11] J. Luettin, N. Thacker, and S. Beet. Statistical lip modelling for visual speech recognition. In VIII European Signal Processing Conference, Trieste Italy, 1996.