ROBUST VIDEO PROCESSING FOR LIPREADING APPLICATIONS

Jacek C. Wojdeł and Leon J. M. Rothkrantz
Knowledge Based Systems Group,
Delft University of Technology,
Zuidplantsoen 4, 2628BZ Delft, The Netherlands
Phone: +31 15 278 8543
Fax: +31 15 278 7141
E-mail: J.C.Wojdel@cs.tudelft.nl, L.J.M.Rothkrantz@cs.tudelft.nl

ABSTRACT

In this paper we present a robust model for video sequence processing that is aimed at extraction of the data describing the mouth shape in the image. The method allows computationally light data extraction from the video sequence. The extracted data can then be used in lip-reading applications. The paper describes results of two different experiments with artificial neural network recognizers trained on the recognition of speech boundaries and vowel-consonant discrimination.

INTRODUCTION

When perceiving speech, human beings use not only the auditory signal; they use also other modalities such as lip movements, hand and body gestures etc. The importance of those modalities grows immensely in noisy environment or in case of hearing disabilities. It was shown that even well hearing subjects do lip-reading as a part of the speech recognition process. The influence of lip-reading is present even in perfect auditory conditions and it grows steadily with the degradation of the auditory signal (Adjoudani et al., 1997).

Automated lip-reading can be used in combination with audio-based speech recognition to improve the recognition rate of a system (Nakamura et al., 1997). In a little different approach, the visual information can be used to enhance the quality of the speech signal when transmitted on the distance (Girin et al., 1997).

Because of the huge amount of data contained in the visual signal it is usually not feasible to apply the signal processing techniques on it directly. Commonly, specific features or parameters are derived from each video frame and further processed. It is usually done by fitting a geometrical model on the image and using model parameters as the input for the recognition system (Coianiz et al., 1995; Luettin et al., 1995). In the following sections we will describe step by step a feature extraction method that does not rely on any geometrical lip model.

LIP-SELECTIVE IMAGE FILTERING

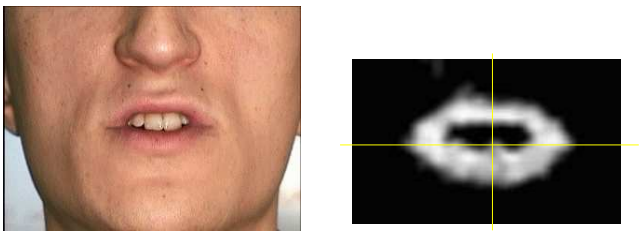
As the first step in video processing, we have to somehow locate the mouth in the image. Fortunately for us, when using the video containing only the face the lips have a distinct coloring that allows finding them without a need for complicated object recognition techniques. In order to utilize this fact, we need to apply some sort of *lip-selective* filter to the image. In our current research we use several different filters depending on the illumination conditions and the quality of the recorded video sequences.

Hue-based filtering

This is probably the simplest possible way of performing *lip-selective* filtering. It was first proposed in (Coianiz et al., 1995) and incorporates a parabolic shaped filter that cuts off the colors that fall outside a given interval:

$$F_{hue}(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & , \quad |h - h_0| \leq w \\ 0 & , \quad otherwise \end{cases} \qquad (1)$$

The filter is defined by the center of the interval h_0 and its half-width w. Filtering the image starts with calculating the *hue* color component of a given pixel and then



(a) original (b) hue-filtered

Figure 1: Parabolic shape hue-based lip-selective filtering example, (a) original image, (b) hue-filtered image with the mean value of resulting distribution depicted ($h_0 = 0$, $w = 0.25$).

passing it through the F_{hue} function. In order to select lips in the image, the value of h_0 must be around 0 (red color) and w about 0.2 (see Fig. 1). Both of the values have to be adjusted for a specific video sequence in order to retain the quality of the filtering.

We can use also the *hue* based filter in combination with the similar parabolic shaped *value* filter if the difference between the *hue* components of the face and lips is not sufficient for robust lip detection. Using a product of those two filters, one can for example remove the noise of the dark or bright background, where the *hue* values behave rather randomly. Both dark and bright ends of the HSV (*hue*, *saturation*, *value*) color space allow for big changes in both *hue* and *saturation* components with relatively small changes in the overall color perception. Therefore the *hue* based filter only is practically useless there.

Parameters of both *hue* and *value* based filters can be set automatically by optimizing filter's separation of color histograms prepared for lip-region and the rest of the face. We use here a simple interaction with the user in which the lip-region in the initial video-frame must be selected. From the selected region we extract the color histogram and perform the optimizations. Resulting filters are then presented to the user for confirmation (see Fig. 2). The total calibration time for both filters does not exceed 5s, and has to be performed only once for each recording setup.

Filtering with ANNs

A *hue* based filter is very simple and computationally effective. Unfortunately, during our experiments we have found out that in many cases, when the illumination of the face is not perfect, the *hue* component itself is not sufficient for proper lip selection (see Fig. 3b). Instead of putting additional constraints on the filtered color (such as a threshold on its *value* or *saturation*), we decided on using a black-box approach. We trained a very simple feed-forward neural network and used it as a filter. The network that is used has only 5 neurons in a hidden layer and one output neuron. It is fed with the *RGB* values from a pixel in the image. The small size of the network allows for on-line training, so that we can use different networks for different video sequences as long as the first frame in the video sequence is prepared with the target outputs for the network. In the same user interface as for calibrating the parabolic filters described earlier, we use the user-selected region to mark pixels as belonging to lips. The training of the ANN based filter takes significantly more time than optimizing parabolic filters, but it does not usually exceed 30s.

ROI extraction

Even with optimal filtering of the mouth, it is possible that in some cases the filtered image contains unwanted artifacts. A typical example can be seen in Fig. 3b-c, where the shadow under the chin cannot be distinguished from the lips. In order to reduce the impact of such occurrences, we must extract a region of interest (ROI) from the whole image. We do it by using vertical and horizontal signatures of the image. Each signature is a vector containing sums of the pixel-values for specific row or column. The boundaries of ROI can be found by looking for raising and falling edges in each of the vectors. Even if the artifacts are big enough to appear on the image signatures, their influence can be removed by assuming temporal consistency of ROI boundaries between consecutive frames. Only the part of the filtered image that lays inside ROI is further processed.

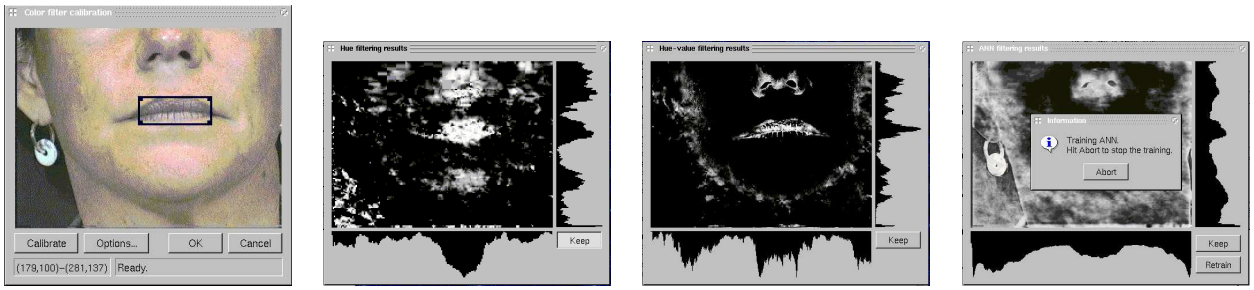
FEATURE VECTORS EXTRACTION

From this point on, the filtered image is being treated as a bivariative distribution $I(X, Y)$ (after normalization). The mean of this distribution: $[EX, EY]$ approximates accurately and in a stable way the center of the mouth. Using this value, we transform the image into polar coordinates:

$$J(a, r) = I(EX + r\cos(a), EY + r\sin(a)) \qquad (2)$$

and we concentrate on the conditional distributions $J(a = \alpha, r)$. We define therefore the following functions of mean and variance values for any angle α:

$$\begin{aligned} M(\alpha) &= \frac{\int_r J(\alpha,r)\cdot r\cdot dr}{\int_r J(\alpha,r)\,dr} \\ \sigma(\alpha) &= \frac{\int_r J(\alpha,r)\cdot(r-M(\alpha))^2\,dr}{\int_r J(\alpha,r)\,dr} \end{aligned} \qquad (3)$$



(a) (b) (c) (d)

Figure 2: User-supported calibration of lip-selective filters. (a) Region selection, (b) *hue* based filter, (c) combination of *hue* and *value* based filters, (d) ANN during training.

(a) original (b) hue-filtered (c) ANN-filtered

Figure 3: Using different lip-selective filters, (a) original image, (b) hue-filtered image with optimal filter parameters, (c) ANN filtered image.

As the image is discrete rather that continuous, all of the values are obtained from summation rather than integration, so we only operate on estimations of those values, namely $\widehat{M}(\alpha)$ and $\widehat{\sigma}(\alpha)$. The vectors resulting from sampling of those functions for one of the video-sequences can be seen in Fig. 4.

The value of $\widehat{M}(\alpha)$ for a specific angle relates directly to the distance of the lip from the center of the mouth in a given direction. Therefore a vector constructed from those values for a number of angles describes the shape of the mouth on a given image. Such shape profiles were gathered from a video sequence of spoken Dutch and grouped by the visemes. The Fig. 5 shows three profiles for visemes `[A]`, `[o u O y Y 2: 9:]` and `[p b m]`. In order to obtain scale independency the vectors were scaled so that all their values fit in $\langle 0, 1\rangle$ interval. It can be seen that the `[ou]` viseme can be distinguished from the others by its relatively high $\widehat{M}$ value in the middle. There is however no feasible way of distinguishing between `[A]` and `[p b m]` visemes. The lip shapes that correspond to those visemes are completely different (mouth closed versus mouth opened), but scaling of feature vectors removed any of those differences.

As the values of $\widehat{M}$ alone are not sufficient for useful data extraction from the video-sequence, we use also additional information about variance of conditional distributions ($\widehat{\sigma}$). Those values are also scaled in order to obtain size independency, but the scaling factor is directly determined by the values of $\widehat{M}$ rather than $\widehat{\sigma}$. Therefore the resulting feature vectors provide additional information about the lips. As an example, the `[A]` viseme clearly shows much lower $\widehat{\sigma}$ values than the other visemes (see Fig. 6). Obviously, for a wide stretched mouth, the lips appear thinner than for a mouth closed, when related to overall size of the mouth.

The last stage of feature vectors extraction is subsampling of the $\widehat{M}$ and $\widehat{\sigma}$ functions into vectors that can be further processed. Obviously, the choice of sampling the rate (vector dimensions) is a compromise between accuracy and processing efficiency. The longer the vectors, the more information on the original distribution they contain but also the longer it takes to extract and process them. After some experiments with the data we chose to use the 18-dimensional vectors for both $\widehat{M}$ and $\widehat{\sigma}$ (see (Wojdeł and Rothkrantz, 2000))

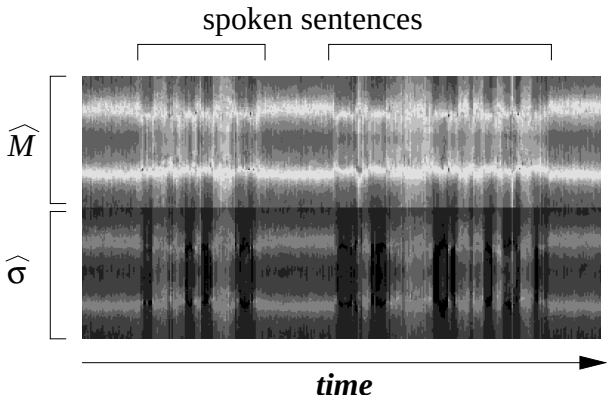


Figure 4: Pairs of $\widehat{M}(\alpha)$ and $\widehat{\sigma}(\alpha)$ vectors extracted from a video-sequence. The periods of silence around two spoken sequences can be seen clearly.

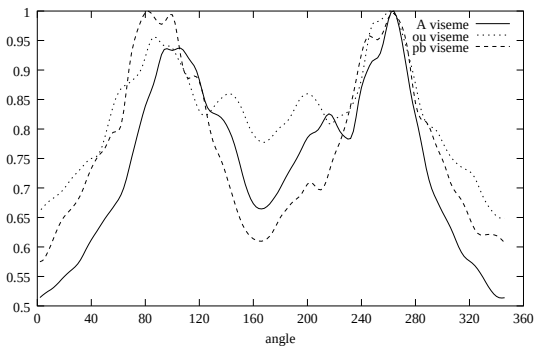


Figure 5: Average $\widehat{M}(\alpha)$ values for specific visemes in one of the video sequences.

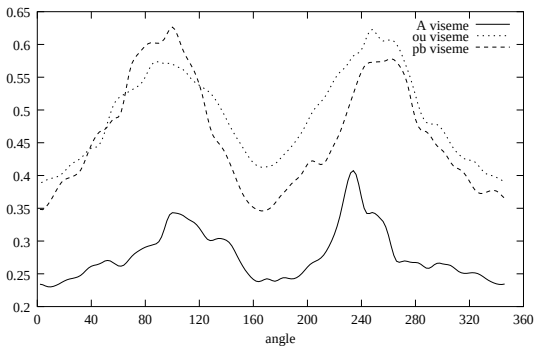


Figure 6: Average $\widehat{\sigma}(\alpha)$ values for specific visemes in one of the video sequences.

VALIDATION OF EXTRACTED DATA

In order to validate the correctness of the extracted data, we performed two different experiments related to lip-reading. The first one aimed at finding the boundaries of the spoken sentences in the video sequence, the other was to discriminate between vowels and consonants in it. In both cases we used a corpus of video recordings consisting of 30 Dutch sentences spoken by three different people. The subjects were asked to speak most of the sentences normally, some in a slower rate and to whisper the rest. In this way we obtained a varied data set that was then manually labeled with transcriptions of the spoken utterances.

Speech boundaries

As it can be seen in Fig. 4, the periods between sentences can be spotted in the data set even just by eye-inspection of the data stream. Therefore developing a classifier that would localize them is not a big problem. After experimenting with several ANN architectures (Wojdeł and Rothkrantz, 2000), we settled with a Jordan recurrent NN. Table 1 summarizes the results of testing the networks that were trained on a different subsets of our data. All networks were tested on about 3m40s of recordings containing all speech patterns (normal / slow / whisper). The table contains the average output error of the network as compared to target labeling of the data, the average delay of onset/offset detection (switching between two output states) and the total time of periods in which the network produced inappropriate output.

Vowel/consonant discrimination

In a second experiment, we investigated a problem of discrimination between two classes of the phonemes being spoken: vowels and consonants. We performed several experiments with different architectures of ANNs that were trained to put the incoming data into one of three different classes: vowel, consonant and silence between utterances.

The Elman Hierarchical Neural Network containing 3 hidden layers with respectively 20, 40 and 20 neurons in each of them proved to be the most efficient in vowel/consonant discrimination. The overall results of recognition are summarized in Tab. 2. Together with the three classes that can be recognized by the network (*0*, *v*, *c*) this table contains also a fourth column labeled (–), which corresponds to the output patterns that should not be classified as any of the three classes. Such patterns occur in our dataset for example in slow speech be-

Table 1: Speech onset/offset detection results for networks trained on different subsets of the recordings.

	training set		
	normal	slow	whisper
avg. output error	0.157	0.068	0.123
avg. detection delay	0.37s	0.17s	0.34s
misclassified	10.56s	2.08s	5.13s

Table 2: Vowel/consonant discrimination results. Symbols are: '0' - inter-sentence silence, 'c' - consonant, 'v' - vowel, '-' - not labeled.

	Network response			
Target	0	c	v	-
0	**70.5%**	14.3%	3.8%	11.4%
c	2.6%	**72.2%**	19.5%	5.2%
v	0.0%	26.5%	**67.3%**	6.1%
-	1.7%	45.2%	12.2%	**40.9%**

tween some phonemes that with a space between them that cannot be considered a silence yet. In the experiments we assumed that the network response was unspecified if none of the network's outputs had a higher activation than a threshold value of 0.1.

CONCLUSIONS AND FUTURE WORK

The proposed method of data extraction from the video sequence proves to be feasible in lip-reading related applications. Despite the small size of the dataset used in the experiments, the recognition rates are relatively high, which indicates that the proposed data representation is suitable for lipreading tasks.

The method's main advantages are it's simplicity and computational effectiveness. Those advantages are crucial when it comes to a real-time applications. The whole video processing is done as a single flow of rather low-level operations without need for per-frame computational heavy tasks (such as model optimization or template matching).

There are still some problems that have to be solved before applying this approach to the lip-reading system. For example the accuracy of *lip-selective* filtering depends on some parameters (h_0 and w or set of ANN weights) which cannot be set automatically yet. If we treat them however as a part of a calibration process that has to be done only once on the beginning of session, we have a choice of using more sophisticated (and possibly computationally expensive) methods of setting these parameters.

In the near future, a bigger corpus of video recordings must be gathered in order to proceed with the development of the system for continuous lip-reading. Having a significantly larger set of recordings (and so also prompts) will allow also investigation in how much the results depend on the contextual information extracted by an Elman network.

References

Adjoudani, A., Guiard-Marigny, T., Goff, B. L., Reveret, L., and Benoit, C. (1997). A multimedia platform for audio-visual speech processing. In (Kokkinakis et al., 1997).

Coianiz, T., Torresani, L., and Caprile, B. (1995). 2D deformable models for visual speech analysis. In (Stork and Hennecke, 1995).

Girin, L., Feng, G., and Schwartz, J. (1997). Noisy speech enhancement by fusion of auditory and visual information: a study of vowel transitions. In (Kokkinakis et al., 1997).

Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors (1997). *Proceedings of ESCA, Eurospeech97*, Rhodes, Greece. ESCA.

Luettin, J., Thacker, N. A., and Beet, S. W. (1995). Active shape models for visual speech feature extraction. In (Stork and Hennecke, 1995).

Nakamura, S., Nagai, R., and Shikano, K. (1997). Improved bimodal speech recognition using tied–mixture HMMs and 5000 word audio–visual synchronous database. In (Kokkinakis et al., 1997).

Stork, D. G. and Hennecke, M. E., editors (1995). *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, Berlin. Springer Verlag.

Wojdeł, J. C. and Rothkrantz, L. J. M. (2000). Visually based speech onset/offset detection. In *Proceedings of Euromedia2000*, Antwerp, Belgium.