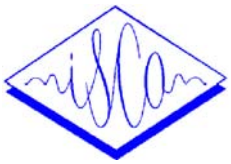

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

REAL-TIME LIP-TRACKING FOR LIPREADING

Rainer Stiefelhagen, Uwe Meier, Jie Yang
{stiefel|uwe}@ira.uka.de, yang+@cs.cmu.edu
Interactive Systems Laboratories
University of Karlsruhe — Germany, Carnegie Mellon University — USA

ABSTRACT

This paper presents a new approach to lip tracking for lipreading. Instead of only tracking features on lips, we propose to track lips along with other facial features such as pupils and nostril. In the new approach, the face is first located in an image using a stochastic skin-color model, the eyes, lip-corners and nostrils are then located and tracked inside the facial region. The new approach can effectively improve the robustness of lip-tracking and simplify automatic detection and recovery of tracking failure. The feasibility of the proposed approach has been demonstrated by implementation of a lip tracking system. The system has been tested by a database that contains 900 image sequences of different speakers spelling words. The system has successfully extract lip regions from the image sequences to obtain training data for the audio-visual speech recognition system. The system has been also applied to extract the lip region in real-time from live video images to obtain the visual input for an audio-visual speech recognition system. On test sequences we have achieved a reduction of the number of frames with tracking failures by a factor of two using detection and prediction of outliers in the set of found features.

1. INTRODUCTION

It has been demonstrated that not only humans benefit from visual input in speech perception and language understanding, but also the performance of automatic speech recognition systems can be significantly improved by adding lip movement information to the acoustic data [1, 2, 3]. A major problem of the previous systems was the intrusive way in which the user had to interact with the speech-reading systems. In order to acquire the visual data the user had to wear head-mounted cameras or reflective markers on his/her lips, or the relevant lip-region had to be extracted manually.

Over the last 4 years, the Interactive Systems Laboratory has been developing lip-reading techniques to enhance speech recognition [4, 5, 6]. We have attempted to achieve non-intrusive human-computer interaction and free the users from these interferences from very beginning. We have been developing lipreading systems based on a modular MS-TDNN structure. The visual and acoustic TDNNs are trained separately, and visual and acoustic information are combined at the phonetic level. In the system described in [4], the data acquisition was automatic and without any marker on the user's face, but the process required the speaker to position his lips in a window shown on a workstation screen. In the system described in [5], we integrated a face tracking and a lip-localization module into our system which allowed the user to freely position himself/herself in the view of the camera. The system is for speaker dependent continuous spelling of German letters. The system has achieved error reduction of up to 55% comparing with the system that only uses acoustic information. However, the neural net based lip-localization module did not perform in real time and could not be used online. The facial region had to be stored and lip-localization and extraction was done off-line, which slowed down the system significantly.

In this paper, we present a new approach to lip tracking for lipreading. Instead of only tracking features on lips, we propose to track lips along with other facial features such as pupils and nostril. In the new approach, the face is first located in an image using a stochastic skin-color model, the eyes, lip-corners and nostrils are then located and tracked inside the facial region. A 3D model which contains information about the relative 3D positions of these features in a human face is used to detect outliers in the set of found features, to predict the true positions of outliers and to detect complete tracking failure. Therefore, the new approach can effectively improve the robustness of lip-tracking and simplify automatic detection and recovery of tracking failure. We have implemented the approach in the lip-tracking module of our lipreading system. The new lip-tracker is able to find and track lips in real-time and allows online extraction of the lip region, which constitutes the visual part of information for the speech-reading system.

In the following section we will describe the methods for locating and tracking the face, eyes, lip-corners and nostrils. In section 3. we will discuss detection and prediction of outliers. In section 4. we will show results of the lip-tracker on image sequences that we recorded in our lab and on sequences of different speakers from a database.

2. TRACKING THE FACIAL FEATURES

2.1. Why Tracking Six Features ?

For a lipreading system, it is essential to track the lip region of the speaker. This can be achieved by tracking the lip-corners. If we are only interested in tracking the lip-corners, then why should we try to track *more* than just the lip-corners? First, it is difficult to locate or track lip-corners alone. In order to find the lip-corners within a face, we might have to search other facial features using certain constraints and heuristics. Some facial features are easier to locate than lip-corners. For example, within a face, the pupils are two dark regions that satisfy certain geometric constraints, such as position inside the face, symmetry according to the facial symmetric axis and minimum and maximum width between each other. Once the eyes are located, the locations of the lip-corners can be predicted. Second, tracking more features than necessary can improve the robustness of the tracking system. Since we know the relative positions of facial features, it is possible for us to use a 3D model to check if the found features are in some way consistent and can detect outliers

Figure 1. Iterative thresholding of the search window

in the set of found features and predict their positions in the next frame.

In the proposed approach, we first locate a face, then search the eyes inside the facial area, next locate the lips based on the found face size and eye locations, and finally find the nostrils. We will discuss these tracking procedures below.

2.2. Searching the Face

To find and track the face, we use a statistical color-model consisting of a two-dimensional Gaussian distribution of normalized skin colors [7]. The input image is searched for pixels with skin colors and the largest connected region of skin-colored pixels in the camera-image is considered as the region of the face. The color-distribution is initialized so as to find a variety of skin-colors and is gradually adapted to the actual found face.

2.3. Tracking Eyes

Assuming a frontal view of the face initially, we can search the pupils by looking for two dark regions that satisfy certain anthropometric constraints and lie within a certain area of the face.

For a given situation, these dark regions can be located by applying a fixed threshold to the grayscale image. However, the threshold value may change for different people and lighting conditions. To use the thresholding method under changing lighting conditions, we developed an iterative thresholding algorithm. The algorithm iteratively thresholds the image until a pair of regions that satisfies the geometric constraints can be found. Figure 1 shows the iterative thresholding of the search window for the eyes with thresholds k_i. After three iterations, both pupils are found.

Because the thresholding value is adjustable, this method is able to apply to various lighting conditions and to find the pupils in very differently illuminated faces robustly.

Once the pupils are found, they can be tracked by simple darkest pixel finding in small search windows in the following frames. These search windows furthermore can be predicted using linear extrapolation over positions in previous frames.

2.4. Tracking Lip Corners

To search the lips initially, the approximate positions of the lip corners are predicted, using the positions of the eyes, the face-model and the assumption, that we have a near-frontal view. A generously big area around those points is extracted and used for further search. Figure 2 shows the search windows for eyes and lips and predicted lip-corners marked with small rectangles. The found features are marked with crosses.

Finding the vertical position of the line between the lips is done by using a horizontal integral projection P_h of the grey-scale-image in the search-region. Because lip line is the darkest horizontally extended structure in the search area, its vertical position can be located where P_h has its global minimum.

The horizontal boundaries of the lips can be found by applying a horizontal edge detector to the refined search area and regarding the vertical integral projection of this horziontal edge image. The positions of the lip corners

Figure 2. Initial search areas for the lips, and found lip-corners. The small rectangles mark the predicted positions of the lip-corners.

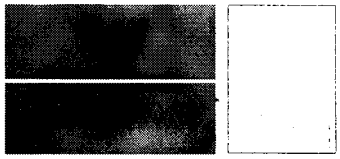

Figure 3. Integral projection of the search-window to find the vertical position of the lips

can be found by looking for the darkest pixel along the two columns in the search area located at the horizontal boundaries. This approach to search the lip corners using integral projections is based on ideas already described for example in [8].

We have developped a method to track lip corners in real-time in an illumination-independent way. Our approach consists of the following steps:

1. Search the darkest pixel in a search-region right of the predicted position of the left corner and left of the predicted position of the right corner. The found points will lie on the line between the lips
2. Search the darkest path along the lip-line for a certain distance d to the left and right respectively, and choose positions with maximum contrast along the search-path as lip-corners

Because the shadow between upper and lower lip is the darkest region in the lip-area, the search for the darkest pixel in the search windows near the predicted lip corners ensures that even with a bad prediction of the lip corners, a point on the line between the lips is found. Then the true positions of the lip corners can be found in the next step. Figure 5 shows the two search windows for the points on the line between the lips. The two white lines mark the search paths along the darkest paths, starting from where the darkest pixel in the search windows have been found. The found corners are marked with small boxes.

2.5. Tracking Nostrils

Searching and Tracking the nostrils in our system is also done by iteratively thresholding the search-region and looking for 'legal' blobs. Similar to searching the eyes, the nostrils can be found by searching for two dark regions, that satisfy certain geometric constraints. Here the search-region is restricted to an area below the eyes and above the lips. Again, iterative thresholding is used to find a pair of legal dark regions, that are considered as the nostrils.

Whereas we have to search a relatively big area in the initial search, during tracking, the search-window can be positioned around the previous positions of the nostrils, and can be chosen much smaller. Furthermore, the initial threshold can be initialized with a value that is a little lower than the intensity of the nostrils in the previous

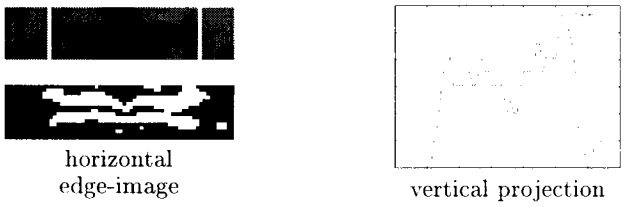

Figure 4. Finding horizontal borders of the lips, using a vertical projection of the horizontal edge-image of the lips

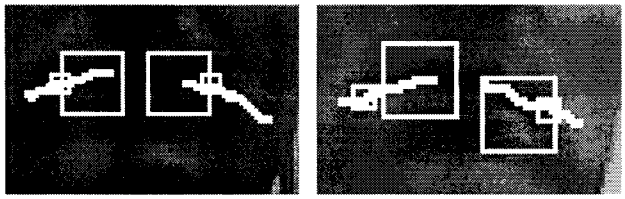

Figure 5. Search along the line between the lips

frame. This limits the number of necessary iterations to be very small.

3. DETECTING OUTLIERS AND RECOVERY FROM FAILURE

In order to build a robust usable tracking system, the system has to be able to detect tracking failure and to recover from it. In addition, to increase the robustness of the system, we try to find outliers in the set of found feature points, and predict their true position in the next frame.

Given the 3D locations of the facial feature points in a simple head-model and their locations in the camera image, the rotation and translation (pose) of the head with respect to the camera can be computed from these 3D- to 2D-point correspondences [9, 10]. The computed pose can be described through a 3x3 rotation matrix $\mathbf{R}$ and translation vector $\vec{t}$ which maps the head coordinate system onto the camera coordinate system.

Finding outliers in the set of found features can be done by computing the head-pose using different subsets of all found features for the pose computation and selecting a best or most consistent subset of feature points. The features that are not included in this best set are considered as outliers. To find a best subset we used a method proposed by Gee & Cipolla [11]. Using this method, the subset that leads to the pose implying the smoothest motion of the head is chosen as the best subset.

Once the best subset of features is found, the true position of an outlier can be easily predicted by projecting its model point into the image, using the computed pose. This prediction allows the system to recover from tracking errors and leads to a more robust tracking of the feature points.

To detect tracking failure, the average distance between the back-projected head-model points and their actual found locations in the image can serve as a measure of confidence. Once this average distance exceeds a certain threshold, tracking failure is considered and the system searches all features again.

4. EXPERIMENTAL RESULTS

We have evaluated the lip-tracking accuracy on two different test sets: First we recorded four sequences of one speaker in our lab (set 1) to hard disk and compared hand labelled lip-corner positions with automatically tracked lip-corner positions. In these sequences the camera image contained not only the users face but also some background (our lab) and the user moved and rotated his head a lot, but did not speak. Figure 6 shows some sample images of this set. Average location error for the lip-corners on these sequences was 3 pixel in x-direction and 2 pixel in y-direction as shown in table 1.

In addition, to obtain new training data for our current audio-visual speech recognition system [12], we ran the lip-tracking system on all sequences of a database of ten different male and female speakers, which were spelling words. The database contains ninehundred greyscale image sequences with a total of around 80.000 frames. In these images faces of the speakers covered the whole images. Because these images were only available in greyscale, we didn't apply the search for the face (based on color) but set the facial area manually to the whole image for all images and did the search for facial features on the complete images. Figures 7 show some sample images of this database. We labelled the lip-corners in ten of these sequences by hand and compared it to the positions obtained by the lip-tracking system. Average location error on these sequences was 4.3 pixel in x-direction and 1.9 pixel in y-direction (see table 1).

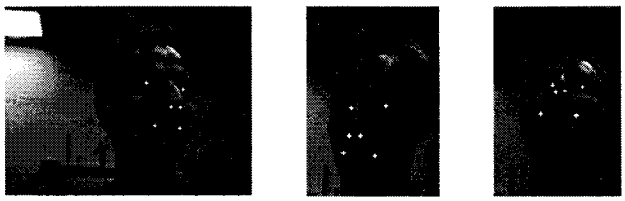

Figure 6. Samples from set 1 (see text)

Figure 7. Samples from set 2 (see text)

	# frames	avg. face size	$Error_x$	$Error_y$
set 1	950	110x140	3.2	2.1
set 2	839	256x256	4.3	1.9

Table 1. Average vertical and horizontal lip-corner localization errors in pixel (see text)

These results show, that we can track lip-corners in images with different resolution of the lip-area and with different illumination accurately. Even when the speaker is moving or rotates his head, the lip-corners can successfully be tracked. On camera images of size 160x120 pixel as used for our current speech-reading system [12] the lip-tracker runs at frame rates of around 25 frames per second.

In addition, we have examined the number of frames where tracking failure occurred in the test sequences in set 1 (sequences with head movement). The results are shown in table 2. It can be seen, that by trying to detect outliers in the set of found features and predicting their true position in the camera image, as described in section 3., the number of frames with tracking failure can be reduced by almost a factor of two. This clearly shows the usefulness of the technique in order to enhance robustness of feature tracking.

method	# frames w. failure	relative
no outlier-detection	79	15 %
outlier-detection	140	8 %

Table 2. Number of frames with tracking failure on set 1

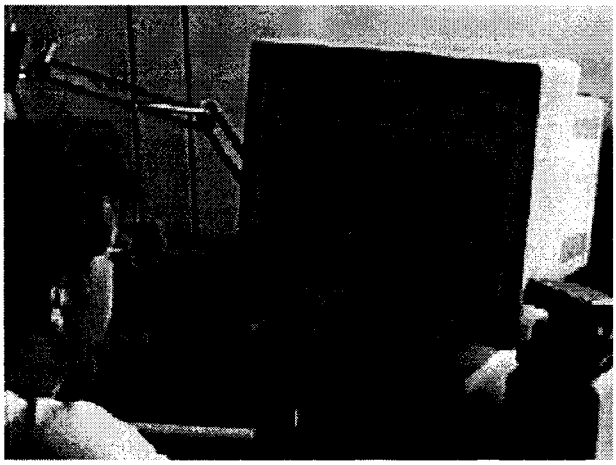

Figure 8. The setup of the lipreading system

5. CONCLUSION

We have presented a new approach to track lip-corners in real-time in a robust way. The basic idea of the new approach is that not only the lips but also other facial features such as eyes and nostrils are tracked, which enables the system to detect outliers and recover from failures using a 3D model. We have implemented the new approach in the lip-tracking module of our lipreading system. The new lip-tracker can extract the lip regions of a speaker in a sequence of images from a video camera for our lipreading system in a non-intrusive way. Using the lip-tracker, the lip-corners of the speaker are located and tracked as soon as he/she appears in the view of the camera. The system has been tested on a database that contains 900 image sequences of different speakers spelling words. The system has successfully extracted the lip regions from the images to obtain training data for our lipreading system. We have also evaluated the robustness of the new approach. The experiment has shown that the number of frames with tracking failures can be reduced by a factor of two using outlier detecting and predicting technique for the test sequences. We are currently working on evaluating performance of our lipreading system with the new lip-tracker.

ACKNOWLEDGEMENTS

This research was sponsored by the Advanced Research Projects Agency under the Department of the Navy, Naval Research Office under grant number N00014-93-1-0806 and by the state of Baden-Württemberg, Germany (Landesschwerpunkt Neuroinformatik). The views and conclusions stated in this paper are those of the authors.

REFERENCES

[1] K. Mase and Alex Pentland. Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22(6):67 – 76, 1991.

[2] E.D. Petajan. Automatic lipreading to enhance speech recognition. In *Proceedings of IEEE Communications Society Global Telecom. Conference*, November 1984.

[3] D. G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *Proceedings of IJCNN*, 1992.

[4] Paul Duchnowski, Uwe Meier, and Alex Waibel. See me, hear me: Integrating automatic speech recognition and lipreading. In *Proceedings of ICSLP*, 1994.

[5] Paul Duchnowski, Martin Hunke, Dietrich Büsching, and Uwe Meier. Toward movement-invariant automatic lip-reading and speech recognition. In *Proceedings of International Conf. on Acoustics, Speech, and Signal Processing*, 1995.

[6] Uwe Meier, Wolfgang Hürst, and Paul Duchnowski. Adaptive bimodal sensor fusion for automatic speechreading. In *Proceedings of International Conf. on Acoustics, Speech, and Signal Processing*, 1996.

[7] Jie Yang and Alex Waibel. A real-time face tracker. In *Proceedings of WACV*, pages 142–147, 1996.

[8] Takeo Kanade. Picture processing by computer complex and recognition of human faces. Technical report, Kyoto Univ., Dept. Inform. Sci., 1973.

[9] Daniel F. DeMenthon and Larry S. Davis. Model based object pose in 25 lines of code. In *Proceedings of Second European Conference on Computer Vision*, pages 335 – 343. Springer Verlag, May 1992.

[10] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. A model-based gaze tracking system. In *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, pages 304 – 310, 1996.

[11] Andrew H. Gee and Roberto Cipolla. Fast visual tracking by temporal consensus. Technical Report CUED/F-INFENG/TR-207, University of Cambridge, February 1995.

[12] Uwe Meier, Rainer Stiefelhagen, and Jie Yang. Preprocessing of visual speech under real world conditions. In *Proceedings of European Tutorial & Research Workshop on Audio-Visual Speech Processing*, 1997.