Hough Transform-based Mouth Localization for Audio-Visual Speech Recognition

Gabriele Fanelli[1]
fanelli@vision.ee.ethz.ch
Juergen Gall[1]
gall@vision.ee.ethz.ch
Luc Van Gool[1,2]
vangool@vision.ee.ethz.ch

[1] Computer Vision Laboratory
ETH Zürich, Switzerland
[2] IBBT, ESAT-PSI
K.U.Leuven, Belgium

Abstract

We present a novel method for mouth localization in the context of multimodal speech recognition where audio and visual cues are fused to improve the speech recognition accuracy. While facial feature points like mouth corners or lip contours are commonly used to estimate at least scale, position, and orientation of the mouth, we propose a Hough transform-based method. Instead of relying on a predefined sparse subset of mouth features, it casts probabilistic votes for the mouth center from several patches in the neighborhood and accumulates the votes in a Hough image. This makes the localization more robust as it does not rely on the detection of a single feature. In addition, we exploit the different shape properties of eyes and mouth in order to localize the mouth more efficiently. Using the rotation invariant representation of the iris, scale and orientation can be efficiently inferred from the localized eye positions. The superior accuracy of our method and quantitative improvements for audio-visual speech recognition over monomodal approaches are demonstrated on two datasets.

1 Introduction

Speech is one of the most natural forms of communication and the benefits of speech-driven user interfaces have been advocated in the field of human computer interaction for several years. Automatic speech recognition, however, suffers from noise on the audio signal, unavoidable in application-relevant environments. In multimodal approaches, the audio stream is augmented by additional sensory information to improve the recognition accuracy [22]. In particular, the fusion of audio and visual cues [19] is motivated by human perception, as it has been proven that we use both audio and visual information when understanding speech [16]. There are indeed sounds which are very similar in the audio modality, but easy to discriminate visually, and vice versa. Using both cues significantly increases automatic speech recognition performance, especially when the audio is corrupted by noise.

To extract the visual features, a region-of-interest [18], a set of feature points [27], or lip contours [14], need to be localized. Although the lip contours contain more information about the mouth shape than the appearance within a bounding box, they do not necessary encode more information valuable for speech recognition, as demonstrated in [21]. In addition, extracting a bounding box is usually more robust and efficient than lip tracking approaches.

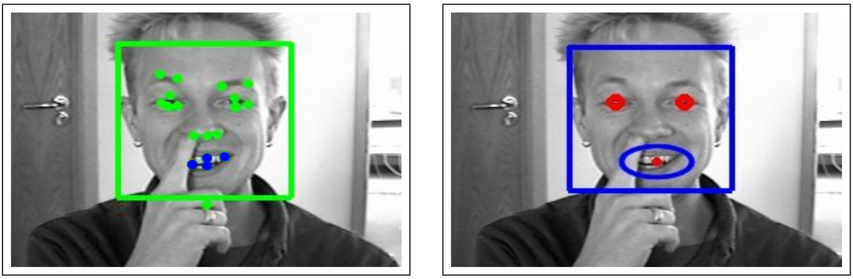

(a) (b)

Figure 1: a) Facial points like mouth corners *(blue dots)* are sensitive to occlusions. b) Our Hough transform-based approach localizes the center of the mouth *(red dot)* even in the case of partial occlusions. The *ellipse* indicates the region of interest for speech recognition.

While standard approaches extract mouth corners to estimate scale, position, and orientation of the mouth, we propose a Hough transform-based method for mouth localization. A certain feature point or patch might be difficult to detect due to occlusions, lighting conditions, or facial hair, therefore our method accumulates the votes of a set of patches into a Hough image where the peak is considered to be the mouth center. This facilitates the localization of the mouth even when a facial feature like a lip corner cannot be detected, as shown in Figure 1. To make the process faster, we exploit the different shape properties of eyes and mouth: a) being the shape of the iris unique and rotation invariant, it can be very efficiently localized using isophote curvature [25]. b) Knowing the approximate orientation and scale of the face from the eye centers, the various shapes of the mouth can be learned using randomized Hough trees [6]. Without the eye detection, scale and orientation would have to be handled by the mouth detector, yielding higher computational cost.

2 Related Work

Audio-visual speech recognition (AVSR) has been pioneered by Petajan [19] and it is still an active area of research. Most approaches focus on the mouth region as it encodes enough information for deaf persons to achieve a reasonable speech perception [24]. In order to fuse audio and visual cues, we employ the commonly used multi-stream hidden Markov models (MSHMM) [29], but other approaches could be used, based for example on artificial neural networks [11], support vector machines [7], or AdaBoost [28].

As visual features, lip contours [14], optical flow [9], and image compression techniques like linear discriminant analysis (LDA), principal component analysis (PCA), discrete cosine transform (DCT), or discrete wavelet transform (DWT) [22], have been proposed. Within monomodal speech recognition (lip reading), snakes [3] and active shape models [12, 15] have been intensively studied for lip tracking. Most of these approaches assume that a normalized mouth region can be reliably extracted, which is addressed in this work. Lip contours-based methods do not encode all possible geometric information (like the tongue), therefore space-time volume features have been proposed for lip-reading in [17]. In order to build these macro-cuboïd features, it is again necessary to reliably extract the mouth regions.

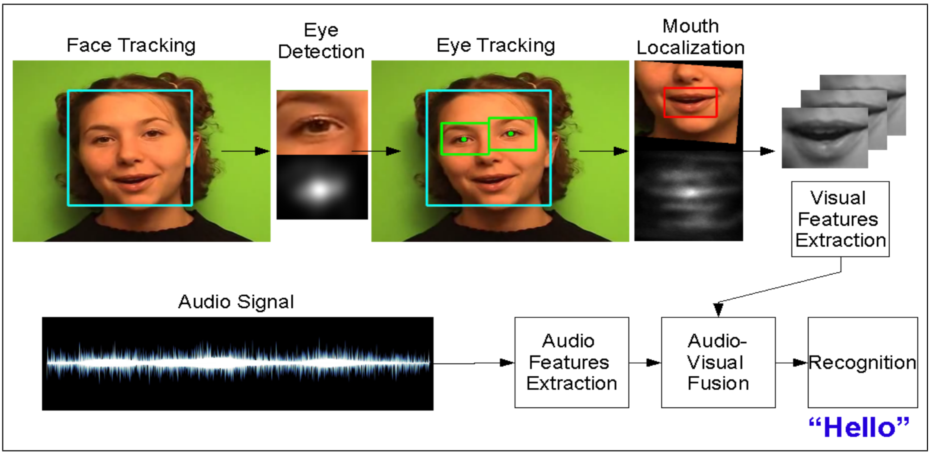


Figure 2: Overview of our AVSR system. The visual pipeline is shown at the top: face tracking, eye detection and tracking, mouth localization on images scaled and rotated according to the eye positions. At the bottom right, the features extracted from the stream of normalized mouth images and from the audio signal are fused allowing the actual speech recognition.

3 Overview

The pipeline of our AVSR system is depicted in Figure 2. The first necessary step is face detection, where we use the algorithm proposed by Viola and Jones [26]. To cope with appearance changes, partial occlusion, and multiple faces, we employ an online-boosting tracker [8] that uses the currently tracked patch and its surroundings respectively as positive and negative samples for updating the internal classifier. Assuming the face to be nearly frontal, the bounding box returned by the tracker allows us to estimate the rough positions of the eyes by using anthropometric relations. The scale and in-plane rotation of the face are then estimated by filtering the positions of the detected irises (Section 4.1). With this information at hand, we crop the lower part of the face image and normalize it such that the mouth is horizontal and has a specific size, thus being able to run the mouth detection (Section 4.2) at only one scale and rotation, speeding up drastically the computation time. Finally, features are extracted from the stream of normalized mouth images and from the audio signal in order to recognize the spoken words (Section 5).

4 Normalized Mouth Region Extraction

4.1 Eye Localization

We use the method of Valenti *et. al.* [25] for accurate eye center localization, based on isophote curvature. The main idea relies on the radial symmetry and high curvature of the eyes' brightness patterns. An isophote is a curve going through points of equal intensity, its shape being invariant to rotations and linear changes in the lighting conditions.

For each point p in the image, a displacement vector is computed as:

$$D(x,y) = -\frac{{L_x}^2 + {L_y}^2}{{L_y}^2 L_{xx} - 2L_x L_{xy} L_y + {L_x}^2 L_{yy}} (L_x, L_y), \tag{1}$$

where L_x and L_y are the image derivatives along the x and y axes, respectively. The value of an accumulator image at the candidate center $c = p + D$ is incremented by the curvedness of p in the original image, computed as $\sqrt{{L_{xx}}^2 + 2{L_{xy}}^2 + {L_{yy}}^2}$. In this way, center candidates coming from highly curved isophotes are given higher weights. Knowing that the pupil and the iris are generally darker than the neighboring areas, only transitions from bright to dark areas are considered, *i.e.*, situations where the denominator of equation (1) is negative. The eye center is finally located by mean shift.

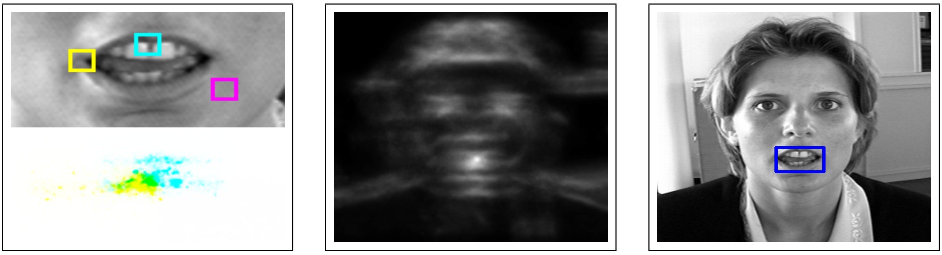
The above method fails when the iris is not visible, *e.g.*, due to closed eyes or strong reflections on glasses. When tracking a video sequence, this can lead to sudden jumps of the detections. Such errors propagate through the whole pipeline, leading to wrong estimates of the mouth scale and rotation, and eventually worsening the overall AVSR performance. To reduce these errors, we smooth the pupils' trajectories using two Kalman filters, one for each eye center. The prediction of the eye position for the incoming frame is used as the center of the new region-of-interest for the pupil detection.

4.2 Mouth Localization

Hough transform-based methods model the shape of an object implicitly, gathering the spatial information from a large set of object patches. Thanks to the combination of patches observed on different training examples, large shape and appearance variations can be handled, as it is needed in the case of the mouth, greatly changing its appearance between the states open and closed. Furthermore, the additive nature of the Hough transform makes these approaches robust to partial occlusions. For localization, the position and the discriminative appearance of a patch are learned and used to cast probabilistic votes for the object center as illustrated in Figure 3 a). The votes from all image patches are summed up into a Hough image (Figure 3 b), where the peak is used to localize the mouth region (Figure 3 c). The whole localization process can thus be described as a generalized Hough transform [2]. The so-called implicit shape model (ISM) can be modeled either by an explicit codebook as in [13] or within a random forest framework [6]. An approach similar to [13] was employed for facial feature localization in [5]. Since the construction of codebooks is expensive due to the required clustering techniques and the linear matching complexity, we follow the random forest approach where learning and matching are less computationally demanding.

A random forest consists of several randomized trees [1, 4] where each node except for the leaves is assigned a binary test that decides if a patch is passed to the left or right branch. Random forests are trained in a supervised way, and the trees are constructed assigning each leaf the information about the set of training samples reaching it, *e.g.*, the class distribution for classification tasks. At runtime, a test sample visits all the trees and the output is computed by averaging the distributions recorded during training at the reached leaf nodes.

Learning Each tree in the forest is built based on a set of patches $\{(\mathscr{I}_i, c_i, \mathbf{d}_i)\}$, where $\mathscr{I}_i$ is the appearance of the patch, c_i the class label, and $\mathbf{d}_i$ the relative position with respect to the mouth center, computed from the annotated positions of the lip corners and outer lips' midpoints. For mouth localization, we use patches of size 16×16 (Figure 3 a) where the



(a) (b) (c)

Figure 3: a) For each of the emphasized patches *(top)*, votes are cast for the mouth center *(bottom)*. While lips *(yellow)* and teeth *(cyan)* provide valuable information, the skin patch *(magenta)* casts votes with a very low probability. b) Hough image after accumulating the votes of all image patches. c) The mouth is localized by the maximum in the Hough image.

appearance $\mathscr{I}$ is modeled by several feature channels I^f, which can include raw intensities, derivative filter responses, etc. The training patches are randomly sampled from mouth regions (positive examples) and non-mouth regions (negative examples), where the images are normalized according to scale and orientation. The samples are annotated with the binary class label $c \in \{p, n\}$ and the center of the mouth in the case of positive examples.

Each tree is constructed recursively starting from the root. For each non-leaf node, an optimal binary test is selected from a set of random tests evaluated on the training patches that reach that node. The selected test splits the received patches into two new subsets which are passed to the children. The binary tests $t(\mathscr{I}) \rightarrow \{0,1\}$ compare the difference of channel values I^f for a pair of pixels (p,q) and (r,s) with some handicap τ:

$$t_{f,p,q,r,s,\tau}(\mathscr{I}) = \begin{cases} 0, & \text{if } I^f(p,q) - I^f(r,s) < \tau \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

A leaf is created when the maximal depth of the tree, *e.g.* 15, or the minimal size of a subset, *e.g.* 20, are reached. Each leaf node L stores information about the patches that have reached it, *i.e.*, the probability $p_{mouth}(\mathscr{I})$ of belonging to a mouth image (the proportion of positive patches that have reached the leaf) and the list $D_L = \{\mathbf{d}_i\}$ of corresponding offset vectors. The leaves thus build an implicit codebook and model the spatial probability of the mouth center $\mathbf{x}$ for an image patch $\mathscr{I}$ located at position $\mathbf{y}$, denoted by $p\big(\mathbf{x}|\,\mathscr{I}(\mathbf{y})\big)$. Such probability is represented by a non-parametric density estimator computed over the set of positive samples D_L and by the probability that the image patch belongs to the mouth:

$$p\big(\mathbf{x}|\,\mathscr{I}(\mathbf{y})\big) = \frac{1}{Z} p_{mouth}(\mathscr{I}) \left(\frac{1}{|D_L|} \sum_{\mathbf{d} \in D_L} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{||(\mathbf{y}-\mathbf{x})-\mathbf{d}||^2}{2\sigma^2} \right) \right), \tag{3}$$

where $\sigma^2 \mathbf{I}_{2\times 2}$ is the covariance of the Gaussian Parzen window and Z is a normalization constant. The probabilities for three patches are illustrated in Figure 3 a).

Since the quantity in (3) is the product of a class and a spatial probability, the binary tests need to be evaluated according to class-label uncertainty U_c and spatial uncertainty U_s. We use the measures proposed in [6]:

$$U_c(A) = |A| \cdot Entropy(\{c_i\}) \quad \text{and} \quad U_s(A) = \sum_{i:c_i=p} (\mathbf{d}_i - \bar{\mathbf{d}})^2, \tag{4}$$

where $A = \{(\mathscr{I}_i, c_i, \mathbf{d}_i)\}$ is the set of patches that reaches the node and $\bar{\mathbf{d}}$ is the mean of the spatial vectors $\mathbf{d}_i$ over all positive patches in the set[1]. For each node, one of the two measures is randomly selected with equal probability to ensure that the leaves have both low class and spatial uncertainty. The optimal binary test is selected from the set of randomly generated tests $t^k(\mathscr{I})$ by

$$\underset{k}{\operatorname{argmin}} \left(U_\star(\{A_i | t^k(\mathscr{I}_i){=}0\}) + U_\star(\{A_i | t^k(\mathscr{I}_i){=}1\}) \right) \tag{5}$$

where $\star = c$ or s, *i.e.*, by the quality of the split.

Localization In order to localize the mouth in an image, each patch $\mathscr{I}(\mathbf{y})$ goes through all the trees in the forest $\{\mathscr{T}_t\}_{t=1}^T$, ending up in one leaf, where the measure (3) is evaluated. The probabilities are averaged over the whole forest [1, 4]:

$$p\big(\mathbf{x} | \mathscr{I}(\mathbf{y}); \{\mathscr{T}_t\}_{t=1}^T\big) = \frac{1}{T} \sum_{t=1}^{T} p\big(\mathbf{x} | \mathscr{I}(\mathbf{y}); \mathscr{T}_t\big). \tag{6}$$

These probabilistic votes are then accumulated in a 2D Hough image, see Figure 3 b)[2]. The location where the generalized Hough transform gives the strongest response is considered to be the center of the mouth (Figure 3 c).

5 Audio-Visual Speech Recognition

In order to fuse the audio and visual cues for speech recognition, we rely on the commonly used multi-stream hidden Markov models [29]. Each modality s is described by Gaussian mixtures, *i.e.*, the joint probability of the multimodal observations $O = (o_1, \cdots, o_t)$ and the states $Q = (q_1, \cdots, q_t)$ is given by

$$p(O, Q) = \prod_{q_i} b_{q_i}(o_i) \prod_{(q_i, q_j)} a_{q_i q_j} \quad \text{where} \quad b_j(o) = \prod_{s=1}^{2} \left(\sum_{m=1}^{M_s} c_{j_{s,m}} N(o_s; \mu_{j_{s,m}}, \Sigma_{j_{s,m}}) \right)^{\lambda_s}, \tag{7}$$

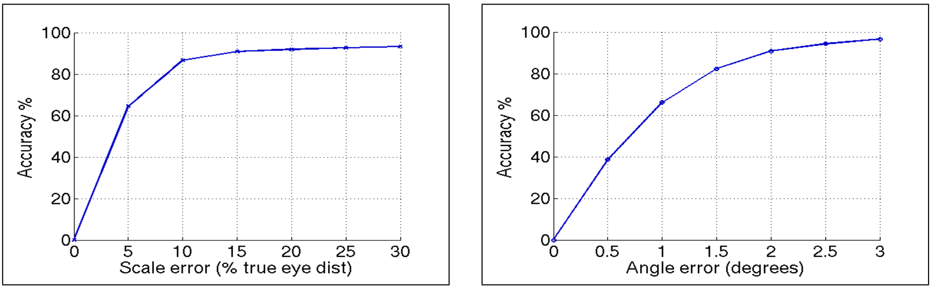
where $a_{q_i q_j}$ are the transition probabilities, $N(o; \mu, \Sigma)$ are multi-variate Gaussians with mean μ and covariance Σ, and $c_{j_{s,m}}$ are the weights of the Gaussians. The model parameters are learned for each modality independently. The parameters $\lambda_s \in [0,1]$ control the influence of the two modalities with $\lambda_1 + \lambda_2 = 1$. As cues, we extract mel-frequency cepstral coefficients from the audio stream and DCT features from the normalized mouth images where only the odd columns are used due to symmetry [20, 22]. For both features, the first and second temporal derivatives are added and the sets normalized as to have zero mean.

6 Experiments

We evaluate our system testing each component separately. First we assess the quality of the scale and orientation estimated from the eye detection method, then we move on to the mouth localization accuracy and compare our results with a state-of-the-art method for facial feature points detection, finally we show the applicability of our system for an AVSR task.

[1] $Entropy(\{c_i\}) = -\bar{c} \log \bar{c} - (1 - \bar{c}) \log(1 - \bar{c})$, where $\bar{c} = |\{c_i | c_i = p\}| / |\{c_i\}|$.

[2] In practice, we go through each image location y, pass the patches $\mathscr{I}(\mathbf{y})$ through the trees, and add the discrete votes $p_{mouth}(\mathscr{I})/|D_L|$ to the pixels $\{(\mathbf{y} - \mathbf{d} | \mathbf{d} \in D_L\}$ for each tree. The Gaussian kernel is then applied after voting.



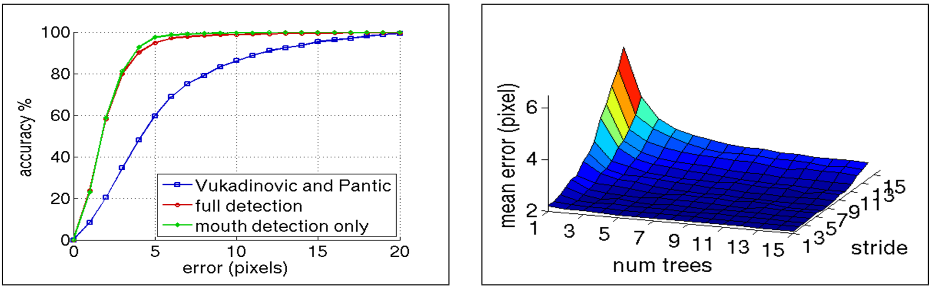
(a) (b)

Figure 4: a) Accuracy vs. eye distance error (scale). b) Accuracy vs. angle error (rotation). The plots show the percentage of correctly estimated images as the error threshold increases.

Estimation of Scale and Orientation We run our tests on the BioID face database [23], composed of 1521 greyscale images of 23 individuals, acquired at several points in time with uncontrolled illumination, and at the resolution of 384x288 pixels. Subjects were often photographed with their eyes closed, showing different facial expressions, and many of them wore glasses. Manually annotated ground truth is provided for the pupils and for 18 other facial points. We divide the database in four sets, training the mouth detector on three, testing on the fourth, and averaging the results of all combinations.

As first experiment, we run the whole pipeline: we detect the face in each image (taking the largest in case of multiple detections), then we detect the eyes in the two upper quarters of the face rectangle and compute the errors for the eye distance (scale) and the angle formed by the line connecting the eyes and the horizontal axis (rotation). Figure 4 shows the accuracy for the two measures, *i.e.*, the percentage of correct estimations as the error threshold increases. In 4 a) the accuracy is plotted against the error between the detected eye distance $dEye$ and the ground truth dGT, as $err = \frac{abs(dEye-dGT)}{dGT}$. In 4 b) the accuracy is plotted against the error of the estimated angle in degrees. It is worth noting that, for 17 images (1.12% of the total), no face was detected at all; we do not consider those for the analysis. Moreover, sometimes the face detector gave wrong results, getting stuck on some clutter in the background; this partly explains the curve in Figure 4 a) never reaching 100%.

Mouth Localization Using again the BioID database, we evaluate the accuracy of the mouth detection and compare our results to the output of the facial points detector (FPD) of Vukadinovic and Pantic [27], for which the code is made available. As we localize the mouth center rather than the corners, we compute the center from the four mouth corners provided by the ground truth and the FPD. As already mentioned, face detection does not always succeed; indeed the FPD failed in 9.67% of the cases. We only take into account images where both methods detect a face, however, there are still some false detections which increase the error variance. In order to decrease the influence of errors originated in the eye detection part, we perform a second test concentrating on the mouth localization only, using the ground truth of the eye positions. As the curve in Figure 5 a) shows, our method outperforms the FPD for the mouth localization task, both in the "full detection" (face, eyes, mouth), and "mouth only" type of experiment. Figure 6 shows some sample results; the successes in the first row indicate that the full pipeline can cope with difficult situations like the presence of glasses, facial hair, and head rotations, however, failures do occur, as shown in the second row. We also run the "mouth only" test varying two parameters of the Hough-



(a) (b)

Figure 5: a) Accuracy vs. mouth center localization error (in pixels) between Facial Point Detector [27] (blue), our full pipeline (red), and the mouth localization given the eye position from ground truth. b) Mouth localization error in pixels vs. stride and number of trees.

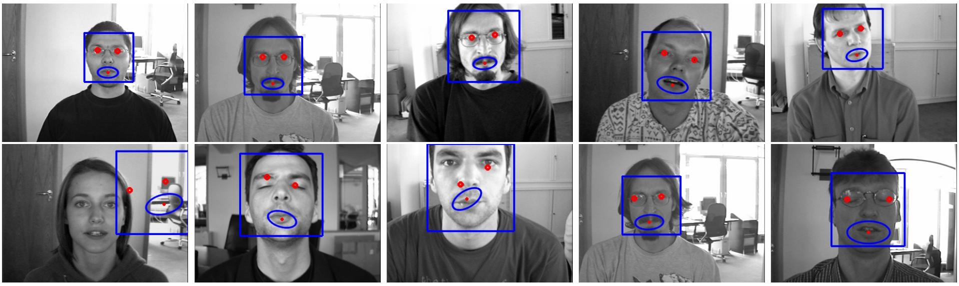
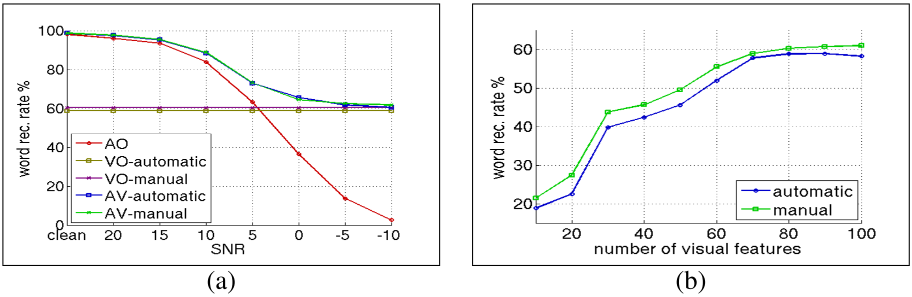


Figure 6: Some examples of successes (top row) and failures (bottom row) of the system.

based detector: the stride and the number of trees; the results in Figure 5 b) show that the mean error (in pixel) remains low (around 2) even for a large stride and few trees.

Speech Recognition As the goal of our system is to automatically provide mouth images for AVSR purposes, we test it on the CUAVE database [18], consisting of videos recorded in controlled audio-video conditions, at 29.97fps interlaced, with a resolution of 740x480. Each of the 36 subjects repeats the digits from "zero" to "nine" in American English. We concentrate on the subset of the database where subjects appear alone, keeping the face nearly frontal, and use a mouth detector trained on the BioID database. The CUAVE videos are deinterlaced and linearly interpolated to match the frequency of the audio samples (100Hz). The power of AVSR is clear when the audio channel is unreliable, we therefore add white noise to the audio stream. We train on clean audio and test at different levels of Signal to Noise Ratios. To run the speech recognition experiments, we use the system of [10], without the automatic feature selection part. For the audio-visual fusion, we keep the audio and video weights λ_1 and λ_2 fixed for each test, and run several trials varying the weights from 0.00 to 1.00 in 0.05 steps, at the end we pick the combination giving the best recognition rate for each SNR. The accuracy is defined as the number of correctly recognized words, C, minus the number of insertions, I (false positives detected during silence), divided by the number of words, N [29]. We split the 36 sequences in 6 sets and perform cross-validation by training on 5 groups while testing on the sixth and averaging the results of all combinations. Figure 7 a) shows the performance for a fixed number of visual features (80), at several SNR levels. We compare to the results obtained from manually extracted mouth-regions, which give the upper bound for the accuracy obtained with automatic extraction. The multimodal approaches always outperform the monomodal ones, moreover, our automatic method for mouth ROI extraction performs only slightly worse than the manual one. In Figure 7 b), we show the accuracy of the recognizer when only video features are used as their number increases: our approach performs best with 80 visual features (58.85%), while for greater sets the performance decreases slightly.



(a) (b)

Figure 7: a) Word recognition rate for the audio-visual system evaluated with 80 visual features at different noise levels, for automatically and manually extracted mouth images, compared to monomodal results. b) Influence of the number of features in video-only speech recognition.

Processing Speed When analyzing videos on a 2.8 GHz machine, the presented system (implemented in C++ without optimization efforts) runs at about 4fps. Most of the computation is concentrated in the mouth localization part, indeed the face plus eyes tracking parts together run at 53fps. A sensible decrease in processing time with a low price in accuracy can be achieved by loading a smaller number of trees and introducing a stride: for 10 trees and a stride of 4, we achieve 15fps.

7 Conclusion

We have presented a novel and efficient method for mouth localization which provides the accuracy needed for audio-visual speech recognition (ASVR). Our experiments show that it outperforms a state-of-the-art facial points detector and that the achieved word recognition rate for ASVR is near to the boundary obtained by employing manually cropped mouth regions. In order to achieve nearly real-time mouth localization, scale and orientation of the face are estimated from filtered irises' detections. A further speed-up with a small price in accuracy can be achieved by reducing the number of trees and sampling rate by introducing a stride. The proposed method is not only relevant for AVSR but also for lip reading and facial expression recognition where a normalized region-of-interest is usually required. The approach is independent of the employed recognition system as it does not necessarily have to be coupled with multi-stream hidden Markov models.

Acknowledgments The authors gratefully acknowledge support by Swiss SNF NCCR project IM2 and EU project HERMES (FP6-027110). Furthermore, we would like to thank Dr. Mihai Gurban for providing source code for speech recognition.

References

[1] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

[2] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[3] C. Bregler and S. Omohundro. Nonlinear manifold learning for visual speech recognition. In *International Conference on Computer Vision*, pages 494–499, 1995.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *British Machine Vision Conference, London, England*, pages 277–286, 2004.

[6] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[7] M. Gordan, C. Kotropoulos, and I. Pitas. A support vector machine-based dynamic network for visual speech recognition applications. *EURASIP J. Appl. Signal Process.*, 2002(1):1248–1259, 2002.

[8] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *British Machine Vision Conference*, volume 1, pages 47–56, 2006.

[9] M. S. Gray, J. R. Movellan, and T. J. Sejnowski. Dynamic features for visual speechreading: A systematic comparison. In *Neural Information Processing System Conference (NIPS)*, pages 751–757, 1996.

[10] M. Gurban and J. P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. *IEEE Transactions on Signal Processing*, 2009.

[11] M. Heckmann, F. Berthommier, and K. Kroschel. A hybrid ann/hmm audio-visual speech recognition system. In *International Conference on Auditory-Visual Speech Processing*, 2001.

[12] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *European Conference on Computer Vision*, pages 376–387, 1996.

[13] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008.

[14] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[15] I Matthews, J. A. Bangham, R. Harvey, and S. Cox. A comparison of active shape model and scale decomposition based features for visual speech recognition. In *European Conference on Computer Vision*, pages 514–528, 1998.

[16] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[17] S. Pachoud, S. Gong, and A. Cavallaro. Macro-cuboids based probabilistic matching for lip-reading digits. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy. Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus. *EURASIP J. Appl. Signal Process.*, 2002(1):1189–1201, 2002. ISSN 1110-8657.

[19] E.D. Petajan. Automatic lipreading to enhance speech recognition. In *IEEE Communication Society Global Telecommunications Conference*, 1984.

[20] G. Potamianos and P. Scanlon. Exploiting lower face symmetry in appearance-based automatic speechreading. In *Audio-Visual Speech Process.*, pages 79–84, 2005.

[21] G. Potamianos, H. P. Graf, and E. Cosatto. An image transform approach for hmm based automatic lipreading. In *International Conference on Image Processing*, pages 173–177, 1998.

[22] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. *Issues in Visual and Audio-Visual Speech Processing*, chapter Audio-Visual Automatic Speech Recognition: An Overview. MIT Press, 2004.

[23] BioID Technology Research, 2001. http://www.bioid.de/.

[24] Q. Summerfield. Lipreading and audio-visual speech perception. In *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, volume 335, pages 71–78, 1992.

[25] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[27] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 1692– 1698, 2005.

[28] P. Yin, I. Essa, and J. M. Rehg. Asymmetrically boosted hmm for speech reading. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 755–761, 2004.

[29] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic Ltd., Cambridge, 1999.