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

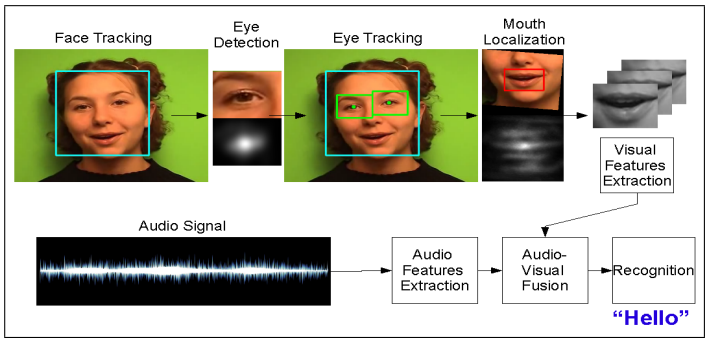


Figure 1: Overview of our AVSR system. The visual pipeline consists of: face tracking, eye detection and tracking, mouth localization on images scaled and rotated according to eye locations. At the bottom right, the features extracted from the stream of normalized mouth images and from the audio signal are fused allowing the actual speech recognition.

Speech is one of the most natural forms of human communication, which makes reliable speech-driven user interfaces very desirable. Speech recognizers based on audio cues only are sensitive to noise, therefore audio-visual speech recognition (AVSR) systems have become popular.

In this paper we propose a Hough transform-based method for mouth localization, in contrast to feature points-based approaches which can be more easily affected by occlusions, lighting conditions, or facial hair. We prefer to extract a normalized bounding box around the mouth rather than to track the lip contours, as appearance features perform better for speech recognition tasks.

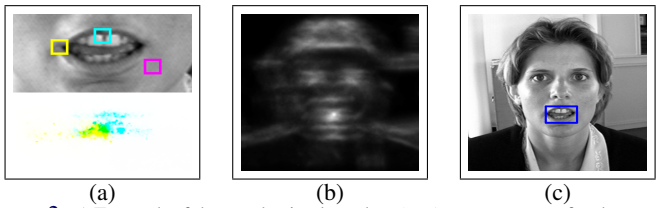
Figure 1 shows our system. After the face is detected, we track it using a method based on online boosting. To localize the eyes, we select two regions specified by anthropometric relations and apply the method based on isophote curvature of [3], exploiting the iris' radial symmetry and high curvature. To tackle blinking, we smooth the pupils' trajectories using Kalman filters. Knowing the eyes' location tells us scale and rotation of the mouth, allowing us to run the mouth localization at one scale and orientation only. Fusing the features coming from the stream of normalized mouth images and the audio signal finally allows word recognition.

In order to localize the mouth in an image, we use a Hough transform-based method, which models the shape of an object implicitly, gathering the spatial information from a large set of patches. Such method can handle large shape and appearance variations and is robust to partial occlusions. The position and the discriminative appearance of a patch are learned and used to cast votes for the object center as illustrated in Fig. 2 a). The votes from all patches are summed into a Hough image (Fig. 2 b), and the peak is taken as the mouth center (Fig. 2 c). We model such implicit shape model as a random forest [2], for which learning and matching are not too computationally demanding.

A random forest consists of many randomized trees [1] where each non-leaf node decides by a binary test to which branch to pass a patch. Each leaf stores information about the patches that reached it during training: the probability of belonging to a mouth, $p_m(\mathscr{I})$ (ratio of mouth patches), and the list $D_L = \{\mathbf{d}_i\}$ of offset vectors (distance to the mouth center). The leafs build an implicit codebook and model the spatial probability of the mouth center $\mathbf{x}$ for a patch $\mathscr{I}$ located at position $\mathbf{y}$, as:

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

where $\sigma^2 \mathbf{I}_{2\times 2}$ is the covariance of the Gaussian Parzen window and Z is a normalizing constant. Sample probabilities are shown in Fig. 2 a).



(a) (b) (c)

Figure 2: 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 peak in the Hough image.

Figure 3: Some mouth localization results.

The binary tests in the trees need to be evaluated according to both a class-label uncertainty U_c and a spatial uncertainty U_s:

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

where A is the set of patches that reaches the node and $\bar{\mathbf{d}}$ is the mean of the offsets $\mathbf{d}_i$ over all positive patches in the set. 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.

During search, each patch $\mathscr{I}(\mathbf{y})$ goes through all the trees in the forest, reaching one leaf. The values of $p(\mathbf{x}|\mathscr{I}(\mathbf{y}))$ are averaged over the whole forest. These votes are summed in a Hough image, whereof the maximum is the mouth center.

For speech recognition, we extract mel-frequency cepstral coefficients from the audio and DCT features from the stream of normalized mouth images. For both, first and second temporal derivatives are added and the sets normalized to have zero mean. We use multi-stream hidden Markov models for the fusion: each modality s is described by Gaussian mixtures, *i.e.*, the joint probability $p(O,Q)$ of the observations O and the states Q is given by $\prod_{q_i} b_{q_i}(o_i) \prod_{(q_i,q_j)} a_{q_i q_j}$, where

$$b_j(o) = \prod_{s=1}^{2} \left(\sum_{m=1}^{M_s} c_{j_{s,m}} N(o_s; \mu_{j_{s,m}}, \Sigma_{j_{s,m}}) \right)^{\lambda_s}, \quad (3)$$

$a_{q_i q_j}$ are the transition probabilities, and $N(o;\mu,\Sigma)$ are multi-variate Gaussians weighted by $c_{j_{s,m}}$. The model parameters are learned for each modality independently. The parameters $\lambda_s \in [0,1]$ control the influence of the two modalities with $\lambda_1 + \lambda_2 = 1$.

In the paper we extensively evaluate the system at several levels: scale and orientation estimation, mouth localization (see examples in Fig. 3), and speech recognition. Our system outperforms a state-of-the-art facial features detector for the mouth center localization task, and is comparable to manually annotated mouth regions for the speech-recognition task.

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[2] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, 2009.
[3] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *CVPR*, 2008.