Mouth Region Localization based on Gabor Features and Active Appearance Models

Luis R. Sapaico∗
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
Hamid Laga†
Global Edge Institute
Tokyo Institute of Technology
Masayuki Nakajima‡
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Abstract

We propose a combined knowledge, feature and appearance-based method for accurate localization of the mouth region from a face image obtained using a web camera. First, using a knowledge-based approach based on geometrical properties of a regular face, we prune the mouth search region by extracting a rough segment where the mouth should be located in the given face. Next, assuming the mouth is closed, we extract near-to-horizontal features that resemble those of the dark shadow line observed in the intersection between the upper and lower lip. We use for this a set of Gabor feature-based filters, with specifically oriented and scaled parameters that are sufficient for our purpose. We then find strong features in the resultant Gabor set of images, and we apply the Hough Transform on the fused binary image in order to obtain the near-to horizontal line approximating the intersection between the upper and lower lip. Next, we use an Active Appearance Model (AAM) to extract an accurate mouth region. For this purpose, we have built a database of mouths containing 365 images of 20 different users, and we have used this dataset for training a 15-point mouth model. The extremes of the detected line are used as two of the model initialization points and after we apply a AAM fitting process, we can finally extract a very accurate region that contains the closed mouth. Our results show that the proposed method is able to detect a line in all of the images of our database. The accuracy of the mouth corners localization is approximately 95%. We also show that the detection of these points proves to be critical for the output of the AAM fitting.

Keywords: mouth segmentation, gabor wavelets, active appearance models

1 Introduction

Face Feature Detection is a problem for which many solutions have been proposed [Hjelmars and Low 2001]. The literature is so vast that it is highly unlikely that a problem has not been solved yet. However, frequently the solutions given are ad-hoc, i.e., they do not perform equally well under circumstances for which they were not prepared. It is no secret the relevance that the lip movement has when communicating with other people for understanding each other, or the vivacity that a gesture such a smile can portray in someone’s face. For these reasons, we aspire to build a system capable of detecting and extracting accurately the location of the mouth given an image. This detection could be input to a variety of applications, such as affective computing, audio-visual speech recognition systems, or biometric systems. Therefore, in this paper we provide a framework for the automatic extraction of one of the most important features in the face, the mouth, under the only constraint that it should be closed. A block diagram of the system we propose is given in Figure 1.

In Section 2 we will briefly indicate some related work in Mouth Segmentation, next in Section 3 we describe our approximate mouth region segmentation method, followed by a technique for the extraction of important features in Section 4; in Section 5 we present an strategy for the accurate mouth segmentation and we show some of the results of our approach, and we conclude with a brief discussion in Section 6.

2 Related Work

Being the mouth such an interesting feature, there has been an extensive research pursuing its segmentation from the complete face image. We can roughly divide them in three categories:

2.1 Model-based Segmentation

A shape and a texture model are learned by training a database where feature points surrounding regions of interest such as the eyes, mouth, etc. have been previously marked for each image. [Cootes et al. 2001]. A set of parameters is then used for adjusting a new input image so as to minimize the error with respect to the mean shape and texture. The fitting process requires a good set of initialization points for being successful.
2.2 Color-based Segmentation

Color information is used for building Skin Likelihood Maps that in turn, will segment the image into skin and non-skin organs. The mouth is found due to the fact that it belongs to the latter category [Zhang et al. 2002] [Lucey et al. 2002]. As other color-based methods, this approach suffers from color constancy problems, which can cause the misdetection of people whose skin color are outliers of the previous model.

2.3 Feature-based Segmentation

This approach searches for features common to most of the people, being spread the use of the integral projection method for reducing the grayscale face image into a 1-D model, in order to find the peaks of this histogram, which correspond to distinctive face features such as eyebrows, eyes and mouth [Jiao et al. 2002] [Lyons et al. 2000]. This method requires that the scan is done with an up-right face, with as less background as possible.

2.4 Knowledge-based Segmentation

This technique relies on a-priori geometric information about the object of interest, e.g., for a face we know “anthropometric” distances among important points such as the eyes, nose, etc. Therefore, we use this measures for locating a certain feature. [DeCarlo et al. 1998]. Since the search is very constrained, this method will fail for objects with “uncommon” shape.

Alternatively, Gabor Wavelets has been used so as to extract important features from the face; nevertheless, its use has been linked to Face Recognition systems [Shen and Bai 2006] and, to the best of our knowledge, not to the detection of arbitrary features in the face image, as we propose.

3 Approximate Mouth Region Segmentation

In this section we will describe the procedure for an initial segmentation of a face region which contains the mouth. This region still needs further processing for the accurate mouth localization. Our method uses the highly-effective Adaboost-based [Viola and Jones 2004] Face Detection algorithm for the location and extraction of the face of the user from the rest of the scene, due to its strength for finding frontal faces. In our implementation, we allow only one face to be processed at a time, being the criterion to find the biggest of all the possible faces that result from applying [Viola and Jones 2004], as we assume that there is only one person who wishes to make use of the system, situated in front of it.

After we have obtained the face region, we need to further segment the image in order to minimize the search region for the mouth. For this, we locate an approximate position of the eyes by using a Gaussian distribution. We could rely on a model-based technique for approximately locating a region where the important mouth features such as the corners would be located based on the eyes location; however, we would first need to train a very large database of manually labeled face images and then provide optimal initialization points for the model-fitting process, as discussed in Section 2.1. Furthermore, as it occurs as well with some approaches based on Anthropometric measurements, we have some problems with users whose face shape is an outlier of the previous statistical descriptor, as discussed in Section 2.4.

Therefore, we adopt a much simpler, yet perceptual hypothesis: that given an upright face image, the mouth should be located in the lower third part of such a face. In summary, we use the eye detection output for limiting the width of our mouth search region, and we use the lower third of the detected face for limiting its height. Figure 2 illustrates the proposed method. Using these criteria, we were able to segment a region in which the mouth is present, for every image of our Face Database. Since we are working with low-resolution images, post-processing is needed in order to remove noise; hence, we apply a bilateral filter that preserves edges, fol-
Approximate Segmentation. a: Width Segmentation. b: Height Segmentation. c: Final Mouth Region Segmentation followed by a sharpening filter for their enhancement.

4 Mouth Features Extraction

In this section, considering that we have seriously shortened our search area, we use a set of Gabor filters for extracting horizontal features and the Hough Transform for finding the line that joins those features, which represents the dark shadow line between the upper and lower lips, that connects both mouth corners, as illustrated in Figure 3. We have chosen to use Gabor Wavelets because despite being computationally expensive, we are able to choose a proper orientation and scale values according to our search needs.

4.1 Feature Extraction by means of Gabor Wavelets

As we mentioned in Section 2, we will make use of a family of Gabor kernels, such as [Wiskott et al. 1997]:

$$\psi_j(x) = \frac{k^2}{\sigma^2} \exp(-\frac{k^2x^2}{2\sigma^2})[\exp(ik_j x) - \exp(-\frac{\sigma^2}{2})] \quad (1)$$

where:

$$k_j = \left( \begin{array}{c} k_{jx} \\ k_{jy} \end{array} \right) = \left( \begin{array}{c} k_{v} \cos \varphi_{\mu} \\ k_{v} \sin \varphi_{\mu} \end{array} \right), k_{v} = 2^{(\frac{1-\nu}{\nu})}, \varphi_{\mu} = \mu \pi \frac{8}{8}$$

Being the parameters $\mu$ and $\nu$ the ones that define the orientation and scale of the Gabor kernel, and $\sigma$ controls the width of the Gaussian window. This kernel is DC-free, and hence is not invariant to illumination changes. Given that we wish to extract horizontal features, we only take parameters $\mu$ such as the orientation of the filters is horizontal or almost-horizontal; because this will permit us to find the line that lays between both lips. Additionally, given that our images come from a webcam working at low-resolution images, and therefore we do not need to find the features for relatively high scales, which in turn shortens the necessary values for our parameter $\nu$.

Consequently, we consider parameters $\mu \in [3,4,5]$, corresponding to $\pi/2, 3\pi/8$, and $5\pi/8$ values. Accordingly, we choose $\nu \in [0,1,2]$, corresponding to the values $\pi/2, \pi/2\sqrt{2}$, and $\pi/4$; for which we obtain the set of Gabor waves that are applied to our mouth image, as shown in Figure 4. We noticed that the points we are searching for, are also the points at which the energy presents a higher response, as we can see in Figure 4.

4.2 Image Binarization by Feature Extraction

From the previous section, we have obtained nine different outputs for one single image; hence, we need to obtain a fused binary image in order to apply the Hough Transform in Section 5. Therefore, we merge all nine results into one single binary image.

Considering that mouth sizes and textures from user to user are variable, instead of finding the “n” maximum feature points common to all images, we adopt the feature point localization given in [Kepeneck 2001], which automatically finds the points with high information from the image, regardless of our a-priori knowledge about the number of points we expect and the place they should be located. In this method, peaks are found by searching with a window $W_o$ of size $W \times W$ at each point centered at $(x_o,y_o)$, by performing sequentially the following procedures:
Figure 5: Features that have been extracted for each of the nine images shown in Figure 4.

\[ R_j(x_o, y_o) = \max_{(x, y) \in W_o} R_j(x, y) \]  
\[ R_j(x_o, y_o) > \frac{1}{N \times M} \sum_{x=1}^{N} \sum_{y=1}^{M} R_j(x, y) \]

where \( N \times M \) indicates the size of the image. In Equation 2 we obtain a new image such as at each pixel we have the maximum value corresponding to the window \( W_o \). Next, by using Equation 3 we assure that the values we have found correspond to a global maximum, and not to a local high. In our experiments, utilizing a window such as \( W = 7 \) gives the best performance for the feature point localization. Results of the feature extraction procedure for the images shown in Figure 4 are presented in Figure 5.

Next, given the fact that:

1. We are searching for high response points, thus they should be present in all the images.
2. The orientation changes minimally from \( 3\pi/8 \) to \( 5\pi/8 \), thus, high response points should continue having high energy regardless the orientation of the filter we have used.

The last step is to obtain just one final “high information” merged image by using the nine binary images obtained after applying the process described in Equation 3 for each Gabor image. We achieve this by applying an \( \text{AND} \) logical operator to the set of nine images, which produces the output shown in Figure 6.

5 Accurate Mouth Segmentation

Once that we have found a single image that represents the mouth region, the last step consists in using these features for extracting an accurate mouth template. Our feature of interest is the line that connects the mouth corners, which will be used for a good initialization of the Active Appearance Model fitting process.

5.1 Hough Line Detection

Considering that our purpose is to build a system that works in a real-time environment, we decided to use the Progressive Probabilistic Hough Transform (PPHT) proposed in [Matas 2000] because it obtains results comparable to those of the Hough Transform, while minimizing the amount of computation needed to detect features by “exploiting the difference in the fraction of votes needed to reliably detect lines with different number of supporting points”, i.e., for long lines, it is not necessary to find every single point that lies on it. Its implementation also allows us to define a minimum length for a line to be returned, as well as a minimum distance that separates lines.

In our approach, given a binary image, we apply the PPHT and we find the longest horizontal line present. Then we extract its extreme points, as shown in Figure 7.

5.2 Active Appearance Model Fitting

Active Appearance Models (AAM) is a technique that tries to fit an average shape, and an average texture model to an image by tuning certain parameters, which are found after a training stage. We make use of such method for finding the contour that surrounds the mouth region.
5.2.1 Model Training

We have built a database containing 364 samples of mouth shapes and textures, including our own frontal images captured by using a normal web-camera, as well as some frontal images from the BioID Face Database [Jesorsky et al. 2001] in which the person has the mouth closed. Each image was acquired under indoor illumination conditions in office-alike environments. For representing the mouth region, we have chosen to use a 15-point model, as in Figure 8. We have manually registered a model for each of our images and trained our AAM offline. As a result, our model needs 30 parameters for explaining 95.14% of the database. When a new image is input, by adjusting the parameters of the training, an initial 15-point model will search locally for the best fit in order to minimize both the shape and texture error.

5.2.2 Mouth Contour Fitting

The fitting process requires an initial location for the points of the model, which means that this step is critical for obtaining an optimal result. Thus, for the AAM Initialization we consider the position of the mouth corners, as in Figure 7, for placing the model in the image. In addition, we use the length of the line obtained previously for rescaling the model. The AAM fits a contour that is highly likely to contain the mouth, and consequently, a final mouth region is segmented.

Results of the entire method for different users are shown in Figure 9, after applying the proposed method. The final mouth contour has been drawn in white color in the last column.

6 Conclusions and Future Work

We have proposed a method that combines a knowledge-based, a feature-based, and a model-based approach for the precise detection of the mouth given a low resolution image, under the unique constraint that the mouth must be closed. Our system is fully automatized and does not require additional input from the user. Using a-priori knowledge regarding the location of the mouth in a frontal face allows us to initially segment a rough region that contains the mouth in all the cases in our database. Our perceptual feature extraction
procedure by means of Gabor Wavelets has extracted horizontal features that resemble the shadow line that appears when we have our mouth closed, regardless of the ethnicity, gender or illumination conditions of the images existent in our database. After the line formed by the features has been detected; and, thus, good initialization points were provided to the AAM Fitting, we were able to segment an accurate mouth contour from which we can ultimately extract the mouth region. For our database, the proposed approach is able to detect the mouth corners in 95% of the face images, and extract a very precise mouth contour in more than 90%.

The segmentation takes in average 200ms. for a 320x240 pixel OpenCV Library [OpenCV], available for Windows XP-SP3 OS. The system could perform faster if using a more powerful hardware, however we are currently working on optimizing the code in order to reduce the processing time so the system can perform closer to real-time. Alternative options for a speed increase include to execute the most demanding procedure, the Gabor Wavelets, on a GPU.

Future work related to the method itself includes to use the system in applications such as affective recognition or mouth-based human-computer interaction, as well as extending the line detection so it can perform for non-upright and non-frontal faces. Given that the size of the bounding box that contains our mouth images is in average $40 \times 20$ pixels, it is also feasible to increase the accuracy of the mouth contour detection by using AAM enhanced for working with low-resolution images, as in [Liu et al. 2006].

References


