AN UNCONSTRAINED METHOD FOR LIP DETECTION IN COLOR IMAGES

Evangelos Skodras and Nikolaos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory,
Department of Electrical and Computer Engineering, University of Patras, Patras, Greece
{evskodras, fakotaki}@upatras.gr

ABSTRACT

The use of visual information derived from accurate lip extraction, can provide features invariant to noise perturbation for speech recognition systems and can be also used in a wide variety of applications. Unlike many current automatic lip reading systems which impose several restrictions on users, our efforts are towards an unconstrained system. In this paper we introduce a method using *k*-means color clustering with automatically adapted number of clusters, for the extraction of the lip area. The method's performance is improved by previously applying nearest neighbor color segmentation. The extracted lip area is morphologically processed and fitted by a best-fit ellipse. The points of interest (keypoints) of the mouth area are extracted, while a corner detector for fine tuning of mouth corners is applied. Experimental tests have shown that the algorithm works very well under natural conditions and accurate extraction of lip keypoints is feasible.

Index Terms— Lip detection, Color segmentation, *k*-means clustering, Lip reading

1. INTRODUCTION

Lip detection has attracted a lot of attention lately in the computer vision community. This increasing interest stems from the wide range of applications in which visual information is an integral part of, or can improve the performance and robustness of the overall system. These applications include audio-visual speech recognition, lip synchronization, synthetic talking faces and facial expression analysis. Nevertheless, accurate and robust lip detection is a non-trivial task due to large variations caused by the high deformable level of lips, different lip color tone, illumination conditions, appearance of teeth and tongue, presence of facial hair and so forth. During the last years many techniques have been proposed to achieve lip detection. Edge information was one of the first methods used to achieve lip segmentation [1]. When there is no shape or smoothness constraint the segmentation can be often very rough and lip boundary edges can be very low in magnitude and overwhelmed by strong false edges. A large category of techniques referred to as model-based, build a model of the lips and its configurations are described by a set of model parameters. These techniques include snakes [2], active contour models [3], deformable templates [4] and several other parametric models [5]. The advantage of these techniques lies on the fact that important features are represented in a low-dimensional parameter space. Also, they are invariant to rotation, scaling and illumination. Nevertheless, the construction of these models is often very challenging and a large training set is needed to cover the high variability range of lips. Moreover, the tuning of parameters is usually very difficult to achieve and many of them require manual selection and initialization. Color provides additional information which proves to be very useful for the task of lip detection and has been used widely [1, 4, 5, 6, 7, 8, 9].

In this paper we present our research efforts towards an unconstrained system for lip detection. Unlike most lip detection systems which impose certain constraints on users [6], such as wearing a head mounted camera [5], painting the subject's lips or having to operate in highly controlled environment, thus precluding practical applications, our method avoids all these preconditions. The only requirement of our system lies on the way the lip segmentation problem is approached. Namely, the lip area must be chromatically distinguishable from the rest of the skin area, given that it is the one with the greater redness along the face. Small areas of red artifacts do not affect the segmentation result and are automatically eliminated.

The paper is organized as follows. In Section 2 details of the proposed algorithm are presented. Section 3 presents the experimental results obtained using the proposed algorithm. Finally, in Section 4, conclusions are drawn.

2. PROPOSED METHOD

Lip detection is a complex problem because of the high variability range of lip shapes and color. To overcome this problem, in many methods, large training sets are used for training [9], several parameters (sometimes very sensitive to initialization) need to be tuned [3], or time consuming preprocessing steps must be taken. In our system, we avoid such pre-processing tasks, thus making it more general and independent of the database set.

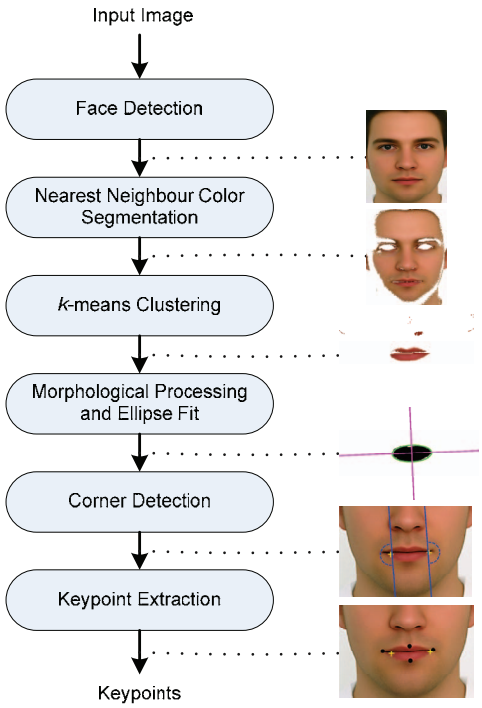

Figure 1 The processing steps of the proposed algorithm.

Our approach uses color segmentation in order to find an optimum partition of the given face image into lip and non-lip regions. The partition is based on the color difference between the lip and non-lip regions and is independent of the particular intensities. For this purpose, the image is transformed into the L*a*b* color space and then a combined technique of nearest neighbor color segmentation followed by a color-based *k*-means clustering with adaptive number of clusters is applied. Thereafter, binary morphological processing is applied and the lip object detected is approached by a best-fit ellipse. Finally, the most important points of interest are extracted (Fig. 1).

Although *k*-means segmentation has already been tried out for this purpose, the results were not very satisfactory as the number of clusters had to be determined manually, affected by many factors as the visibility of teeth, facial hair, uneven illumination etc. This is the main reason why the current *k*-means approaches fail to operate fully automatically [8]. One of the novelties of our system lies on the fact that it utilizes *k*-means clustering with an automatically adaptive number of clusters, whose performance and accuracy is improved by the use of nearest neighbor segmentation. Moreover, for the mouth corners, where color is unreliable because of dark areas, a corner detector for fine tuning is used.

2.1. Face Detection

The task of face detection, which constitutes the first step of the lip detection system, was carried out using a real time face detector implemented in openCV. The face detector is based on the detection of features called Haar-like features, which encode the existence of oriented contrasts between regions in the image. Details of the face detection system can be found in [10]. The face detector achieved very satisfactory results with a correct detection rate of 99% and false positive rate of 10% due to complex background or other faces in the image, tested on Caltech Image Database (CID) [11]. The 1% failure rate was due to very poor illumination conditions or heavy occlusions. On GTAV Face Database [12] all frontal and near frontal faces, without heavy occlusions, or great roll angles were correctly detected. The lack of background in those images caused null false positives.

2.2. Lip Segmentation

In order to extract the visual features in the face image acquired from the previous step, an accurate extraction of the lip area is essential. In our approach we use color information for segmenting the lip area, transforming the RGB face image into L*a*b* color space to increase the color contrast between lip and non-lip regions. L*a*b* color space has the advantage of being a perceptually uniform color space, matching the perceived color difference with qualitative distance in color space. This makes up a very useful cue for the following clustering algorithm, which uses a Euclidean-liked similarity measure. In the L*a*b* color space luminance information (L component) is separated from the chrominance information (a*, b* components), which we utilize for the segmentation. Intensity variation due to uneven illumination has minimal effect on the chrominance components, a fact that is very useful for our application.

Using color for locating lip regions and separating them from non-lip regions presents several problems. Although the color composition of human skin and lips differs surprisingly little across individuals [7, 8], total intensity of the reflection varies over a wide range [9]. Color values also depend strongly on the camera, frame grabber and illumination. Gaussian mixture models obtained from training [8], or fixed proportions between color space components (i.e. between Cb and Cr [5, 7]) sometimes fail to overcome these issues. In our case, the clustering method used is an unsupervised learning method where neither prior assumption about the underlying feature distribution nor training is needed [4].

Before applying the *k*-means clustering method, nearest neighbor segmentation in the L*a*b* color space is applied to the face image. It acts as an aid for the *k*-means algorithm by discarding unwanted non-lips pixels and thus decreasing the *k*-means clusters to a number belonging to a limited values space (2-5 clusters). This fact increases segmentation accuracy and speed. The nearest neighbor algorithm classifies each pixel in the face image by calculating the Euclidean distance between that pixel and a color marker, using both the a*, b* chromatic components. Lip pixels

have the feature of very high a* values (great redness) combined with low b* values (little greenness). We use two color markers for segmentation, made of the mean ten maximum a* values and mean ten minimum a* values, calculating also the correspondent b* values. The classification result contains all lip pixels and part of the skin region as seen in Fig. 1. It also reduces the influence of factors such as facial hair and the visibility of teeth, which would require an additional number of clusters to be segmented into.

The *k*-means clustering is applied starting with a number of clusters which is found experimentally to be the most common (4 clusters) and if necessary, adapting this number, until a relevant lip area size criterion is met. The criterion we have set involves the area of the cluster with the greatest a* mean value (lip area) to be within the range of 1.5 - 4 % of the whole face image. A slightly less than 1.5% percentage lip area (>0.8%) when the number of clusters reaches the minimum (2 clusters), is acceptable. These threshold values stem from extensive experimental tests on the CID and frontal and near frontal faces of GTAV face Database. The mean area values and mean number of clusters for the 27 people in CID are shown in Fig. 2.

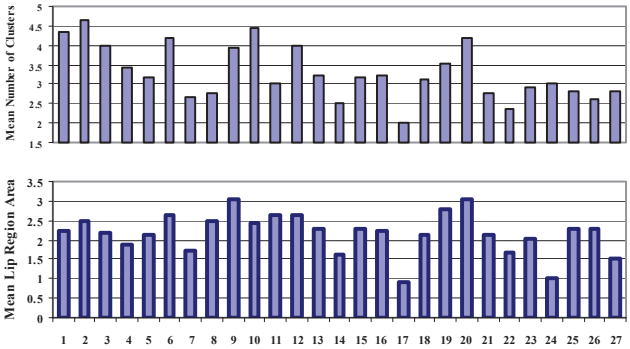

Figure 2 Mean number of clusters and mean lip region area for the people in Caltech Image Database.

2.3. Morphological Processing

From the above lip segmentation procedure a binary image is created, containing the segmented region with the max a* value (lip region). After applying morphological closing (to fix possible shuttered objects) and connected component labeling, we pick the largest area object as the lips region. For each region, in order to calculate its features, we compute the ellipse that has the same second moments as the region. The parameters of the ellipse, i.e., the center of mass, the major and minor axis as well as their orientation are computed [4].

A sensitive issue at this point is whether the picked object constitutes the whole lips area or just the lower lip, as lips area is often recognized as two separate objects. The criterion is based on the angle between the lines fitting the extreme left point *Left(x,y)* coordinates with the middle upper *MidUp(x,y)* and the middle lower *MidLow(x,y)* coordinates defined as

$$angle = atan\left(\frac{m_2 - m_1}{1 + m_2 m_1}\right) \quad (1)$$

where

$$m_1 = \nabla[Left(x,y), MidUp(x,y)]$$
$$m_2 = \nabla[Left(x,y), MidLow(x,y)] \quad (2)$$

Figure 3 Examples of three different angle cases. a) The detected object is just the lower lip (angle = 9°). b) The detected object is the whole lips area (angle = 43°). c) Borderline case, the detected object is the whole lips area (angle = 27°).

If the angle calculated is less than 25° we assume that the lip area consists of only the lower lip. If so, we incorporate in the lip area the object in the binary image which is above our lower lip object and meets some distance and orientation requirements set, i.e. the Euclidean distance between the centers of mass cannot be greater than the length of major axis divided by 2 and orientation must be about the same, with an allowed inclination of ±10° (Fig. 3).

2.4. Keypoint Detection

The points of interest (keypoints) widely used for lip reading and other applications are shown in Fig1. The upper and lower keypoints are found as the intersection points between the minor axis line and the upper and lower lip boundary, respectively. Unlike the upper and lower keypoints, which are precisely detected from the color segmented lip object, the mouth corners (left and right keypoints) are more difficult to detect because of their location in dark areas, where chromatic information is not visible. In order to detect them, we use the extreme left and right points of the lip object as starting points and search for corners in the proximity area using Harris corner detector. The proximity areas, where corners are searched out, are depicted in Fig 1.

From the corners detected in each side, we choose the one with the smallest Euclidean distance from the corresponding extreme object points as the left and right keypoints, respectively. If no corners are detected in either side, the corner strength threshold is automatically reduced until at least one corner is detected. In the case where only on the one side a keypoint is found, the corresponding keypoint on the other side is assumed using symmetry towards the minor axis.

3. EXPERIMENTAL RESULTS

In order to test our algorithm we used 421 images of 27 different people on CID [11] and 848 frontal and near

frontal face images of 44 different people on GTAV Face Database [12]. They were acquired under various lighting conditions without any particular make-up. Images very poorly illuminated, where color was hardly visible, were eliminated. Figure 4 shows representative results of our algorithm for different people. The yellow cross denotes the extreme right and left points of the lip region derived from the color segmentation. The black dots are the final keypoints after the corner fine tuning process. We can observe that the keypoints fit very well to the corners of the mouth as well as to the upper and lower lip boundary (perfect detections). Moreover, the method is robust even in challenging cases such as non-uniform lightning, bearded speakers, low color contrast between lip and non-lip area, or if teeth are visible. It is also unaffected by the yaw, pitch and roll angle as long as the lip region is visible. Our algorithm failed to extract accurate results in cases of heavily uneven illumination or when weak mouth corners were overwhelmed by strong beard corners.

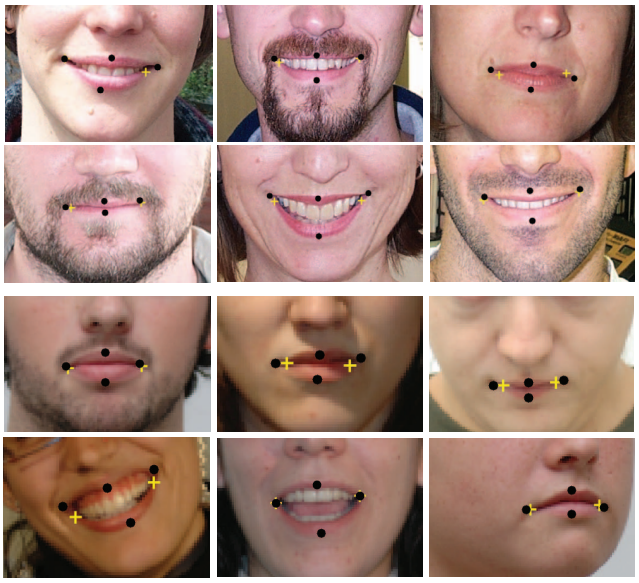

Figure 4 Several results on various people from CID (upper two rows) and from GTAV face Database (lower two rows). The yellow crosses denote the mouth corners before the fine tuning process.

Experimental results for the CID and GTAV databases are shown in Table 1. It is observed that in both cases lip detection is over 96%. Acceptable detection denotes perfect detections plus the cases where even though the keypoints are not placed accurately, the results are perceived as acceptable for lip reading purposes. For the GTAV database the detection figures are slightly higher. Since the proposed method does not contain complex and time consuming algorithms, it is very cost effective and with a proper C or hardware implementation, requirements of real time processing can be met.

Table 1 Lip detection results for the CID and GTAV databases

Database	Failed Detection	Perfect Detection	Acceptable Detection
CID	3.8%	94.3%	96.2%
GTAV	2.5%	93.3%	97.5%

4. CONCLUSIONS

The reliable extraction of visual information from lips, used widely in many applications, is a difficult problem and requires a robust lip detection approach. In this paper we have presented an automatic, accurate lip segmentation algorithm. It makes use of color information for segmentation of lip pixels with an automatic *k*-means clustering method. After morphological processing and ellipse fitting of the mouth object, a Harris corner detector is applied for fine tuning of the mouth corners. The use of corner detection makes the method suitable for applications which require a high level of precision, such as lip reading. Future efforts are directed towards finding the inner lips points of interest, denoting the visibility of teeth and testing its performance in audio-visual speech recognition.

5. REFERENCES

[1] X. Zhang, and R.M. Mersereau, "Lip Feature Extraction Towards an Automatic Speech Reading System", *Int. Conf. on Image Processing (ICIP'00)*, Vancouver, 2000.

[2] P. Delmas, P.Y. Coulon, and V. Fristot, "Automatic Snakes for Robust Lip Boundaries Extraction", *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Phoenix, 2002.

[3] X. Liu, Y. Cheung, M. Li, and H. Liu, "A Lip Contour Extraction Method Using Localized Active Contour Model with Automatic Parameter Selection", *Int. Conf. on Pattern Recognition (ICPR'10)*, Istanbul, 2010.

[4] A.W-C. Liew, S.H. Leung, and W.H. Lau, "Lip Contour Extraction from Color Images Using a Deformable Model", *Pattern Recognition 35*, Elsevier, pp. 2949-2962, 2002.

[5] N. Eveno, A. Caplier, and P.Y. Coulon, "A Parametric Model for Lip Segmentation*",ICARCV'O2*, Singapore, 2002.

[6] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel , "Towards Unrestricted Lip Reading", *Int. Journal of Pattern Recognition and Artificial Intelligence (IJPRAI),* 2000.

[7] R.L. Hsu, M.A. Mottaleb, and A.K. Jain, "Face Detection in Color Images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, VOL. 24, NO 5, pp. 696-706, 2002.

[8] M. Sadeghi, J. Kittler, and K. Messer, "Modeling and Segmentation of Lip Area in Face Images", *IEE Proceedings Vision, Image and Signal Processing*, 2002.

[9] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibelia, "Toward Movement-Invariant Lip-Reading and Speech Recognition*", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'95)*, Detroit, 1995.

[10] P. Viola, and M. Jones "Rapid Object Detection Using a Boosted Cascade of Simple Features", *CVPR*, Kauai, 2001

[11] Caltech Image Database, http://www.vision.caltech.edu/html-files/archive.html

[12] F.Tarrés, and A. Rama, "GTAV Face Database", http://gps-tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase.htm