Application of Directional Statistics to Gradient-Based Lip Contour Extraction for Speechreading

Mihaela Gordan*, Constantine Kotropoulos and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Artificial Intelligence and Information Analysis Laboratory,
GR-54006 Thessaloniki Box 451, Greece
{mihag,costas,pitas}@zeus.csd.auth.gr

Abstract. In the relatively large class of lipreading algorithms based on lip contour analysis, lip contour extraction is the first step. This refers to the detection of lip contour in the first frame of an audio-visual sequence, where no initial prediction of the lip contour is available. This task is usually solved manually or pseudo-automatically, requiring some manual human intervention. Past attempts to contour extraction based on typical edge detection schemes failed in the case of frontal gray level natural mouth images due to the weak contrast of these images. The proposed approach for lip contour extraction is still based on edge processing, but instead of using the edge magnitude, we process the edge direction. Clear patterns of edge directions can be observed around the lip contours, and noisy directions in the remaining areas. By using angular measures for edge following, good results are obtained with little manual intervention.

1 Introduction

A relatively large class of lip-reading algorithms are based on lip contour analysis. Examples of such algorithms can be found in [1–3]. In these cases, lip contour extraction is needed as the first step. By *lip contour extraction*, we usually refer to the process of lip contour detection in the first frame of an audio-visual image sequence. Obtaining the lip contour in subsequent frames is usually referred as *lip tracking*. While for lip contour tracking there are well-developed techniques and algorithms to perform this task automatically, in the case of lip contour extraction in the first frame the things are different. This is a much more difficult task than tracking, due to the lack of a good a-priori information in respect to the mouth position in the image, the mouth size, the approximate shape of the mouth, mouth opening, etc. So, while in lip contour tracking we have a good initial estimate of the mouth contour from the previous frame, this initial estimate is not always available for the first frame, but it has to be produced by some means.

Different authors tried different procedures to solve the extraction of a good lip contour in the initial frame. Of course, the goal would be to solve this task automatically. The most straight solution to lip contour extraction seems to be edge detection

* On leave from the Technical University of Cluj-Napoca, Faculty of Electronics and Telecommunications, Basis of Electronics Department, Cluj-Napoca, Romania.

(and/or region-based image segmentation). These methods work indeed quite well in profile images and even in frontal images if the speaker wears lipstick or reflective markers. However, in real mouth images, i.e. in frontal images recorded under natural illumination conditions and without any artificial marking of the lips, especially when these are gray level images, the above-mentioned techniques fail, due to the characteristics of these images: low contrast in the skin/lip area and non-uniformity of luminance inside the skin area/lip area [1], [4]. In these cases, the solution adopted is based on manual marking of more or less points on the lip contour (or even on manual drawing of the entire lip contour). When a large number of points (aprox. 50-100) are marked on the lip contour ([4], [5]), they are either used "as they are" to represent the lip contour (as for example in lip-reading based on Active Shape Models [6] and Active Appearance Models [3]). When a small number of points (e.g., 6 - 12 points) are marked on the lip contour, an interpolation procedure between them is applied to obtain the entire lip contour (as for example B-splines [5]), or these points are used to derive some geometric model parameters for the lip contour (as for example the widely used ellipsoidal model [7], or parabolic model [8]). In the latter case, the accuracy of lip contour extraction is limited by the fitness of the geometric model to the real lip contour. For example, in the case of an asymmetric mouth image (due let's say to a displacement of the video camera), the geometric model-based lip contour representation might be different from the real lip contour. (Such an example of mouth image is the one depicted in Figure 1 from the Tulips1 database).

Since achieving completely automatic lip contour extraction, without having available a first good estimation of this contour, seems to be a very difficult task in natural gray level frontal mouth images, the solution adopted by the researchers so far consists in the manual marking of some (more or less) points on the lip contour; under these circumstances, we aim:

(1) to require as few manual marked points as possible, while

(2) obtaining from these points a lip contour as similar as possible to the real one without any geometric model constrains or limitations, if possible.

This is exactly what we propose to achieve with the approach described in this paper. We focused our attention on the simple class of edge detection-based lip contour extraction methods, but instead of using the edge magnitude, we examined the edge direction in the frontal mouth images, and we observed that

(1) the edge direction shows a certain pattern, piecewise-quasi-constant, on the lip contour;

(2) the same or similar patterns are not present neither inside the lip area, neither inside the skin area. Instead, inside these areas we have "angular (directional) noise", i.e., random and highly variable edge directions (inside the same mouth image and from one mouth image to another).

So, while we cannot distinguish the pixels belonging to lip contours in the edge magnitude domain, we can differentiate them in the edge direction domain, which is an angular data domain. Since in the literature there are reported rather powerful mathematical tools for angular data processing [11], we can make use of these techniques for lip contour extraction based on the processing of the piecewise edge direction patterns. This is exactly what we propose in this paper, the result being an algorithm for lip contour extraction in frontal, gray level, natural mouth images.

However, due to the difficulty of describing the piecewise patterns on the contours without any a-priori information available, we require from the human operator the manual marking of start and end of each pattern on an “artificially created” color map that represents the edge direction by the hue color component. The typical number of points required to be marked is small (12 points). The contour extracted with the proposed algorithm has good quality, according by the visual estimation by human observers. In a future work, the algorithm will be made completely automatic, by learning the directional patterns from a set of training mouth images and using the learned patterns as initial criteria in edge following.

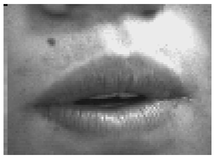

Fig. 1. Example of an asymmetric mouth image from the Tulips1 database [13].

2 HLS Color Maps of Edge Magnitude and Direction in the Mouth Image Area

As briefly stated in Section 1, the proposed algorithm for lip contour extraction is based on the processing of edge image information, mainly in the edge direction domain. Since the mouth images are gray level images, the edges can be easy computed with almost any edge detection scheme. Since the sophisticated edge detection methods mainly aim at estimating as best as possible edge magnitude, and since we are mainly concerned with edge direction processing, there is no need to use a high performance edge detector; simple techniques as gradient-based convolution masks will do. Therefore, in the purpose of edge extraction, we choose the Sobel gradient masks: the horizontal mask h_x and the vertical mask h_y:

$$h_x = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \qquad h_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{1}$$

By convolving the 3×3 neighborhood of each pixel with h_x and h_y, we obtain the horizontal gradient g_x and the vertical gradient g_y for the current pixel. From these we can compute the edge magnitude, $|g|$, and edge direction, tan α_g, thus describing the edge property of the pixel as

$$|g| = \sqrt{g_x^2 + g_y^2}$$
$$\tan \alpha_g = g_y / g_x. \tag{2}$$

Furthermore, due to the fact that we will require from the human observer to manually mark the start and end point of each directional edge pattern along the lip contour, we would need to have some meaningful visual mapping of the edge property of each pixel, described by the pair ($|g|$, tan α_g). When the edge property is described only by its magnitude $|g|$, a gray level image provides a complete representation of the edge map. When the edge property is described by a pair of values, we will need at least two image components for a complete edge representation. The most straightforward is to use a color image for edge mapping. Due to the angular nature of the edge direction, the simplest mapping is onto a color space that also has an angular component, as for example the Hue-Luminance-Saturation (HLS) color space, where the hue H is represented as an angle. In this case, we will use the luminance component L to represent the edge magnitude (thus maintaining the compatibility of representation with the typical magnitude-only edge images). The third component of the color space, i.e., the saturation S, will not be used, so it is set for simplicity to maximum (full saturated colors in the color map). The formulas for edge to HLS color space mapping are

$$\begin{aligned} L &= \frac{255 - |g|}{255}; \quad L \in [0, 1]; \\ H &= 180 + \frac{\alpha_g}{\pi} \cdot 180; H \in [0; 360]; \\ S &= \begin{cases} 0 \text{ if } L = 1 \\ 1 \text{ if } L \neq 1. \end{cases} \end{aligned} \tag{3}$$

Due to the (general) low contrast of gray level mouth images, $|g|$ has small values, so the color map image is usually not very clear. To obtain a better representation of the color map, an edge enhancement preprocessing step applied in $|g|$ is desirable. An example gray level frontal mouth image from the Tulips1 database, with its corresponding magnitude edge image, color map and pre-enhanced color map is shown in Figure 2. From this figure, it is rather easy to visually verify that, although the image of edge magnitude is too poor to allow the extraction of the lip contour, this task is possible on the color map, where the edge direction is represented by hue, due to the appearance of some specific hue patterns corresponding to lip contour. Still, outside these patterns (and, to some extent, inside some of these lip contour patterns), there is some noise present. Reducing this noise by applying a lowpass filter in the angular domain enhances the separability of the patterns around the lip contours.

In this purpose, we applied on the enhanced edge color map a 2-D circular mean filter in a 7 × 7 pixels window, as defined in [11]. The filter's output $y_{i,j}$ is defined as the sample mean direction of the set of 7 × 7 samples, and is given as the solution of the equation [11]:

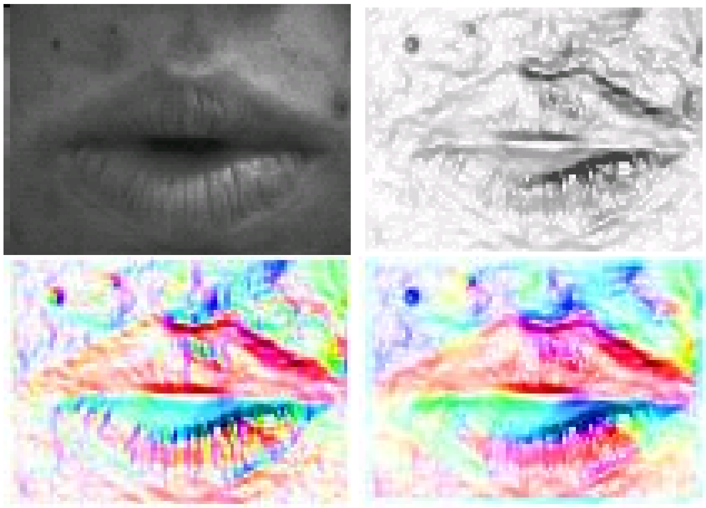

Fig. 2. From left to right: the original mouth image; the corresponding edge magnitude image; the pre-enhanced color map; the low-pass filtered color map in directional domain.

$$\sum_{k=-3}^{3}\sum_{l=-3}^{3}\sin(x_{i+k,j+l}-y_{i,j})=0, \tag{4}$$

which leads to:

$$y_{i,j}=\arctan\frac{\sum_{k=-3}^{3}\sum_{l=-3}^{3}\sin x_{i+k,j+l}}{\sum_{k=-3}^{3}\sum_{l=-3}^{3}\cos x_{i+k,j+l}}. \tag{5}$$

where $x_{i+k,j+l}$ are the hue values around the current pixel under processing, $x_{i,j}$, on the HLS color map prior to filtering, and $y_{i,j}$ is the hue value of the current pixel under processing from the HLS color map, after the 2-D circular mean filtering.

The window size of 7×7 was chosen on an experimental basis, by noticing that a smaller window size (3×3 or 5×5) doesn't filter properly the noise inside the lip region. Even if this relatively large window size breaks to some extent the direction constancy in some regions of the lip contour, this is not really a problem, being relatively easy to solve by an appropriate choice of the cost function in the edge following algorithm. The last image from Figure 2 shows an example of such a filtered HLS color map. This filtered color map will be passed to human operator for manually marking of 12 points, corresponding to the start and end of 6 lip contour regions, as illustrated in Figure 3.

3 The Proposed Edge Direction Processing Scheme for Lip Contour Extraction

In the previous section, we described the processing steps applied to the gray level frontal mouth image, up to the point of obtaining the manual marked HLS color map

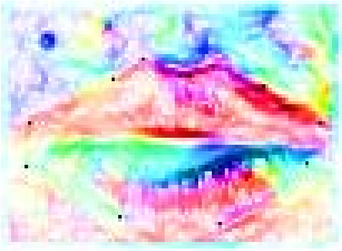

Fig. 3. The manually marked HLS color map of edge magnitude and directions.

of edge magnitudes and directions. From this point forward, we will apply an edge following algorithm inside each manually marked lip contour region defined by its corresponding start point and end point, with an angular distance-based cost function, which will be explained in detail in the next section. Finally, the remaining "gaps" in the lip contour, due to the spatial difference between the manually marked end point of one region and start point of the subsequent region, are filled by linear interpolation. This is generally possible without large discontinuities in the lip contour obtained, because these gaps are small. The resulting lip contour follows good the real lips outlines, but is not very smooth. To improve this aspect of the lip contour extractor, we sample 22 equally horizontally spaced points on the extracted lip contour as in [2], and then apply a quadratic B-spline interpolation between these 22 sampled contour points to obtain the smoothed closed lip contour, as in [5].

The processing steps leading to the HLS color map can be regarded somehow as preprocessing steps, while the subsequent ones, leading to the actually extracted lip contours, as processing ones.

4 The Modified Heuristic Directional Edge Following Algorithm

Having available the HLS color map of the edge magnitude and direction with the six different areas of lip contour patterns manually marked on it, we use the edge direction information from this map to obtain the (almost closed) lip contour. To do this, we apply an edge following algorithm.

The edge following algorithm proposed here is a version of the modified heuristic edge following algorithm given in [9]. The main particularity of our algorithm comes from the specific of the edge information being "followed", i.e., directional data. As a consequence, *the cost function* used in deciding the best path is computed based on directional (angular) differences, not on magnitude differences. Thus, instead of the general cost function of the heuristic edge following algorithm:

$$C(x_1, x_2, \ldots, x_N) = -\sum_{k=1}^{N} |g(x_k)| + a\sum_{k=2}^{N} |\alpha_g(x_k) - \alpha_g(x_{k-1})| + b\sum_{k=2}^{N} |g(x_k) - g(x_{k-1})| \tag{6}$$

we use a cost function containing only the term corresponding to angular differences between the current candidate pixel to be added to the path (x_N) and the previous pixel in the path (x_{N-1}), summed to the previously computed cost of the actual path, C(x_1, x_2, …, x_{N-1}). This cost function corresponds to the second term from the general cost given by (5); just that, taking into account that we work in the angular data domain, instead of using the linear difference of the angles given in (5), we use the circular deviation between the two angles given in [11]. Using this angular measure, the cost function of the proposed modified heuristic edge following algorithm becomes:

$$C(x_1, x_2, \ldots, x_N) = \sum_{k=2}^{N} \pi - |\pi - |\alpha_g(x_k) - \alpha_g(x_{k-1})||. \qquad (7)$$

The second difference of the proposed modified heuristic edge following algorithm compared to the one proposed in [9] regards the search space for the next candidate of the edge path: instead of using the 3 × 3 neighborhood of the last pixel in the path, we can use, in each of the six lip contour regions currently under processing, the a-priori information regarding the following direction on the lip contour (i.e., only from the start point forward to the end point), and this reduces the real possible candidates space to a 2 × 2 neighborhood, with the orientation specific to each lip contour region, as shown in Figure 4.

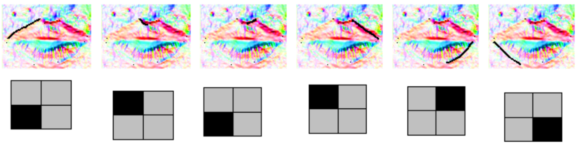

Fig. 4. The six 2 × 2 search neighborhoods: from left to right: NE neighborhood; SE neighborhood; NE neighborhood; SE neighborhood; SW neighborhood; NW neighborhood.

Finally, after obtaining the piecewise extracted lip contour, we perform the linear interpolation between the six lip contour segments corresponding to the manually marked regions.

The algorithm for lip contour extraction based on the proposed modified heuristic edge following in the angular (directional) space can be summarized as follows:

START

• For each region r, $r = 1, \ldots, 6$,

Do:

Step 1. “Read” the starting point (x_{r1}, y_{r1}) and the ending point (x_{rN}, y_{rN}).

Step 2. Select the directional 2 × 2 neighborhood of the region, denoted M_r, to be used in the edge following algorithm.

Step 3. Apply the heuristic edge following:

for (r_k=r_1; $r_k < r_n$;r_k++)

Select:

x_{rk}=$\arg\min_{j\in M_r(x_{r(k-1)})}$ C(x_{r1},x_{r2},...,$x_{r(k-1)}$,x_{rj})
where C(x_{r1},x_{r2},...,$x_{r(k-1)}$,x_{rj}) is given by (6).
• Obtain the closed lip contour by linear interpolation between: ((x_{rN},y_{rN})) and ($x_{(r+1)1}$,$y_{(r+1)1}$)), for $r = 1, \ldots, 5$ and ((x_{6N},y_{6N})) and (x_{11},y_{11}))
END

5 Experimental Results

The proposed lip contour extraction algorithm based on directional data processing was tested on mouth images from the two most used databases in speechreading experiments: Tulips1 [13] and M2VTS [12].

The images from Tulips1 database represent more difficult testing data than the ones from M2VTS, due to their lower contrast and more variable illumination of these images. Due to this reason, most of the experiments for the verification of our algorithm were performed on images from Tulips1.

For the experiments performed on Tulips1 database, we selected 10 mouth images, showing as large variation as possible in respect to: degree of mouth opening; symmetry/asymmetry of mouth image; mouth shape; lip-to-skin contrast; illumination. Based on this criterion, we selected:

- three mouth images (frames) of the subject Anthony;
- two mouth images (frames) of the subject Cynthea;
- one mouth image (frame) of the following subjects: Ben; Candace; Regina; George; Oliver.

Although a quantitative evaluation of the precision of lip contour extraction would be better, this would require all images to be labeled with the correct outline of the lips, and this can be done only manually. In general, we don't have this labeling available, so most authors decide to judge the performance of lip contour extraction by visual inspection [6]. Following the same methodology, we qualified visually the extracted lip contour according to [6]. In every situation from our experiments, the extracted lip contour was visually classified as *good*. Some example images (the original and the extracted lip contour) are given in Figure 5.

For the sake of comparison with the performance of other (somehow similar) lip contour extraction methods, we depicted a frontal speaker image from M2VTS database, we selected manually the rectangular region of interest containing the mouth, and used this ROI as input test image for our algorithm to perform outer lip contour extraction. The result obtained was visually compared with the one reported by the Centre for Vision, Speech and Signal Processing of the University of Surrey [10], using manual marking of 11 points on the outer lip contour and B-splines to interpolate between them. Although there are some differences in the two contours, the visual examination of our result classifies it as good. This result is given in Figure 6. It has to be mentioned here that this small number of manually marked points is suitable for M2VTS images, where the mouth area is small, but will not be enough for an accurate contour interpolation in case of Tulips1 images, which have bigger size. Instead, our method allows sampling as many contour points as desired, without requiring more manually labeled points on the lips.

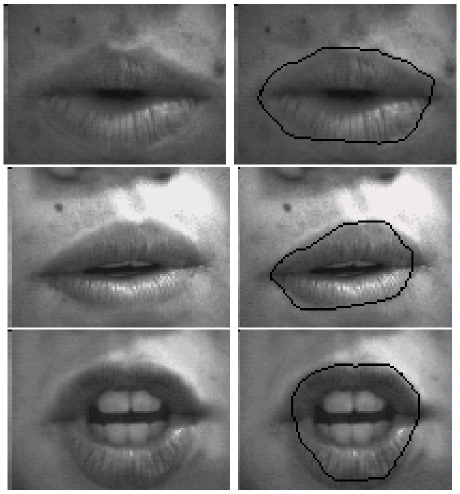

Fig. 5. Three outer lip contour extraction examples, for 3 subjects from Tulips database.

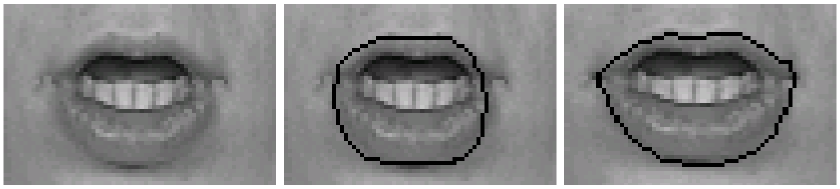

Fig. 6. Outer lip contour extraction example for M2VTS database. From left to right: the original mouth image; outer lip contour tracked with our algorithm; outer lip contour tracked with B-splines.

6 Conclusions

In this paper we proposed a new lip contour extraction solution for frontal, gray level, mouth images, recorded under natural illumination conditions, without any artificial marking of the lips. The proposed approach has the advantage of being computationally very simple, requiring a small extent of human interaction, while providing a good quality of the extracted lip contour. The primary information used is the edge direction map of the original mouth image, interpreted as angular data. These angular data are further processed making use of various measures from directional statistics: circular mean direction; circular deviation between two angles; circular mean filters. Based on these measures we defined a new modified heuristic edge following algorithm for lip contour extraction; the experimental results obtained proved the good functionality of the algorithm.

Our future research will be oriented on automatic detection of angular lip contour patterns present in the mouth image, taking into account the fact that the patterns are mainly independent to the individual mouth image. This allow us to develop some statistical models of these patterns, to be used in a statistical model-based search procedure for lip contour extraction.

7 Acknowledgement

This work was supported by the European Union Research Training Network "Multimodal Human-Computer Interaction (HPRN-CT-2000-00111)".

References

1. Kaucic, R., Dalton, B., and Blake, A.: Real-time lip tracking for audio-visual speech recognition applications. Proc. European Conf. Computer Vision, Cambridge, UK, 1996 376-387.
2. Dupont, S., and Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia **2(3)** (Sept. 2000) 141-151
3. Matthews, I., Cootes, T., Cox, S., Harvey, R., and Bangham, J. A.: Lipreading Using Shape, Shading and Scale. Proc. Auditory-Visual Speech Processing, Sydney, Australia, December 1998 73-78.
4. Luettin, J., Thacker, N. A., and Beet, S. W.: Active shape models for visual speech feature extraction. Speechreading by Humans and Machine. NATO ASI Series, Series F: Computer and Systems Sciences **150** 383-390, Springer Verlag, Berlin, 1996
5. Ramos Sanchez, M. U., Matas, J., and Kittler, J.: Statistical chromaticity models for lip tracking with B-splines. Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication, Crans Montana, Switzerland, 1997 69-76.
6. Luettin, J., and Thacker, N. A.: Speechreading using probabilistic models. Computer Vision and Image Understanding **65(2)** (February 1997) 163-178
7. Hennecke, M. E., Prasad, K. V., and Stork, D. G.: Using deformable templates to infer visual speech dynamics. Proc. 28th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 1994 578-582.
8. Tian, Y., Kanade, T., and Cohn, J. F.: Robust lip tracking by combining shape, color and motion. Proc. ACCV'2000,Taipei, Taiwan, January 2000 1040-1045.
9. Pitas, I.: Digital Image Processing: Algorithms and Applications. John Wiley & Sons, February 2000.
10. *http://www.ee.surrey.ac.uk/EE/VSSP/xm2vtsdb/results/lips/*
11. Nikolaidis, N. and Pitas, I.: Directional statistics in nonlinear vector field filtering. Signal Processing **38(3)** (Aug. 1994) 299-316.
12. Pigeon, S., and Vandendorpe, L.: The M2VTS multimodal face database. Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication (J. Bigun, C. Chollet and G. Borgefors, Eds.) **1206** (1997) 403-409
13. Movellan, J. R.: Visual Speech Recognition with Stochastic Networks. Advances in Neural Information Processing Systems (G. Tesauro, D. Toruetzky, and T. Leen, Eds.) **7**, MIT Pess, Cambridge, MA, 1995