LIP ACTIVITY DETECTION FOR TALKING FACES CLASSIFICATION IN TV-CONTENT

Meriem Bendris[1,2], *Delphine Charlet*[1], *Gérard Chollet*[2]

[1] France Télécom *R&D* - Orange Labs, France
[2] CNRS LTCI, TELECOM-ParisTech, France
{*meriem.bendris,delphine.charlet*}@*orange-ftgroup.com*
gerard.chollet@telecom-paristech.fr

ABSTRACT

Our objective is to index people in a TV-Content. In this context, because of multi-face shots and non-speaking face shots, it is difficult to determine which face is speaking. There is no guaranteed synchronization between sequences of a person's appearance and sequences of his or her speech. In this work, we want to separate talking and non-talking faces by detecting lip motion. We propose a method to detect the lip motion by measuring the degree of disorder of pixel directions around the lip. Results of experiments on a TV-Show database show that a high correct classification rate can be achieved by the proposed method.

Index Terms— Audiovisual identity indexing, video search, visual speaker detection.

1. INTRODUCTION

With the increase of internet use, we see a proliferation of multimedia content (Video On Demand, TV websites interfaces). While there are many available technologies capturing and storing of multimedia content, technologies to facilitate access and manipulation of multimedia data need to be developed. One way of browsing this type of data is to use audio-visual indexing of people, allowing a user to locate sequences of a certain person. In our study, we focus particularly on audio-visual indexing of people in popular TV-programs. Identifying people in this video context is a difficult problem due to many ambiguities in audio, in video and in their association. First, concerning the audio, the speech is spontaneous, shots are very short and often people are speaking simultaneously. Secondly, concerning the visual information, faces appear with many variations in lighting conditions, position and facial expressions. Finally, associating audio and visual information in this context introduces many ambiguities. The main one is the asynchrony between sequences of speech and face appearance of a person. Then, it is difficult to determine which face is speaking in the cases of multi-faces shots or shots where the speaker face is not detected (not visible).

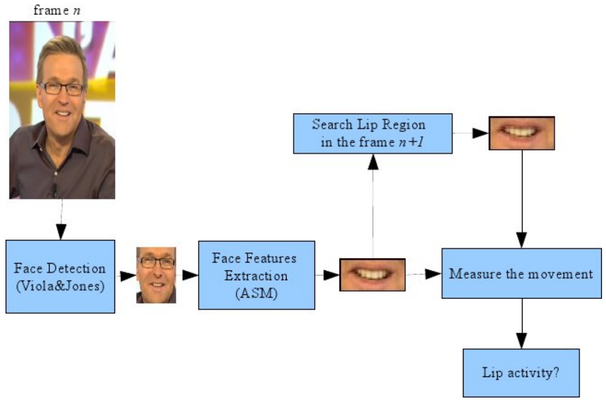
In this work, the objective is to detect whether each face in a video shot is speaking or not using visual information in order to associate the correct face to the speaker. We chose to accomplish this by detecting lip activity. In the literature, there are several methods which study visual information of speech activity to improve speech recognition systems [1, 2] and audio-visual synchrony [3]. Most of these methods requires a high level of lips representation. Few authors have focused only on detecting mouth activity to localize the speaker. In our context, because of the variability of the face appearance, it is very difficult to extract the shape of lips with great reliability. In [4, 5] the authors use a difference between mouth region to detect lip activity. Our contribution in this work is to develop a lip activity detection using the disorder of the directions of pixel.

This paper is organized as follows : in section 2, a system of lip activity detection in TV-context is proposed. Section 3 presents briefly the TV-Show database used to perform our experiments. Finally, experiments are reported in section 4

2. LIP ACTIVITY DETECTION

Our objective is to detect a lip activity in order to classify faces as talking/non talking in TV-Context. The first challenge is to identify the information to be extracted to detect the lip activity. In the domains of lip reading, synchrony and visual speech speaking detection, there are 2 types of mouth region representations : grey-level information [1, 3] and high level visual information (geometrical) like lip width, height, surface, mouth opening [2]. In [1], the authors combine acoustic features with visual features represented by lip contour and gray scale of the mouth region in order to improve the speech recognition performance. In [2], visual speech parameters are represented by the outer and inner lip width, outer and inner height and lip surface. In [3], a *Discrete Cosine transform (DCT)* coefficients of the grey-level lip area extracted and combined with *MFCC* coefficients to measure the audio-visual synchrony. The high level features are not appropriate in a TV-Context because it is very difficult to extract the shape of lips with great reliability due to the quality of face appearance. We choose to represent the mouth region as a rectangle of pixels.

Figure 1 presents the details of the lip activity detection for talking faces classification. For a given visible face shot, the *Viola*&*Jones*[6] face detector is applied on each frame to localize faces. Then, a mouth is localized using the facial features detector. After that, the mouth region is searched in the next frame. This step is important to align the mouth region even in the case of moving faces. For each shot, the final measure of movement is the average of the measures calculated between two consecutive mouth regions. Details of each step are presented below.



frame n
Search Lip Region in the frame $n+1$
Face Detection (Viola&Jones)
Face Features Extraction (ASM)
Measure the movement
Lip activity?

Fig. 1. Details of the lip-activity detection system

2.1. Facial feature detection

The facial feature detector is based on the Active Shape Model (*ASM*s) described in [4]. The shape model is learned from a set of manually annotated shapes of faces (unobstructed frontal views) as follows : the first step is to align all shapes of the learning data to an arbitrary reference by a geometric transformation (rotation, translation and scaling). The second step is to calculate the average shape. These two steps are repeated until convergence by minimizing the *average Euclidean distance* between shape points. At the end of the process, the facial shape model is obtained by principal component analysis (*PCA*) of the average of the aligned shapes. Thus, we obtain the principal modes of shape variations.

To extract the facial features from a new image, first, faces are detected using the *OpenCV* implementation of the *Viola*&*Jones* face detection algorithm [6]. After that, the shape model is positioned on the face, and iteratively deformed until it sticks to the face of the image. We used *Stasm* software package [7] for locating 68 landmarks : each eye is represented by 5 points (left and right corners, top and bottom eye opening and the iris center) and each eyebrow is represented by 6 points. The lip contour is described by 19 points (18 points for the upper and lower lips and a point representing the center of the mouth). The nose is represented by 12 points and the face contour by 15 points.

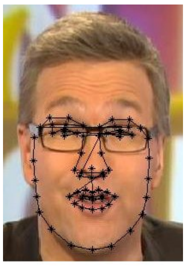


Fig. 2. Example of Face Features Detection - 68 landmarks

Figure 2 shows an example of the facial features detection using *Stasm*. The facial features can be located with high reliability in standard lighting conditions, frontal face position and classical facial expression. Unfortunately, the system is not reliable enough with pose changes such as faces looking up or down and expressions such as a mouth wide open or wince.

2.2. Selection of mouth region

To minimize variation due to the face features localization, mouth regions must be aligned. For a given frame, the mouth region is located using the facial feature detector. The rectangle around the lips is enlarged in the next frame to search the best region aligned with the previous one by taking the region that minimizes the *mean squared difference (MSD)*. Figure 3 illustrates the mouth regions alignment.

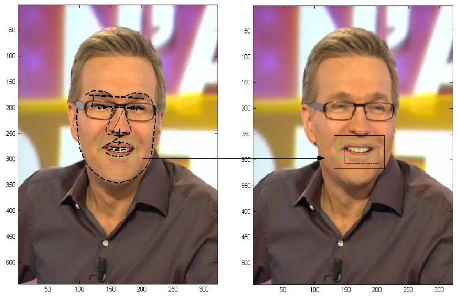


Fig. 3. Search region around the mouth

2.3. Lip activity measures

The goal is to measure the lip activity in order to identify the face corresponding to the speaker in shots. At this step, for each visible face in a video shot, a collection of aligned lip regions is detected. To measure a lip movement we propose to measure de degree of disorder of pixels direction in mouth region. To measure the pixels directions we used optical flow technique [5]. Each pixel is associated to a *2D-vector* representing the estimated direction of movement.

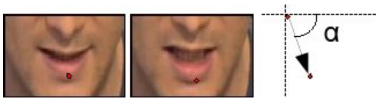


Fig. 4. Estimated directions of pixel motion with optical flow

For each face detected in a video frame X_t, the entropy of the moving directions of pixels around the mouth is used to measure the lips movement. Between 2 mouth regions in 2 consecutive frames X_t and X_{t+1}, we compute the entropy of the pixels directions of lip region as follows :

$$\mathrm{Entropy}(\mathrm{X_t}, \mathrm{X_{t+1}}) = -\sum_i P(\alpha_i) \times log(P(\alpha_i)) \qquad (1)$$

where $P(\alpha_i)$ is the probability that a random pixel chosen from the lips region will have the direction α_i. Figure 4 shows an example of the direction of movement estimated for a pixel. For each face detected in the segment X (being the sequence of N frames $(X_t, .., X_{t+N-1})$), the lip activity measure is calculated as follows :

$$\mathrm{Mv(X)} = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathrm{Entropy}(X_{t+i-1}, X_{t+i}) \qquad (2)$$

Then, the decision of talking face is taken by comparing $Mv(X)$ to a given threshold.

In order to classify talking faces in TV content, a method is proposed in [8, 4] where the *Mean Squared Difference MSD* is used to detect lip motion. In order to evaluate our lip activity measure, we implemented the method proposed in [8]. The mean squared difference between pixels is computed between aligned consecutive mouth regions. For each face detected in a shot, the lip movement is measured by the average of the *MSD* obtained. A fixed threshold is applied to determine whether the lips are moving or not.

3. TV-SHOW DATABASE

Few works have been focused on real TV-Context database. To our best knowledge, there are no public data annotated with both voice and facial appearance. It was necessary to construct and annotate a real TV-Context database. We selected a TV-Show in which people appear often in order to have a large number of examples. The experiments are done on the live TV-Show *”on n’a pas tout dit”*, a French TV-program presented by *Laurent Ruquier*[1] on the public channel ”*France2*”. Many commentators discussed, with some celebrities, the news of the moment in a good mood. Figure 5 shows examples of shots. There are several types of shots : focus on one face, multi-faces shots, public and general shots around table and video edited shots between two speakers. Except for *Laurent Ruquier*, who has a longer presence, shots are often very short. In this data, speech is spontaneous and faces move with quick changes of poses.

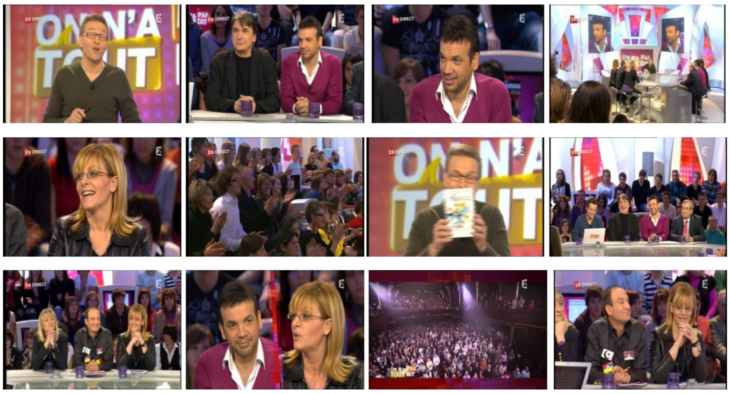


Fig. 5. TV-Show Database - examples of shots collected

One show contains typically 10 people who speak during the $50mn$ of the show. Four shows have been manually annotated as follows :

– *Audio annotation :XML* format files containing information from beginning and end of each speech sequence, the identity of the speaker and the spoken text. The applause and interposition of two or more speakers were also annotated with text if it is understandable. The tool used for the audio annotation is *transrciber*[2].

– *Video annotation :* each participant (host, commentators and guests) was annotated from the time he appears to the shot end. The information manually annotated is the identity of the person, coordinates of the face region in the shot, and the position of the face relative to the camera (Right, left, front, quarter right, quarter left, top, bottom and the face occultation). The tool used for the audio annotation is the *Elan*[3] software.

In shows, people do not necessarily speak and appear simultaneously : each person is visible more than 60% of his speaking time and the speaking time of a visible face is about 35% of the time. Thus, for these TV-shows, the probability that a speaker is visible is much higher (almost twice as much) than the probability that a face is speaking. Those points explain the context of working in TV-content where no synchronization between speech and appearance sequences is guaranteed.

[1] http ://www.ruquier.com/tv/onapastoutdit.html

[2] http ://trans.sourceforge.net

[3] http ://www.lat-mpi.eu/tools/elan/download

4. EXPERIMENTS AND EVALUATION

4.1. Protocol

To evaluate the ability of the lip motion detector to separate talking faces and non talking faces, we select the following database : for each speaker, we selected shots where the face is visible (the face may appear front, quarter right or left). Among these shots, we selected the *Talking face shots* where the face speaks during the entire shot and *Non-talking face shots* where the face does not speaks throughout the duration of the shot. We recovered a total of 582 speaking face shots (mean duration $3.45sc$) and 667 non-speaking face shots (mean duration $2.17sc$). It is important to note that in this type of TV-Context, the shots are short and face movements are very fast. Shots of talking faces are relatively longer than non-speaking faces. In order to improve the detection of activity, the movement is measured between a given frame n and $n + 3$.

4.2. Results and discussion

Figure 6 shows an example of mouth regions alignment. In TV-content, faces move very quickly, and it is necessary to align the mouth regions to calculate a variation of the lips without taking into account face movements.

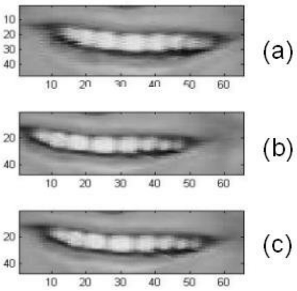
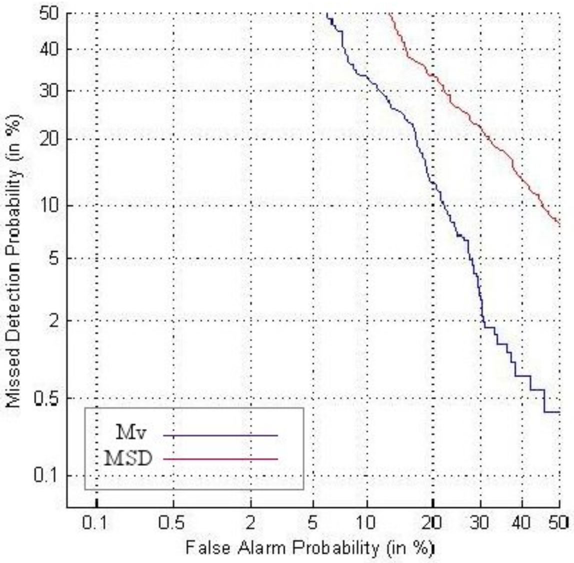


Fig. 6. Alignment of mouth regions (a) : mouth region detected in the frame n, (b) : mouth region in the next frame without alignment (c) : mouth region in the next frame with alignment

Figure 7 summarizes the performance of the lip activity detection for talking faces classification system using the measures mean squared difference (*MSD*) and the disorder of pixels directions (*Mv*) presented in section 2. The *DET* curve describes the false rejection rate (*FRR*) according to the false acceptance rate (*FAR*) by varying the threshold of decision. The equal error rate (*EER*) is the point where the false reject rate is equal to the false acceptances rate. The performance obtained by measuring the disorder of the movement directions are significantly better (in terms of *EER*) than that obtained by a mean squared difference of the pixels.



	EER
Mv	17.52%
MSD	25.64%

Fig. 7. DET curve - The performance of the lip activity detection for speaking face detection module

The proposed method based on measuring the degree of disorder of pixels achieves an *EER* of 17.5%, which is much better than the method based only on the difference of the pixels. Errors can be explained by the fact that the method makes the assumption that a mouth activity means speech activity. Our method is not reliable in the case of winces, sudden gestures of mouth or giggles. This can be explained by the unreliability of the facial features detectors to extract feature because the shape model was learned from faces without expression. Figure 8 shows an example of errors detection.

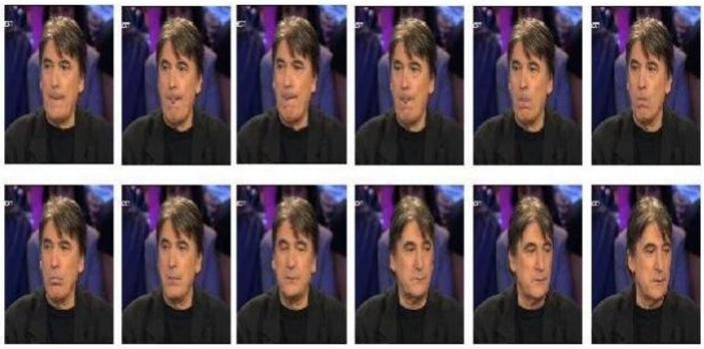


Fig. 8. Example of error detection - Person is classified as having a mouth activity, however does not speak

5. CONCLUSION

Our ultimate objective is to automatically annotate people in TV-Context database. The audio-visual indexing of people exploits these two information in order to improve the performance. In this context, the synchronization between sequences of a person's appearance and sequences of his or her speech is not guaranteed. To accomplish the fusion, it is necessary to associate them correctly. Hence, we must determine which face is speaking. In this paper, we propose a new method for detecting speech activity using visual information in TV-Context in order to find the association between speech sequences and their corresponding face. The proposed method uses the disorder of pixels direction around the mouth. The evaluation has been performed on manually annotated TV-shows. A significant improvement is observed compared to the difference between pixels. The results show that in the real data, mouth activity indicates, in most cases, speech activity. The annotation of a large database was necessary to achieve our experiments. A way to improve the talking face classification is to investigate on methods to enhance the reliability of the proposed measure to the laughing and wince by using a facial feature detector witch takes into account all kind of distortions of the lips.

In future work, we intend to integrate the lip activity detector as a cue to associate a face to the speaker in audio-visual indexing of people system. In particular, we want to use the lip activity to resolve inconsistencies between audio and visual index of people in order to confirm a talking-face sequences.

6. REFERENCES

[1] Stphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on*, 2000.

[2] Martin Heckmann, Frederic Berthommier, and Kristian Kroschel. A hybrid ann/hmm audio-visual speech recognition system. *In International Conference on Auditory-Visual Speech. Processing*, 2001.

[3] Enrique Argones-Rúa, Hervé Bredin, Carmen García-Mateo, Gérard Chollet, and Daniel González-Jiménez. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. *Pattern Analysis and Applications Journal*, 2007.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 61 :38–59, 1995.

[5] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell. Visual speech recognition with loosely synchronized feature streams. *Tenth IEEE International Conference on Computer Vision, ICCV.*, 2005.

[6] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, April 2001.

[7] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. pages 504–513, 2008.

[8] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello ! my name is... buffy – automatic naming of characters in tv video. *Proceedings of the British Machine Vision Conference*, 2006.

[9] Hongzhou Zhang Lin Wang Chengbo Wang, Yongping Li. Multi-modal biometric verification based on far-score normalization. *International Journal of Computer Science and Network Security (IJCSNS)*, April 2008.

[10] K. Mase and A. Pentland. Lip reading by optical flowlip reading by optical flow. *IEICE of Japan*, pages 796–803, 1990.