BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN 2348-0513
Vol. 3, Issue 2, Feb 2015, 41-46
© BEST Journals

AN ESTIMATION TO SPEAKING FREQUENCY IN VIDEO STREAMING

SHUVRA CHAKRABORTY & ISMAT RAHMAN

Lecturer, Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh

ABSTRACT

This paper presents an algorithm to estimate the state of open or closed mouth in real time video streaming. It detects associated frequency of mouth motion events in specific time interval which helps to identify speaking and yawning. While speaking, accurate lip movement is estimated using 3D cascade classifier and only lip coordinates are stored in the database for memory efficiency. Horizontal and vertical distances of the lip are used as an estimation of lip surface area. Speaking frequency is counted based on comparison of statistical data in database. Our proposed method shows satisfactory performance with a high speaking frequency detection rate on live video streaming.

KEYWORDS: Image Processing, Lip Tracking, Lip Corner, Speaking Mode, Streaming

INTRODUCTION

Now a days, Human-Computer interaction topic has be- come a cutting edge field of research. Now, computers are more ubiquitous than never. It is usual to find them any- where, but still the Human-Computer interaction (HCI) is only based on manipulation. Communication between computers and humans must evolve, thus every time humans engage a transaction with computers, they should be able to interact in a more ”human” way. Current researches on HCI are summing up the efforts from different fields involving Artificial Intelligence, Speech Recognition, or Computer Vision in order to make computers able to accordingly interact with their interlocutors. To achieve this goal, when computers are engaged in an interaction with humans, the acquisition of automatic audio-visual feedback is a key point. This information can be provided in different ways. Speech is the most common one, but spoken words are highly person and context dependent. Nonetheless, speech recognition and speaking detection is a very active and challenging field. In recent years, problems in the automatic speaking recognition (ASR) have drawn the attention of researchers [1]-[3]. With the presence of noise as in real world circumstances, the ASR rate could be dramatically reduced. The ASR system would be able to provide an appreciable performance only under a certain controlled environment. With the inspiration of lips-reading capability from the impaired society and the limitation of the noise robust techniques, the audio-visual speaking recognition (AVSR) has become a research trend and is growing rapidly [4].

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 Identifying people speaking mode in video streaming. It’s 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). Thus, we have associated the video information here to develop a real time algorithm for speaking frequency analysis.

LITERATURE REVIEW

Our objective is to detect a lip activity in order to classify faces as speaking/non speaking in video streaming. 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 two types of mouth region representations: grey-level information and high level visual information (geometrical) like lip width, height, surface, mouth opening. The degree of lip movement can be related to the state open or closed mouth and frequency of mouth motion. Work on mouth shape detection is generally based on lips segmentation. Lip boundary extraction is an important problem that has been studied to some extent in the literature [1]-[4].

Lip segmentation can be an important part of audio-visual speech recognition, lip-synching, modelling of talking avatars and facial feature tracking systems. In audio-visual speech recognition, it has been shown that using lip texture information is more valuable than using the lip boundary information [5]-[6]. However, this result may have been partly due to inaccurate boundary extraction as well, since lip segmentation performance was not independently evaluated in earlier studies. In addition, it is possible to use lip segmentation information complementary to the texture information. Lip boundary features can be utilized in addition to lip texture features in a multi-stream Hidden Markov model framework with an appropriate weighting scheme. Thus, we conjecture it is beneficial to use lip boundary information to improve accuracy in AVSR. Once the boundary of a lip is found, one may extract geometric or algebraic features from it. These features can be used in audio-visual speech recognition systems as complementary features to audio and other visual features.

The visual appearance of the human mouth holds a lot of information about the individual it belongs to. It is not only a distinct part of each person's look the lip shape also serves as mean of expressing our emotions. Moreover, the lips' motion indicates if the person is talking and even allows conclusions about what is being uttered. Localizing the exact lip boundaries in an image or video is demanded. Valuable information for various applications with human computer interaction and in automated surveillance is required in many commercial applications.

Recent techniques use knowledge about the lip’s color or shape to identify and track the lips. Indeed, color differentiation is an effective technique for locating the lips.

Lip feature extraction, or lip tracking, is complicated by the same problems that are encountered with face detection, such as variation among persons, lighting variations, etc. However, lip feature extraction tends to be more sensitive to adverse conditions. A moustache, for example, can be easily confused to be an upper lip. The teeth, tongue, and lack of a sharp contrast between the lips and face can further complicate lip feature extraction.

A relatively large class of lip reading algorithms is based on lip contour analysis. Different authors tried different procedures to solve the extraction of a good lip contour in the initial frame. Region-based image segmentation and edge detection have been proposed. These methods work quite well in profile images and also in frontal images where the speaker wears lipstick or reflective markers.

But of course, our goal would be to solve this task automatically by building an effective system.

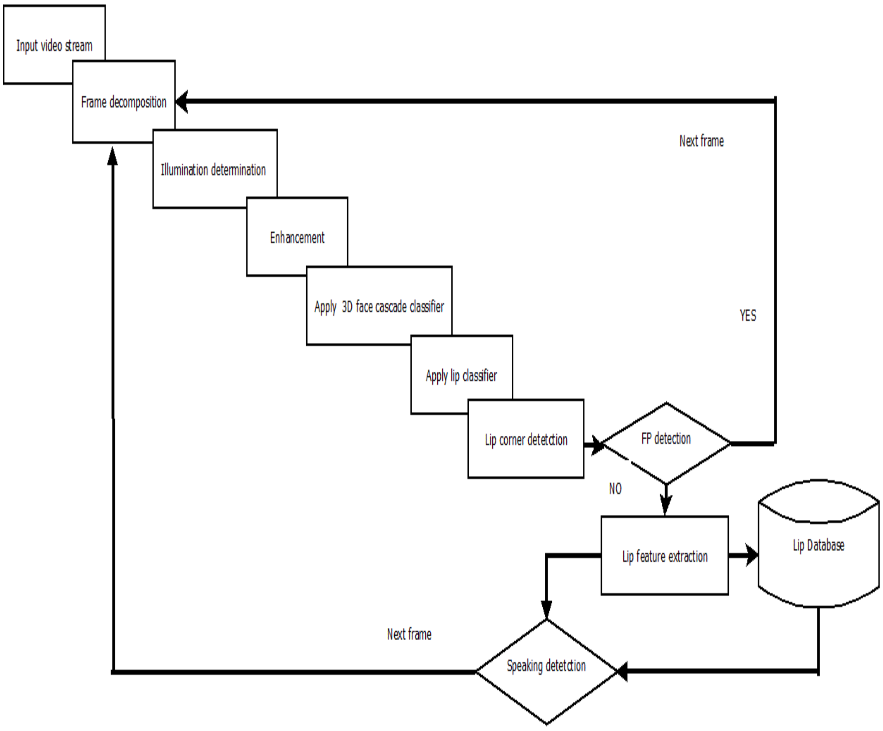


Figure 1: Proposed System Workflow

PROPOSED SYSTEM ARCHITECTURE

Figure 1 presents the detailed system architecture of our speaking frequency count system. For any given video streaming, we follow several steps. At first, we decompose it into frames. If video streaming is live, we analyze it frame by frame. Now, for a given visible enhanced face shot, we apply our own 3D cascade classifier which is compiled from OpenCV cascade classifier and our own training database. For multiple face detection, we apply viola jones face detector again to face detection. After face localization, lip boundary is localized using the facial features detector given by OpenCV.

To avoid False Positive Rate, we detect lip corner from the frame. After that, the process continues for the next frame. We store corresponding lip coordinates in database to keep a dynamic history of speaking information. 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.

SYSTEM HARDWARE AND REQUIREMENTS

For any image processing or pattern recognition algorithm performance evaluation, system configuration is a challenging issue as image processing algorithms are usually heavyweight task. Our system has been experimented on UBUNTU Operating System in Corei7 processor. We have considered each frame size to be 130X150.

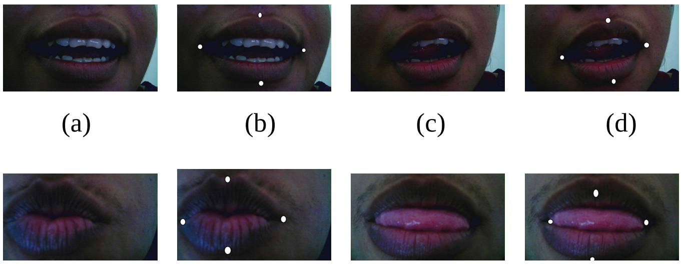
RESULTS AND DISCUSSIONS

For testing the performance of the proposed speaking frequency counter system, we have used 20 live streaming video under different lighting condition. Performance evaluation is as shown in Table 1.

Table 1: Results on Still Frames

Total Frames	False Positive	Correct Result	Face Not Found
22000	4400	16000	1600

*Face not found means face was present but could not be identified properly due to unavoidable lighting effect.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: (a)(c)(e)(f) Input Image ; (b)(d)(f)(h) Detected Feature

Some results of the feature detection (lip coordinates detection) are shown in figure 2. Most significant part of the proposed system is detecting the lip region accurately. Sometime it fails, (as g and h show) due to unavoidable lighting conditions. We could use more accurate algorithm for feature extraction the lip contour extraction algorithm, but we prepared our own 3D classifier to ensure the effectiveness of the proposed system in real time. Unpredictable lighting situation is the most challenging factor here, sometimes images are bright, later they become dark and the abrupt changes continue. Developing an algorithm for a static environment is easier as thresholding remains almost unchanged under same lighting feature. Designing a practical real time system with adaptable lighting is a great challenge towards development. We could not but mention the tradeoffs between perfection and time boundary in real time algorithms of pattern recognition. Here, our proposed system contributed a little towards the objectives. Though it does not work well for multi-face environments, we are working forward to find a remedy of that problem. We are working to extend this system as speech recognizer in future.

REFERENCES

1. Paul Kuo, Peter Hillman and John Hannah, “Improved Lip Fitting and Tracking For Model-Based Multimedia and Coding”, International Conference on Visual Information Engineering Conference, Glasgow, UK, pp. 251-258, 2005.

2. Mohammad Sadeghi, Josef Kittler and Kieron Messer, “Segmentation of Lip Pixels For Lip Tracker Initialization”, International Conference on Image Processing, ICIP, IEEE, Greece, 2001.

3. Rainer Stiefelhagen, Jie Yang, Alex Waibel, “A Model based Gaze Tracking System”, Proc. of IEEE International Joint Symposia on Intelligence and Systems, pp. 304-310, Rockville Maryland, 1996.

4. Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon, “Accurate and quasi-automatic lip tracking”, IEEE Trans. Circuits Syst. Video Technology, vol. 14, no. 5, pp. 706-715, 2004.

5. C. Neti, G. Potamianos, J. Luettin, I. Matthews, H.Glotin, and D. Vergyri, Large-vocabulary audio-visual speech recognition:A summary of the Johns Hopkins Summer 2000 Workshop, Proc. Works. Multimedia Signal Process. (MMSP), pp. 619-624, Cannes, France, 2001.

6. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, Recent advances in the automatic recognition of audio-visual speech, Invited, Proceedings of the IEEE, vol. 91, no. 9, pp. 1306-1326, 2003.

7. Robert Kaucic, Barney Dalton, and Andrew Blake, Real-time lip tracking for audio-visual speech recognition applications, Proc. Of the 4th Euro. Conf. on Comp. Vis.,Vol 2, pp376-387, Springer-Verlag,1996.

8. Xiao Zheng Zhang, Charles C. Broun, Russell M. Mersereau and Mark A.Clements, Automatic speechreading with applications to human-computer interfaces, Eurasip Journal on Applied Signal Processing, Vol. 2002, Issue 11, pp 1228-1247.

9. R. Caneel, Social signaling in decision making, in: Master Thesis, 2005.

10. M. Jones, P. Viola, Robust real-time face detection, in: International Journal of Computer Vision, Vol. 57, 2004, pp. 137–154.

11. http://video.nytimes.com/.

12. T. G. Dietterich, Machine learning for sequential data: A review, in: Proc. on Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2002, pp. 15–30.

13. D. H. Wolpert, Stacked generalization, Neural Net- works 5 (2) (1992) 241–259.

14. W. W. Cohen, V. R. de Carvalho, Stacked sequential learning, in: Proc. of IJCAI 2005, 2005, pp. 671–676.

15. J. Friedman, T. Hastie, R. Tibshirani, Additive logis- tic regression: a statistical view of boosting, in: The annals of statistics, Vol. 38, 1998, pp. 337–374.

16. M. Pantic, A. Pentland, A. Nijholt, T. Huang, Human computing and machine understanding of human be- havior: a survey, in: ICMI ’06: Proceedings of the 8th international conference on Multimodal interfaces, ACM, New York, NY, USA, 2006, pp. 239–248.

17. R. Cai, L. Lu, H.-J. Zhang, L.-H. Cai, Highlight sound effects detection in audio stream, in: ICME ’03: Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME ’03), IEEE Computer Society, Washington, DC, USA, 2003, pp. 37–40.

18. N. Campbell, H. Kashioka, R. Ohara, ”no laughing matter”, in INTERSPEECH-2005, 465-468 (2005).

19. K. P. Truong, D. A. van Leeuwen, Automatic discrimi- nation between laughter and speech, Speech Commun. 49 (2) (2007) 144–158.

20. B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pan- tic, Decision-level fusion for audio-visual laughter detection, in: MLMI ’08: Proceedings of the 5th international workshop on Machine Learning for Mul- timodal Interaction, Springer-Verlag, Berlin, Heidel- berg, 2008, pp. 137–148.

21. S. Petridis, M. Pantic, Fusion of audio and visual cues for laughter detection, in: CIVR ’08: Proceedings of the 2008 international conference on Content-based image and video retrieval, ACM, New York, NY, USA, 2008, pp. 329–338.