LIP DETECTION FOR AUDIO-VISUAL SPEECH RECOGNITION IN-CAR ENVIRONMENT

Rajitha Navarathna, Patrick Lucey, David Dean, Clinton Fookes, Sridha Sridharan

Speech, Audio, Image and Video Technology
Queensland University of Technology
GPO Box 2424, Brisbane 4001, Australia
{r.navarathna, p.lucey, d.dean, c.fookes, s.sridharan}@qut.edu.au

ABSTRACT

Acoustically, car cabins are extremely noisy and as a consequence audio-only, in-car voice recognition systems perform poorly. As the visual modality is immune to acoustic noise, using the visual lip information from the driver is seen as a viable strategy in circumventing this problem by using audio visual automatic speech recognition (AVASR). However, implementing AVASR requires a system being able to accurately locate and track the drivers face and lip area in real-time. In this paper we present such an approach using the Viola-Jones algorithm. Using the AVICAR [1] in-car database, we show that the Viola-Jones approach is a suitable method of locating and tracking the driver's lips despite the visual variability of illumination and head pose for audio-visual speech recognition system.

Index Terms- AVASR, AVICAR database, Viola-Jones algorithm

1. INTRODUCTION

There is a strong need to reduce driver distraction as vehicle navigational and other operational systems become more complex. The use of voice recognition technology has the potential to solve the problem by providing voice based control for the operation of such in-car systems. However, the robustness and effectiveness of voice recognition systems in a car environment is still poor, due to the number of environmental factors such as acoustic noise. To overcome this problem visual information of the driver's face, such as lip movement can be used as a secondary source to improve speech intelligibility in adverse conditions.

Visual information from a speakers lip movement is unaffected by acoustic noise. Utilizing this visual information in conjunction with the audio channel has the potential to improve the performance of speech recognition in vehicles. The field of recognizing speech using both audio and visual inputs is known as Audio Visual Automatic Speech Recognition (AVASR) [2].

A significant amount of research has been conducted in the field of AVASR. However, systems have only been examined in unrealistic scenarios. There are few attempts to incorporate the visual modality [3, 4] in real-time system. Recently, one notable attempt has been the work of Libal et. al [4], who developed a real-time system to recognize visual speech activity on low cost embedded platforms. This system uses a camera mounted on the rearview mirror to monitor the driver. It detect face boundaries and facial features, and finally use lip motion clues to recognize visual speech activity.

Lucey et. al [5] presents that the Viola-Jones [6] algorithm can be used to develop an efficient visual front-end system in a clean smart room environment. However, it is of interest to see the robustness in a "real-world" application. In this paper we present an efficient visual front end system which is able to track and locate the driver's face and lip area in a car environment using the Viola-Jones algorithm [6]. Using an efficient implementation we will show that the Viola-Jones approach is a suitable method of locating and tracking the driver's lip despite the visual variabilities of illumination and head pose.

The paper is organized as follows. Section 2 describes the experimental data and the methodology used to develop a visual front end system to detect the face and lip area in a car environment. The results are present in Section 3. Discussion and future work are reported in Section 4.

2. RESEARCH METHODOLOGY

In this section, the research data and the methodology are outlined.

2.1. Viola-Jones Algorithm

The Viola-Jones algorithm is a rapid object detection algorithm. It is based on cascade of weak classifiers instead of a single strong classifier to detect objects. The Viola-Jones algorithm uses the AdaBoost algorithm [7], to develop the classifiers. The main principle of the Viola-Jones algorithm is to scan sub windows within an image to detect objects of interest across an image. It provides a quick and accurate framework, which can be used in real-time object detection applications. Therefore, this research uses the Viola-Jones object detection algorithm to detect the face and lip area. Viola and Jones describe the

978-1-4244-7166-9/10/$26.00 ©2010 IEEE

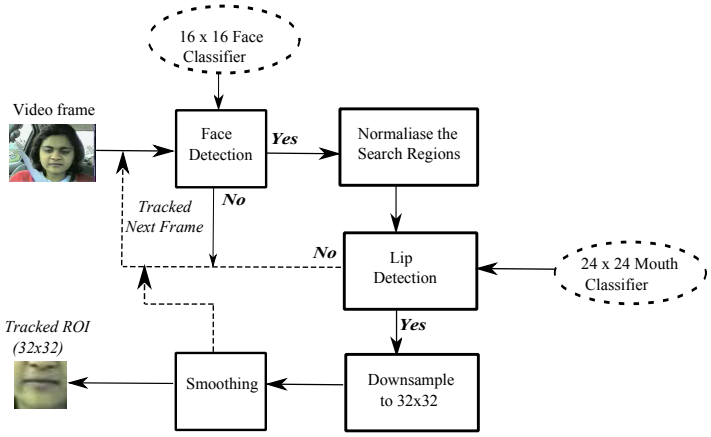

Fig. 1. Block diagram for the visual front end system to detect the face and lip area of a speaker in-car environment

Table 1. Noise Conditions in AVICAR database

Noise	Description
35D	Car travelling at 35mph and windows open
35U	Car travelling at 35mph and windows close
55D	Car travelling at 55mph and windows open
55U	Car travelling at 55mph and windows close
IDL	Car stopped

steps of the algorithm in [6].

2.2. Experimental Data

This section describes the data which were used to perform the experiments. There are several databases which have been developed for AVASR, such as IBM smart-room database [8], CUAVE database [9]. Unfortunately, most of these databases are captured in ideal video conditions. This research uses the AVICAR database, which has the dataset in a "real-world" car environment [1].

The AVICAR database is a publicly available in car speech corpus containing multi-channel audio and video recordings. It was recorded by researchers at the University of Illinois. The database consists of 50 male speakers and 50 female speaker audio and video files. Four cameras and eight microphones were used to capture the audio and video data. Most of the speakers are American English speakers. And the others are from Latin America, Europe, East or South Asia. However, all the recorded speech is in English. Each recording session contains speech under five different driving conditions. The driving conditions are detailed in Table 1.

2.3. Lip Feature Extraction for AVASR System

This section presents the methodology which is used to extract lip features from the speaker in five driving conditions in a car environment. An overview of the visual front-end system is presented in Figure 1. Given a video of a speaker in a vehicular environment, the face classifier is used to find the face. Once the face was located, the mouth classifier was used in the lower part of the isolated face image to extract the mouth region of the speaker. The face and the mouth classifiers were developed as described below.

Fig. 2. Ground truth data points used to derive ROI for facial features on a image. Image from the AVICAR database [1]

The AVICAR database images were categorized into two image categories as testing images and training images. To generate the Region of Interest (ROI), the ground truth points were manually labeled in all the driving conditions in the AVICAR database. These points were left eye, right eye, nose, left corner of the nose, right corner of the nose, right mouth, left mouth, top mouth, bottom mouth, center mouth and chin. Figure 2 gives an example of a ground truth data image.

To train each classifier, positive and negative images were used. The effectiveness and the efficiency of the classifiers depends on the training phase. Therefore, approximately 6000 positive images and around 3000 negative images were used to develop face and the mouth classifiers.

All the positive images for the face classifier were normalized to 16×16 pixels based on the distance between the eyes to increase the speed of the training phase. Figure 3 shows an example of the normalized image. The positive images used to develop the face classifier are shown in Figure 4. The negative images used to train this classifier are chosen from images of environmental scenes. These are simply images with no faces. Example of negative images are shown in Figure 5. The selected negative im-

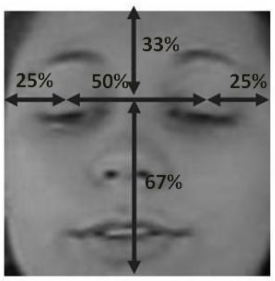

Fig. 3. Normalized face image

Fig. 4. Positive face images used to train the face classifier

ages are high resolution images. Having a high resolution negative image creates abundant background sub windows at the training phase of the Viola-Jones algorithm. This increases the overall performance of the final classifier. Eventhough the use of high resolution images reduces the speed of the training algorithm, this is not significant as most of the negative images are rejected at early stages by the algorithm.

All the positive images for the mouth classifier were normalized to 24×24 pixels. We selected some of the facial part images excluding the lips as the negative images. All the negative images for the lips are high resolution images. The positive and negative images for the lip are shown in Figures 6 and 7 respectively.

After the proper selection of positive and negative images the classifiers were developed using the OpenCV [10] libraries, which are very useful for real-time computer vi-

Fig. 5. Negative face images used to train the face classifier

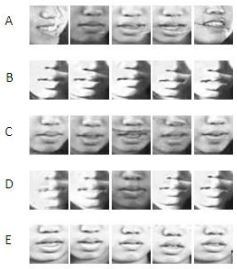

Fig. 6. Example of the positive images used to train the facial classifier (A) Positive images form 35U driving condition (B) Positive images from 35D driving condition (C) Positive images from 55U driving condition (D) Positive images from 55D driving condition (E) Positive images from IDL driving condition

Fig. 7. Example of the negative images used to train the mouth classifier

sion application. The overall visual frond end system for the car environment was developed using Microsoft Visual C++ to detect the face and extract the lip area for AVASR system.

3. EXPERIMENTAL RESULTS

This section describes the results of the experiments performed using the developed system. The experiments were conducted using the AVICAR database, as described in Section 2.2. The classifiers were used to detect the face and lip from the speakers in a car environment. The system was tested in five driving conditions. As depicted in Figure 1, initially the system detects the face of the speaker using the face classifier. Next, the system detects the lip area from the isolated face image. Examples of the detected images in five driving conditions are shown in Figure 8.

The average detection time and the false alarm rate are shown in Table 2. The number of frames per second was 30 in the incoming video. The average detection time is reported in the last column in Table 2. The hit rate is the number of frames the object is detected correctly over the total number of frames. The false alarm ratio gives the fraction of incorrectly detected windows over all detected windows. The hit rate for the face and mouth classifiers are achieved with more than 90% accuracy. The average detection time for the face was high compared to

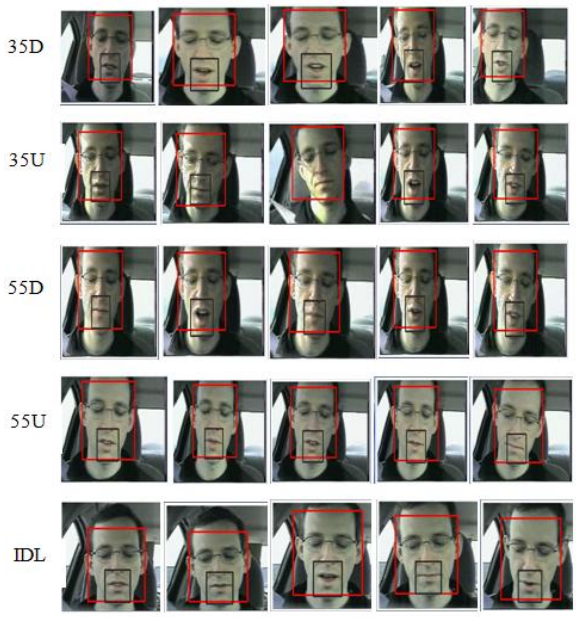

Fig. 8. The resultant frames of a subject in five different driving conditions

Table 2. Overall Results

Feature	Hit Rate/(%)	False Rate/(%)	Average Detection Time/(ms)
Face	96.92	0.94	17.18
Mouth	92.36	26.3	7.95

facial feature (mouth) detection. The main reason for this is need of the face classifier to search the entire frame to detect the face. We were able to reduce the false alarm rate to below 1% using the innovative selection of the negative images at the training phase of the face classifier. The average detection time for mouth region was less due to the small search region compared to the face classifier. However, the false alarm rate of the mouth classifier was higher compared with the false alarm rate of the face classifier. The false alarm rate of the mouth classifier can be reduced by providing a greater variety of negative images to the mouth classifier.

4. DISCUSSION AND FUTURE WORK

Visual lip information from the driver is considered as useful information to improve the robustness and efficiency of speech recognition in a car environment. However, the efficiency of resulting AVASR system depends on the visual front end system. This paper has presented an efficient visual front end system to detect face and lip area using the Viola-Jones algorithm. This research shows that the Viola-Jones approach is a suitable method of locating and tracking the driver's lip area despite the visual variabilities of illumination and head pose that typically occur in a car cabin. We have shown that identifying suitable positive and negative images for the training phase will increase the accuracy of the overall system as well as achieve an improvement in the speed of the training stage. The paper also describes a technique to reduce false alarm rate of facial classifiers by presenting suitable image regions.

Our current work is directed in develop a complete AVASR system using the extracted lips and the audio information, to overcome the problems of using audio only speech recognition in a vehicular environment.

5. ACKNOWLEDGMENT

This work was supported by the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

6. REFERENCES

[1] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "Avicar: Audio-visual speech corpus in a car environment." *In Proc. Interspeech 2004*, Jeju Island, Korea.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *in Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[3] J. Huang, G. Potamianos, J. Connell, and C. Neti, "Audio-visual speech recognition using an infrared headset," *Speech Communication*, vol. 44, no. 4, pp. 83–96, 2004.

[4] V. Libal, J. Connell, G. Potamianos, and E. Marcheret, "An embedded system for invehicle visual speech activity detection," in *Proceedings of the International Workshop on Multimedia and Signal Processing*, Chania, Greece, 2007, pp. 255–258.

[5] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," *in Proceedings of the International Workshop on Multimedia and Signal Processing*, pp. 24–28, 2006.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," vol. 1, 2001, pp. I–511–I–518 vol.1.

[7] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *in Computational Learning Theory: Eurocolt '95*, pp. 23–27, Springer-Verlag, 1995.

[8] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL,USA ,2002.

[9] G. Potamianos and P. Lucey, "Audio-visual asr from multiple views inside smart rooms," *Proc. Int. Conf. Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 35–40, 2006.

[10] *Open Source Computer Vision Library*, Std. [Online]. Available: http://www.intel.com/research/mrl /research/opencv