Facial Feature Detection for In-Car Environment

Rajitha Navarathna and Patrick Lucey
Speech, Audio, Image and Video Smart System Theme
Queensland University of Technology
GPO Box 2424, Brisbane 4001, Australia
{r.navarathna, p.lucey}@qut.edu.au

***Abstract*—Acoustically, vehicles are extremely noisy environments and as a consequence audio-only in-car voice recognition systems perform very poorly. Seeing that 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. However, implementing such an approach requires a system being able to accurately locate and track the driver's face and facial features in real-time. In this paper we present such an approach using the Viola-Jones algorithm. Using this system, we present our results which show that using 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.**

***Index Terms*—Face & facial features; False alarm rate; AVICAR database; Viola-Jones algorithm**

I. Introduction

There is a strong need to reduce driver distraction as in-vehicle navigational and other operational systems become more complex. The use of voice recognition technology has the potential to provide solutions to this problem by providing voice based control for the operation of such in-car systems. However, the robustness and effectiveness of voice recognition systems in car environment is still poor, due number of environmental factors such as acoustic noise. A major benefit of using visual information from a speakers lip movement for speech recognition is that the visual modality is unaffected by acoustic noise. Utilizing this visual information in conjunction with the audio channel has the potential to improve the performance speech recognition in vehicles. The field of recognizing speech using both audio and visual inputs is known as Audio Visual Automatic Speech Recognition (AVASR) [1].

A significant amount of research has been conducted in the field of AVASR. However, systems only been implemented in unrealistic scenarios. There are few attempts to incorporate the visual modality [2], [3] in-car or in real-time system. Recently, the notable attempt has been the work of Libal et. al [3], where they developed a real-time system on low cost embedded platforms which uses a camera mounted on the rear-view mirror to monitor the driver, detect face boundaries and facial features, and finally use lip motion clues to recognize visual speech activity. In 1997, Eriksson et al. [4] presented a visual front-end system which locates and tracks the driver's eyes.

There are several databases which have been developed for use for AVASR, such as CUAVE database [5], IBM smart-room database [6]. Unfortunately, most of these databases are captured in ideal video conditions. The cost of capturing data in realistic environments and storing video data are major problems in the collection of a AVASR databases. The AVICAR [7] is one of the publicly available database which has the datasets in a real-time car environment. The AVICAR database was used in this research.

Over the past decade a plethora of work has been done on face and facial feature detection. This has been an active fields of research in computer vision. In 2001, Viola et al. [8] presented a novel method to object detection, which has the ability to detect the objects in a real-time environment. This research used the Viola-Jones object detection algorithm.

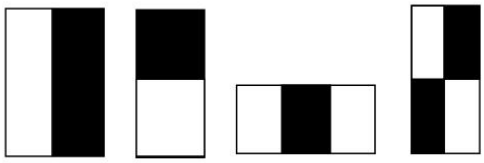

Fig. 1. Basic haar-like features [8]

The outline of the paper is as follows. Section II describes the Viola-Jones algorithm and the main steps of the algorithm in subsections. Section III describes the AVICAR database used for the experiments. Section IV presents a methodology for improve the computational speed of face and facial feature detection and to reduce false alarm rates of Viola-Jones algorithm. The results are present in Section V. Conclusions are reported in Section VI.

II. Viola-Jones Algorithm

The Viola-Jones algorithm [8], is a rapid object detection algorithm proposed by Viola and Jones in 2001. It is based on a boosted cascade of simple classifiers. The main principle of the 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. The main steps of the Viola-Jones algorithm are described in briefly in the following sub section.

A. Feature Representation

The Viola-Jones algorithm uses a "Haar-like" feature representation of the images instead of pixels. The basic type of features are shown in Figure 1 [8]. The original features were extended to fourteen features by Lienhart and Maydt [9] by introducing new features which were rotated by 45°.

B. AdaBoost Algorithm

The Viola-Jones algorithm uses the AdaBoost algorithm [10], which was developed by Freund and Schapire in 1996 to develop the classifier. It is an efficient and effective learning algorithm. The main concept of the algorithm is to produce a strong classifier which has a high detection performance by linearly combining weak classifiers. The steps of the algorithm are briefly described here [8], [11].

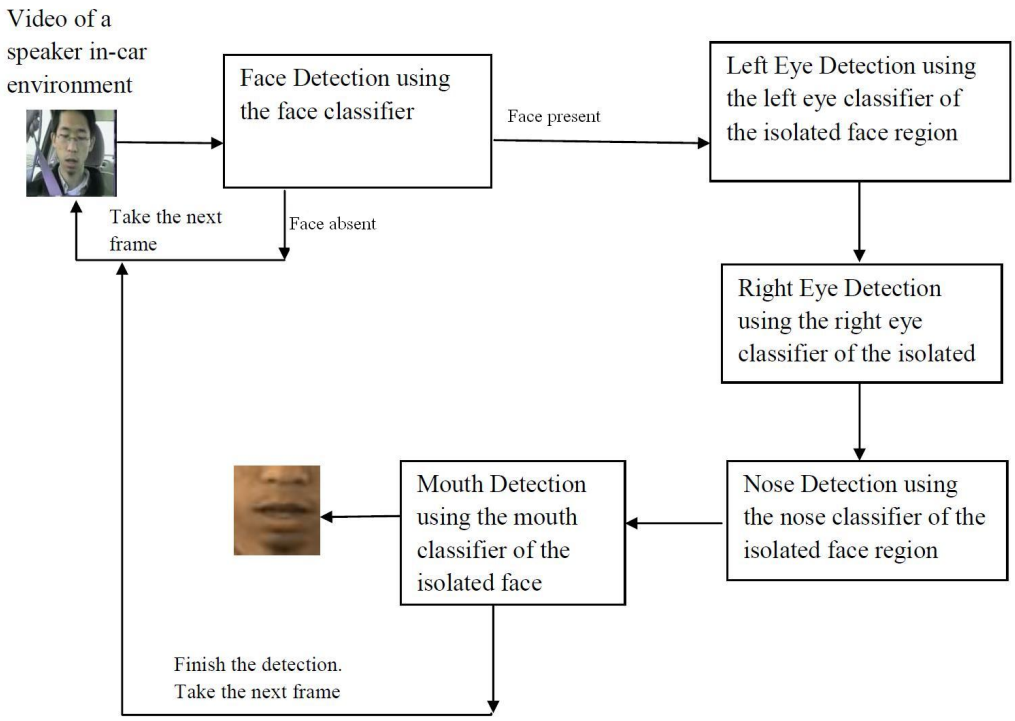

Fig. 2. Block diagram for the visual front end to detect the face and facial features in-car environment

1) **Given n example images** $(x_1, y_1), \ldots, (x_n, y_n)$ **where** x **is the sub-window of the entire image and** $\mathbf{y}_i = 0, 1$ **for negative and positive examples respectively.**
2) **Initialize weights** $\mathbf{w}(1, i) = \frac{1}{2m}, \frac{1}{2l}$**, for** $\mathbf{y}_i = 0, 1$ **respectively, where m is the number of negative examples and l is the number of positive examples.**
3) **For** $\mathbf{t} = 1, \ldots, T$
   - **Normalize the weights at each stage, so that** w_t **is a probability function**

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \tag{1}$$

   - **For each feature** j**, train a classifier** h_j **which is restricted to using a single feature. The error is evaluated with respect to** w_t**,**

$$\epsilon_j = \sum_j w_i |h_j(x_i) - y_i| \tag{2}$$

   - **Choose the classifier with** h_t**, with the lowest error** ϵ_t
   - **Update the weights :**

$$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i} \tag{3}$$

   where $e_t = 0$ **if example** x_i **is classified correctly,** $e_t = 1$ **otherwise, and** $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$
4) **The final classifier is :**

$$h(x) = \begin{cases} -1 & : \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & : otherwise \end{cases} \tag{4}$$

where $\alpha_t = \log \frac{1}{\beta_t}$

C. Cascading the Classifications

Viola et al. proposed the use of a cascade of weak classifiers instead of a single strong classifier to detect objects. The complexity of each stage increases in the cascade. A key innovation in having a cascade of classifiers is that it rejects the majority of sub windows that have no objects at early stages rather than allowing these to go to the complex stages.

III. AVICAR Database

This section describes the data which were used to the experiments. Research was used the AVICAR database, which has the dataset in a real-time car environment.

The AVICAR database is a publicly available in car speech corpus containing multi-channel audio and video recordings [7]. It was recorded by researchers at the University of Illinois. The collection was designed to enable low-SNR (*speech-to-noise*) speech recognition through combining multi-channel audio and visual speech recognition.

The AVICAR database consists of 100 speakers audio and video files (50 male and 50 female). 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 noise conditions. The noise conditions are detailed in Table I. The speech data is of isolated digits, isolated letters, phone numbers and TIMIT sentences, all in English.

TABLE I
NOISE CONDITIONS IN AVICAR DATABASE

Noise	Description
35U	Car travelling at 35mph, windows open
35D	Car travelling at 35mph, windows close
55U	Car travelling at 55mph, windows open
55D	Car travelling at 55mph, windows close
IDL	Engine running, car stopped, windows open

Fig. 3. Ground truth data points used to derive ROI for facial features on a image

IV. FACIAL FEATURE DETECTION METHODOLOGY

This section presents face and facial feature detection strategy that increases detection speed and reduces the false alarm rate of a Viola-Jones based classifier. An overview of the visual front-end system is presented in Figure 2. Given a video of a speaker in car environment, the face classifier is used to find the face. Once the face was located, the location of the left eye and right eye was searched for in the upper part of the isolated face region. The nose classifier was used to locate the nose. Finally, the mouth classifier was used in the lower part of the isolated face image to extract the mouth region of the speaker.

Part of the AVICAR database images were categories into two image categories as testing images and training images. To generate the Region of Interest (ROI), the ground truth values were manually labeled for the part of the face images 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 3 gives an example of a ground truth data image.

To train each classifier, two sets of images were used. In a good face and facial detector, the positive and negative images as well as the size of those images play a significant role. 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 each face and facial classifier. Initially work was conducted to generate a face classifier.

All the positive images for the face classifier were normalized to 16×16 based on the distance between the eyes to increase the speed of the training phase. Figure 4 shows the normalized ratio image. The positive images used to develop the face classifier are shown in Figure 5. The negative images used to train this classifier are chosen from environmental places. These are simply images with no faces. Example of negative images are shown in Figure 6.

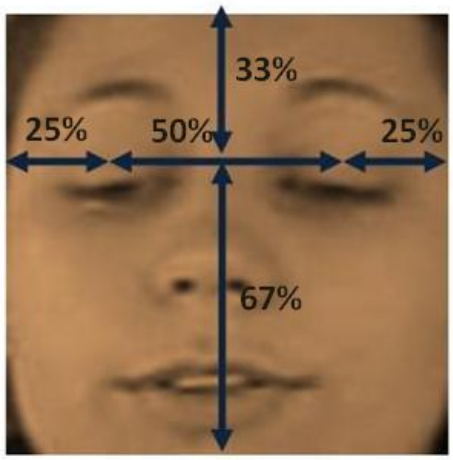

Fig. 4. Normalized face image

Fig. 5. Positive face images used for train the face classifier

The negative images are high resolution images. The reason is that the Viola-Jones algorithm rejects most of the negative images in early stages. Apart form that, having a high resolution negative image creates abundant background sub windows at the training phase of the Viola-Jones algorithm. This reduce the speed at the training phase in the Viola-Jones algorithm, but it increases the overall performance of the final classifier.

Table II shows the normalized image sizes for the left eye, right eye, nose and mouth. An interesting point of generating these classifiers is the selection of negative images. As we know

TABLE II
NORMALIZE IMAGE SIZES FOR THE FACIAL FEATURES

Feature	Normalize image size
Left Eye	20×20
Right Eye	20×20
Nose	20×20
Mouth	24×24

Fig. 6. Negative face images used for train the face classifier

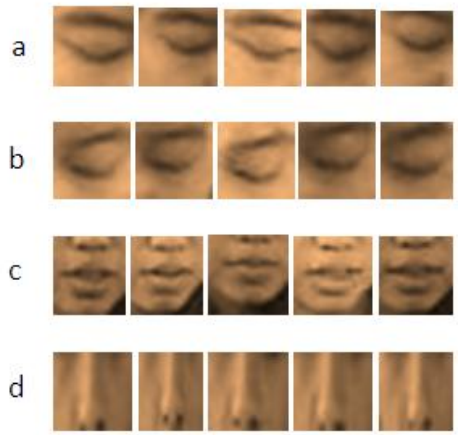

Fig. 7. Example of the positive images used for train the facial classifier (a) Positive images used for train the left eye classifier (b) Positive images used for train the right eye classifier (c) Positive images used for train the mouth classifier. (d) Positive images used for train the nose classifier

the facial features on the face, we selected some of the facial part images as the negative images. For example, if the ROI is the left eye, the negative images selected is the lower part of the face rather than the environmental places as the latter selection would increase the time needed in the training phase as well as increase the false alarm rate. All the negative images for the facial features are high resolution images. Selecting proper negative images for the ROI, showed a good performance of the final classification for that facial feature. The positive and negative images for the facial features are shown in Figures 7 and 8 respectively.

After the proper selection of positive and negative images all the classifiers were developed using the OpenCV (Open Source Computer Vision) [12] libraries, which are very useful for real-time computer vision application.

V. Experimental Results

This section describes the results of the experiments. The experiments were conducted using the AVICAR database. The database was described in Section III. After generating the classifiers, they were used to detect the face and facial features

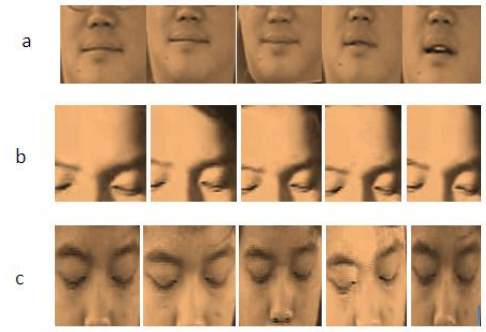

Fig. 8. Example of the negative images used for train the facial classifier (a) Negative images used for train the left eye, right eye classifier (b) Negative images used for train the nose classifier (c) Negative images used for train the mouth classifier

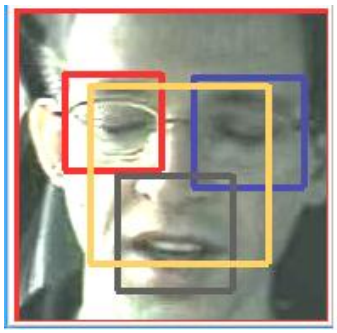

Fig. 9. Diagram of a face and facial feature detection of a AVICAR database image

from the incoming video. As depicted in Figure 2, initially the system detects the face using the face classifier. Next, the system detects the facial features of the isolated face image. As eyes are located on the top part of the face, they further regionalise the detected region from the isolated face. For example the top part used to detect the eyes and the bottom part of the region was used to detect the mouth. Figure 10 shows an example of testing regions for eyes and mouth. As the region is limit to analyze the facial features, it increase the accuracy of the system by reducing the false alarm rate of the captured image. The result of a detected image is shown in Figure 9.

The overall results are shown in Table III. The number of frames per second was 28 in incoming video. The average detection time is reported in the last column in Table III. As

Fig. 10. Example of testing resting regions. (a) Region which used for detect mouth using the mouth classifier (b) Region which used for detect eyes using the left eye and right eye classifier

TABLE III
OVERALL RESULTS

Feature	Hit Rate	False Alarm Rate	Average Detection Time/(ms)
Face	96.92	0.94	17.18
Left Eye	96.63	6.78	8.86
Right Eye	82.75	10.1	12.64
Nose	78.81	23.1	10.19
Mouth	92.36	26.3	7.95

can bee seen , the hit rates for all the classifiers are achieved more than 75%. The hit rate for the face classifier is the highest. However, the average detection time was high compared to facial feature detection. The main reason is, the face classifier searches the entire frame to detect the face. We were able to reduce the false alarm rate to around 1% due to the innovative selection of the negative images used at the training phase to develop the face classifier. The average detection time for facial feature detection were less due to the reduction of the search region. However, the false alarm rate was a little bit higher comapared with the false alarm rate of the face classifer. The reason is the less varieties of negative images. Mainly, the error rate of the feature detection cause by the head movement of the speaker in-car enviroment.

VI. CONCLUSION

This paper has presented an efficient way of detecting faces and facial features using the Viola-Jones algorithm in a car environment. The paper has shown that identifying suitable positive and negative images and selecting suitable image sizes for the positive and negative images at the training phase will increase the performance of the overall system as well as achieve an improvement in speed at the training phase. The paper also describes a technique to reduce the false alarm rate of the facial classifiers by presenting suitable image regions for appropriate facial classifiers.

Face and facial features detection is poor in situation where there is significant head movement causing changes to the pose and illumination. In the future, our research directed towards to overcome this problem in-car environment.

ACKNOWLEDGMENT

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

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] N. Eriksson, M. an Papanikotopoulos, "Eye-tracking for detection of driver fatigue," *Intelligent Transportation System, 1997. ITSC '97., IEEE Conference*, pp. 314–319, Nov, 1997.

[5] 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.

[6] 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.

[7] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "Avicar: An audiovisual speech corpus in a car environment," Jeju Island, Korea, 2004, pp. 2489–2492.

[8] 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.

[9] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, 2002, pp. I–900–I–903 vol.1.

[10] 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.

[11] P. Viola, "Detecting pedestrians using patterens of motion and appearance," *IEEE ICCV*, pp. 734–741, Nice, France, 2003.

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