Trials of the CSIRO Face Recognition System in a Video Surveillance Environment

R.-Y. Qiao, J. Lobb, J. Li and G. T. Poulton
CSIRO Telecommunications & Industrial Physics
Australia
E-mail: Rong-Yu.Qiao@csiro.au

Abstract

SQIS^{TM1} is a face image capturing system combined with a real-time face recognition engine developed within CSIRO. It can automatically capture a face or faces in a video stream and verify against face images stored in a database to inform the operator if a match occurs. The entire system operates on a standard standalone Pentium 4 PC and requires no special hardware apart from a normal video frame grabber. It is capable of operating in real-time at 25 fps with a database of up to 1000 subjects and simultaneously monitoring up to 10 people in a video scene. It is ideal for video surveillance type applications where computer-aided subject recognition is needed for security enhancement. This paper describes the trials of the SQIS^{TM} system undertaken at a number of potential application sites. It shows the system performance in terms of automatic face capture and recognition, as well as the problems identified during the trials.

1 Introduction

In addition to other popular biometric measures such as finger prints and iris scans, the human face is another common biometric feature that can be used for automatic person recognition. It is, however, virtually the only viable biometric for identifying people when cooperation from the subjects is not available. Even though automatic face recognition may never be able to achieve the high level accuracy of a finger or iris-based system, its unique feature of being non-intrusive makes the technology very suitable for open area surveillance. For such applications, automatic face recognition can achieve better performance than human operators, especially when a large number of people is involved.

CSIRO has developed an automatic face recognition system called SQIS^{TM}. It consists of real-time automatic face capture followed by recognition. The system is currently undergoing trials at a number of potential application sites, and aims to gather market information and assess the system performance in various environments. The trials have been very valuable by giving us useful user feedback, enabling us to improve the system in both usability and performance.

2 The SQIS^{TM} System

The schematic diagram of the SQIS^{TM} system is shown in Fig. 1. It has four major parts: a face capture engine, a timeline database for storing all captured faces, a recognition engine and its associated database of enrolled faces. System hardware consists of a standard video frame grabber and a PC. The system can also be configured as a multiple-input system where recognition is conducted on a centralised PC while both face capture and image storing for each video input are performed remotely on a separate PC. Microsoft Access databases are used for both the captured image timeline and the recognition database. All database images are time stamped and saved in compressed JPEG format.

Output from video sources such as camera, VCR,

[1] SQIS is a registered trademark of CSIRO Australia

DVD or TV tuner can be used with the system. Input switching between different input sources is possible, but not desirable for the trial system as automatic updating of system settings has not yet been implemented. If the input video is taken only from a few fixed locations as in the case of fixed-camera surveillance, multiple sets of system settings can be stored in files and loaded when necessary.

Fig. 2 shows the main system interface. The captured image timeline at the bottom of screen represents the captured images in the image store at different scales. It facilitates the image searches and also indicates where a recognition event has happened. When an alarm (recognition) occurs, an alarm bell will ring and the captured image will be displayed together with the matched image in the recognition database for visual verification by system operators.

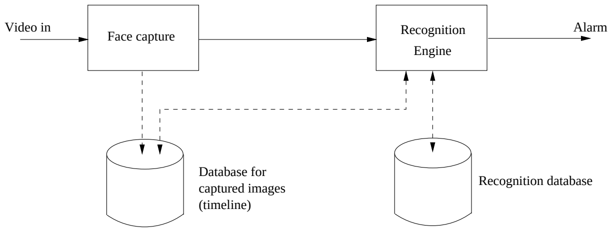

Figure 1. The CSIRO SQIS^{TM} system

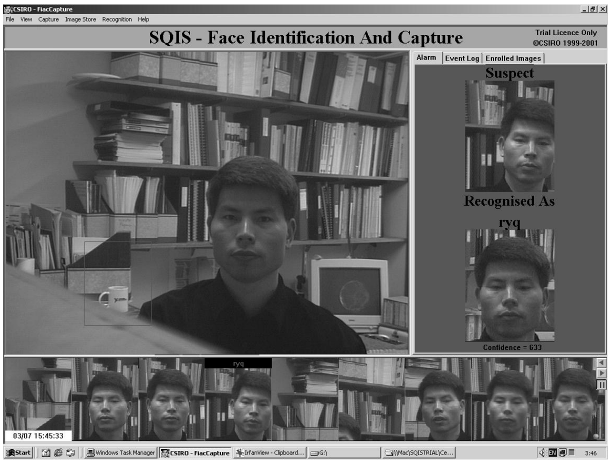

Figure 2. The SQIS^{TM} system interface. This is the main operation screen which consists of live video, timeline showing the captured images and the enrolment/alarm screen.

2.1 Capture Engine

The capture engine relies on attributes such as motion, skin colour and face features to identify faces from a video stream[1]. Multiscale feature checking allows the system to capture faces over a wide range of distances from the camera. Even though the capture engine runs at more than 25 fps, only one face image from each person per second is stored in the timeline database in order to reduce storage requirements. Such images are selected as the closest to a standard frontal face. Different faces are differentiated by their motion and locality in a scene. This minimises the number of people being missed while at the same time avoiding the capture of too many images of the same person.

The combination of motion, colour and facial features works well in an environment where there are no moving background objects and all subjects face the camera. The worst scenario occurs when moving background objects have skin-toned colour and complex texture which results in many non-facial images being captured. To reduce the number of non-faces captured, a number of filters were implemented using various methods such as image similarity and background. Although this scenario causes increased storage because of the larger number of non-faces, the effect on recognition is quite small since the recognition engine is capable of rejecting non-facial images.

An example of faces captured from a video stream is shown in Fig. 3.

2.2 Recognition

Recognition is based on a proprietary method using both global and local facial features. Good performance for both feature types depends largely on having good reference points on faces. In our system, we use the eyes as the reference points because it also gives us a measure of the face size (distance between left and right eyes). Accurately locating eyes is a difficult problem in itself, but we have developed a very fast eye-finder with good performance (up to 95% accuracy depending on image quality).

Good eye positions are defined as those having a pixel error from the actual centre of eye (manually located position) of less than 10% of the distance between left and right eyes. A number of methods are available to further reduce the effect of eye location error. One such method is described in [3].

The system runs in real time on a 1.5 GHz Pentium 4 PC with a Matrox Meteor analogue frame grabber. System performance depends largely on the type of application. This system can simultaneously capture and recognise up to 10 people in real-time at the full frame rate of 25 fps. The system has a built-in higher performance version of recognition which can not operate in real-time because of its high complexity. It activates automatically whenever extra system resources become available to process all the captured images in the timeline.

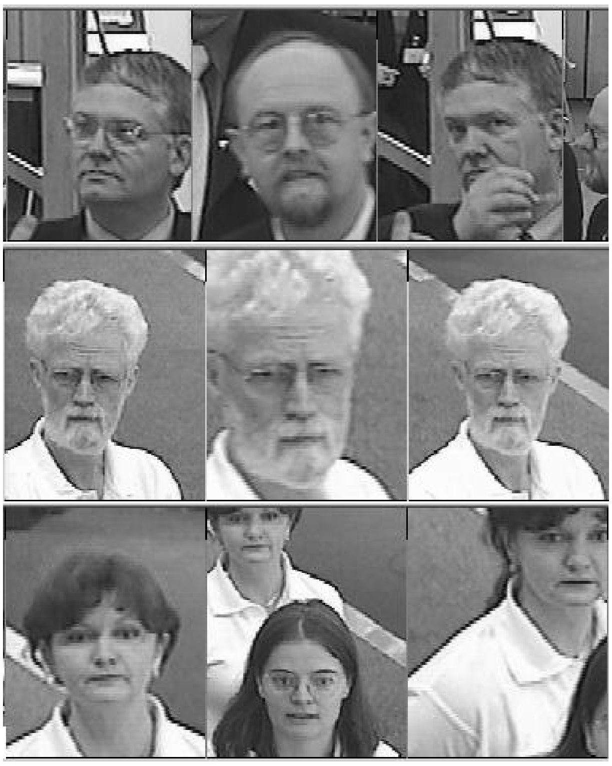

Figure 3. Captured faces from a video scene

3 Operational modes

Capture This mode demonstrates the operation of the face capture engine. All captured face images are displayed beside the live video display. Best images for each individual are also shown on the screen. These are the images sent to the recognition engine if the engine is active. They can also be placed on the timeline so that various recognition tasks like database search or enrolment to the recognition database can be performed. This mode is also useful for system adjustment, because it gives a direct feedback of how the capture engine performs in a particular environment.

There is also a manual capture mode to allow users to grab a face from the screen for recognition or other purposes.

Recognition Captured faces in the timeline are compared with all images in the recognition database in real-time for identification. If an identification event occurs, the system sounds an alarm and displays the pair of matched images and the person's ID. The system can also give a warning when a less confident identification occurs. All events are logged and displayed to enable sorting or searching.

Enrolment Recognition database enrolment can be done using images selected from the captured image store or images imported from other sources such as JPEG photos. There is no restriction on the number of images stored for each individual, but a good representation of the person with different poses, sizes, and lighting variations will significantly increase the recognition rate. For each enrolled image, accurate eye locations are desired for good system performance. Best performance is obtained if the automatically located eye positions are checked by the operator and manually corrected if necessary.

Timeline search For each enrolled individual, the system can perform a search through the whole set of captured images for potential matches. This will list all the matched images in rank order. Operators can check the results and mark the correct images on the timeline. A very useful feature called *bootstrapping* allows users to enrol images picked up from an initial timeline search and repeat the search. This gives improved recognition each time bootstrapping is carried out.

Offline recognition Identification can be also be performed on any previously captured faces. Instead of finding a single best match, the system can rank all the enrolled images according to their recognition distances from the selected image in a descending order. When there are many enrolled faces, this mode helps the user by providing a short-list of possible candidates, and at the same time eliminates possible misses by the real-time recognition when the recognition distance for a correct match is above the recognition threshold.

Setup This mode is used literally for the system setup or when the operation environment (camera, lighting etc) changes. The system provides some default settings for some typical environments such as office or outdoors. Further adjustments are usually necessary for the system to operate effectively. Colour balance, thresholds for various face features and the number of face scales are the main system parameters which can be adjusted.

4 Trials

The trials are being conducted to gather information regarding our system and assess its performance in real world environments. Feedback from users has been valuable for improving the system in terms of performance, usability and functionality.

4.1 Setup

System trials are being conducted simultaneously at the premises of a number of potential customers. We require the users to use the system for their real tasks, give us feedback on performance and usability, and to allow us access to the data. All system parameters and data are logged and collected on a weekly basis.

The trial sites include both indoor and outdoor environments. Outdoor environments have very large lighting variations because of changes in daylight and shadowing by buildings. The camera is usually mounted at a much larger distance than for indoor operations, and is often looking through tinted glass windows. Because of the long range zoom of the camera, the images are often blurred and the view angle of the camera can become so small that subjects are often only visible for a second.

One of the indoor site settings is illustrated in Fig.4. It represents a typical situation of indoor video surveillance. A camera is mounted on a 4m high ceiling, aimed down at an entrance door. The distance from the camera to the door is about 15m and the camera is adjusted so that its field of view is about 2m on each side of the door for an average height person. At one of the indoor sites, the lighting is very stable while on another it is affected by daylight coming through nearby windows.

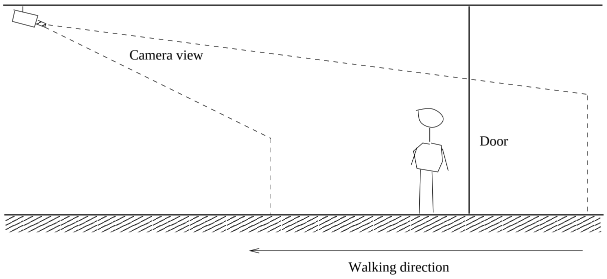

Figure 4. A typical indoor camera setup to monitor an entrance door

4.2 Face capture

The following table shows the performance of the capture engine in the above environment. The relatively low percentage ($\leq$33%) of faces in the captured images shown in the table is due to the complicated background, high camera angle and the operational requirement that the system misses as few faces as possible. Many non-faces are also captured when people are facing away from the camera. Since the capture engine runs at very high frame rate (25 frames/s) and a subject usually stays visible in the camera for more than a few seconds, the probability of failing to capture that person's face is extremely low, unless the face is visible for only a short time. Without stored video, it is extremely difficult to obtain statistics on how many people are missed out, but from our observation and user feedback, the percentage is almost zero.

Correct eye-location rate is around 60% among all the face images. This rate is very good considering the fact that many non-frontal faces are captured because people may walk in any possible directions in front of the camera. Because the system captures faces at 25 frames/s, multiple images are usually captured from each face. Since for surveillance purposes, only one image with good eye locations is required, the correct eye-location percentage over persons is actually much higher. Occasional failures come largely from those persons with only one captured face image.

The redundancy inherent in capturing multiple images for each face helps improve the recognition performance dramatically. In heavy traffic situation where large computing resources are needed to cope with a large number of people, the system will automatically reduce capture frame rate in order to free up necessary resources.

Table 1. Capture Engine Performance

	Indoor #1	Indoor #2	Outdoor
Percentage of faces among all captured images	15–28%	17–33%	7%
Faces with correct eye locations	61–77%	50–58%	30%

4.3 Recognition

Fig. 5 shows the recognition performance for one of the indoor trial sites. The number of false alarms is plotted against the number of correct alarms for different system versions. It is inevitable that when the system is required to correctly recognise more people, the number of wrong recognitions will also increase. At the beginning of the trials, the recognition performance was not satisfactory, but as the trials progressed, suggestions from the users were noted and new ideas developed to improve the performance. As shown by the figure, these measures have helped to improve the recognition rate by more than 50% when equal error rate (shown as a straight line in the figure) is considered.

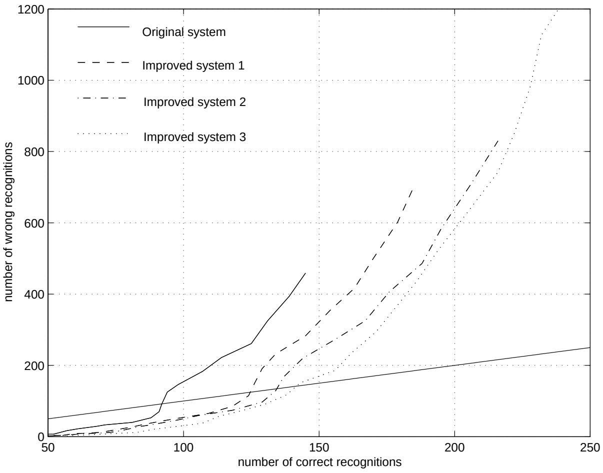

Figure 5. Performance improvement in terms of recognition events during the trials

System performance in terms of face capture and recognition rate depends largely on the working environment. The recognition performance as shown in Fig. 5 also varies with recognition database size and its contents. If the enrolled persons rarely show up, or the database is very big, it is inevitable that the false recognition rate will be high relative to correct recognitions. Other factors include the enrolled image quality, number of enrolled images for each person and their poses. The more information the system has on a person, the more reliable the recognition becomes. A disadvantage of having more images for each person is an increase in false recognition rate. Depending on the application, adjusting the recognition threshold can achieve a desirable balance.

Another way to look at the system performance is to examine false recognition (FP) and false rejection (FN) rate. In order to find the equal FP-FN error rate point, an image database consisting of images captured during the trials was constructed. Images were selected randomly from all the captured image stores, and for each person in the database, multiple images captured at different times were selected. The results are shown in Fig.6 together with results obtained using manually-located eye positions. In terms of false-negative false-positive cross-over rate, the system performance is very close to that with the real eye positions. In real situations, the system operates at a threshold significantly below the cross-over point in order to reduce the number of false positives. It is worth noting that the cross-over errors for surveillance data of this sort are much larger than for images captured under controlled conditions.

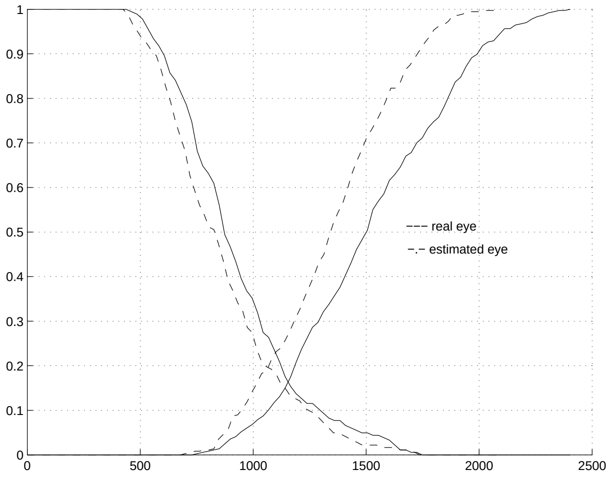

Figure 6. False positive, false negative recognition rates for single image comparison.

Table 2 is a comparison of the performance statistics at the beginning and the end of one of the indoor trials. The FP is shown as a percentage of total captured images, while the FN is a percentage of recognisable persons. Recognition threshold is selected so that the system operates at near-zero-FP point. The table shows that the false rejection rate at the end of trials is reduced quite significantly as compared with the beginning of trials. It is possible to reduce the false rejection rate by operating the system at a higher threshold, but it will be at the expense of a higher false recognition rate. This is shown by the *alarms+warnings* results in the table.

Table 2. Recognition performance (Indoor #1)

	alarms		alarms + warnings	
	FP	FN	FP	FN
Beginning	1.7%	45%	4.4%	40%
End	1.3%	32%	4.5%	25%

5 Conclusion

The trials have proved to be very fruitful. They have helped us to greatly improve the system in terms of usability and performance. User feedback regarding the system has been very positive.

The SQIS^{TM} system was developed specifically for open area surveillance applications. It provides a useful tool for organisations to identify potential criminals or trouble-makers, especially when there are large numbers of people to watch for. Research has shown that human beings are extremely good at recognising friends or known persons, but extremely bad at recognising strangers[4]. When there are more than a couple of faces to remember, a computer-based recognition system is capable of out-performing human operators. This has been confirmed by our trial users.

The system would be ideal for applications such as immigration/customs control, customer check-ins in airports, in casinos, clubs, sporting venues and for police, media, bank, schools and retail shops, etc.

References

[1] D. G. Geers, J. T. Lobb and G. T. Poulton, “Automatic Face Location in an Open Environment: Face-in-a-Crowd”, pp.73-77, DICTA’99, 7-8th December 1999, Perth, Western Australia

[2] G. T. Poulton, “Optimal Feature Sets for Face Recognition”, *Proceedings of IASTED Int. Conf. on Signal Processing & Communications, Feb.11-14, 1998, pp.269-272*

[3] J.Li, “Robust Face Recognition Using Multiple Eye Positions”, (this proceeding)

[4] G. T. Poulton, “Face Recognition in Open Environments: Developments in CSIRO”, IS-IMVSP