Cue Combination for Robust Real-Time Multiple Face Detection at Different Resolutions*

M. Castrillón-Santana, O. Déniz-Suárez, C. Guerra-Artal, and J. Isern-González

IUSIANI Universidad de Las Palmas de Gran Canaria
Spain
{mcastrillon, cguerra, odeniz, jisern}@iusiani.ulpgc.es

1 Introduction

The face detection problem, defined as: *to determine any face -if any- in the image returning the location and extent of each* [Yang et al., 2002], seems to be solved, according to some recent works [Schneiderman and Kanade, 2000] [Viola and Jones, 2001]. Particularly for video stream processing, these approaches focus the problem in a monolithic fashion, forgetting elements that the human system employs: temporal and contextual information, and cue combination.

The work summarized in this abstract describes an approach for robust real-time multiple face detection which combines different cues. The resulting approach achieves better detection rates for video stream processing and cheaper processing costs than outstanding and public available face detection systems.

2 The Face Detection Approach

The approach briefly described makes use for the first detection or after a failure, of two window shift detectors based on the general object detection framework described in [Viola and Jones, 2001], which provide acceptable performance and processing rates. These two brute force detectors, recently integrated in OpenCV, are the frontal face detector described in that paper, and the local context based face detector described in [Kruppa et al., 2003]. The last one achieves better recognition rates for low resolution images and non frontal faces whenever the head and shoulders are visible.

The exclusive use of a monolithic approach based on the Viola framework has the disadvantage of not using a main cue needed for video processing: temporal coherence. Any face detected in a frame provides information which can be used in the next frames to speed up the process. Therefore, for each detected face,

* Work partially funded by research projects Univ. of Las Palmas de Gran Canaria UNI2003/06 and Canary Islands Autonomous Government PI2003/160 and PI2003/165.

the system stores not only its position and size, but also its average color using red, green normalized color space.

Skin color based approaches for face detection have a lack of robustness for different conditions. A well known problem is the absence of a general skin color representation for any kind of light source and camera [Storring et al., 2001]. However, the skin color extracted from the face previously detected by the frontal face Viola detector can be used to estimate facial features position by means of the color blob, which provides valuable information to detect eye positions for frontal faces [Castrillón Santana et al., 2003].

Each face detected in a frame can be characterized by different features $f = \langle pos, size, red, green, leye_{pos}, leye_{pattern}, reye_{pos}, reye_{pattern}, face_{pattern} \rangle$. These features direct different cues in the next frame which are applied opportunistically in an order based on the computational cost and the reliability.

- Eye tracking: A fast tracking algorithm [Guerra Artal, 2002] is applied in an area that surrounds previously detected eyes, if available.
- Face detector: The Viola-Jones detector is applied in an area that covers the previous detection [Viola and Jones, 2001].
- Local context face detector: If previous techniques fail, it is applied in an area that includes the previous detection [Kruppa et al., 2003].
- Skin color: Skin color is searched in the window that contains the previous detection, and the new sizes and positions coherently checked.
- Face tracking: If everything else fails, the prerecorded face pattern is searched in an area that covers previous detection [Guerra Artal, 2002].

These techniques are applied until one of them finds the face, or the process will be restarted using the Viola-Jones based detectors applied to the whole image. Whenever a face is detected, the skin color is used for facial features detection [Castrillón Santana et al., 2003].

The approach extracts different features from each detected face in a frame, therefore multiple face detection is considered. Additionally, a single or multiple faces detected in consecutive frames are related according to their specific features. During the video stream processing, the face detector gathers a set of detection threads, $IS = \{dt_1, dt_2, ..., dt_n\}$. A detection thread contains a set of continuous detections, i.e. detections which take place in different frames but are related by the system in terms of position, size and pattern matching techniques.

The Viola-Jones based detectors have some level of false detections. For that reason a new detection thread is created only if the eyes are detected. The use of the weakest cues, i.e. color and tracking, after a recent detection is reserved to detections which are already considered part of a detection thread. In this way, spurious detections do not launch cues which are not robust enough, in the sense that they are not able to recover from a false face detection. The final results is that for each detection thread, the face detector system provides a number of facial samples, $dt_p = \{x_1, ..., x_{m_p}\}$, which correspond to those detections for which also the eyes were located.

3 Results

For still images the lower boundary is the combination of the Viola-Jones based detectors [Viola and Jones, 2001] [Kruppa et al., 2003]. If both detectors return a face in the same position, it is preferred the frontal face detection as it is more precise. For still images the added value of the approach is the likely eye detection for almost frontal views and the combination of two Viola-Jones based detectors, see Figure 1.

Fig. 1. Detection examples from some CMU database [Schneiderman and Kanade, 2000] samples. The color indicates the technique used: green means that the eyes were detected, yellow means that they were not detected, and red containing a yellow rectangle means that the local context detector was used. The images have been scaled down to fit the extended abstract size. The size of the images are originally 814×820, 256×256, 611×467 and 159×160 respectively in the first row and in the second 500×500, 539×734, 336×484 and 258×218. Obviously for still images there are no detections based on color or tracking.

The benefits of our approach are evidenced in video stream processing. The results achieved processing a set of 70 desktop sequences containing more than 35000 images of different individuals allow, for typical webcam resolutions 320×240, multiple face detection in real-time, an average of 20 fps. The approach provides better recognition rates than the OpenCV implementation of the Viola frontal face detector (3 percentage points greater), faster, and with the added value of eye detection for almost frontal faces. The approach is also suitable for sequences with resolution changes, see Figure 2.

The temporal coherence applied to video stream processing not only speeds up the process but also allows to despise most false detections which appear when a still image is processed, see some false detections in Figure 1. In at least 10 of the sequences some detections were non face patterns, and they were correctly not assigned to any detection thread as the eyes were not found and their position, or their color and size were not coherent with any active detection thread. Or in the worst case, a non face detection was associated to a detection thread, but the system observed soon an incoherence and decided to remove the detection thread and wait for a new one, i.e. a new eye pair detection.

As a conclusion, the approach developed provides good and real-time performance for multiple face detection at different resolutions, making the system suitable for further processing in the field of perceptual user interfaces.

Fig. 2. Frames extracted from a video stream with 720×576 resolution. The color has the same meaning than in Figure 1, but observe that the last frame depicts a blue rectangle which means that tracking was used.

References

[Castrillón Santana et al., 2003] Castrillón Santana, M., Hernández Tejera, F., and Cabrera Gámez, J. (2003). Encara: real-time detection of frontal faces. In *International Conference on Image Processing*, Barcelona, Spain.

[Guerra Artal, 2002] Guerra Artal, C. (2002). *Contribuciones al seguimiento visual precategórico*. PhD thesis, Universidad de Las Palmas de Gran Canaria.

[Kruppa et al., 2003] Kruppa, H., Castrillón Santana, M., and Schiele, B. (2003). Fast and robust face finding via local context. In *Joint IEEE Internacional Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*.

[Schneiderman and Kanade, 2000] Schneiderman, H. and Kanade, T. (2000). A statistical method for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Storring et al., 2001] Storring, M., Andersen, H. J., and Granum, E. (2001). Physics-based modelling of human skin colour under mixed illuminants. *Robotics and Autonomous Systems*.

[Viola and Jones, 2001] Viola, P. and Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*.

[Yang et al., 2002] Yang, M.-H., Kriegman, D., and Ahuja, N. (2002). Detecting faces in images: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58.