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, Edif. Ctral. del Parque Científico Tecnológico,
Universidad de Las Palmas de Gran Canaria, 35017, Spain
mcastrillon@iusiani.ulpgc.es

Abstract. This paper describes a face detection system conceived to process video streams in real-time. Cue combination allows the system to tackle the temporal restrictions achieving a notable detection rate. The system developed is able to detect simultaneously at different resolutions multiple individuals building a feature based model for each detected face.

1 Introduction

If Human Computer Interaction (HCI) were more similar to human to human communication, HCI would be non-intrusive, more natural and comfortable for humans [8]. Human beings are sociable and communicate not only with words but with sounds and gestures. In this context, the human face is a main information channel during the communication process.

The face detection problem is a revisited topic in the literature, and it is commonly defined as: *to determine any face -if any- in the image returning the location and extent of each* [3,13]. According to some recent works [6] [9] [11] the problem seems to be solved. However, those detectors focus the problem using approaches which are valid for restricted face dimensions, and with the exception of the first reference, to a reduced head pose range. Particularly for video stream processing, they try to solve the problem in a monolithic fashion, neglecting elements that the human system employs: temporal and contextual information, and cue combination.

Section 2 describes our approach for robust real-time multiple face detection based on the combination of different cues. The resulting approach achieves better detection rates for video stream processing and cheaper processing costs than outstanding and publicly available face detection systems, as suggested by different experiments in Section 3.

[⋆] Work partially funded by research projects Univ. of Las Palmas de Gran Canaria UNI2003/06, UNI2004/10 and UNI2004/25, Canary Islands Autonomous Government PI2003/160 and PI2003/165 and the Spanish Ministry of Education and Science and FEDER funds (TIN2004-07087).

R. Moreno Díaz et al. (Eds.): EUROCAST 2005, LNCS 3643, pp. 398–403, 2005.

2 The Face Detection Approach

The system developed assumes that a single technique alone can not solve the problem properly for any circumstance. Therefore, our approach combines different single techniques which are not robust and fast enough individually, but as a whole they outperform any individual approach included in terms of correct detection rate and processing cost.

The basic system working makes use of an exhaustive face detection approach, which allows the system to extract different features of the face detected. Those features are used considering temporal coherence in the next frames, obviously only after a detection, reducing the processing cost which any exhaustive approach would have if it were applied to every frame.

On the one side, the approach makes use for the first detection or after a failure, of two window shift detectors based on the general object detection framework described in [11], which provide acceptable performance and processing rates for their particular context although an exhaustive search is performed. These two brute force detectors, recently integrated in OpenCV computer vision library [4], are the frontal face detector described in [11], and the local context based face detector described in [5]. The last one achieves better recognition rates for low resolution images and non frontal faces whenever the head and shoulders are visible.

On the other side, as we mentioned above, the approach extracts different features from each detected face in a frame, therefore multiple face detection is considered. Indeed 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, the system stores not only its position and size, but also its average color using red, green normalized color space. Those features are certainly useful useful to speed up the process, e.g., it is used to define a Window of Attention where the previous detection will likely be, or if the face size is big enough to be worth the application of the object centered detector. In any case, this information is valuable to reduce the time consumption.

Among the different features, the skin color is a valid cue extensively used in the literature. 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 [10]. 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 [1].

Additional features have been considered in order to develop a more robust system. Each face in a frame is 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 [2] 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 [11].
- Local context face detector: If previous techniques fail, it is applied in an area that includes the previous detection [5].
- Skin color: Skin color, defined using red-green normalized color space [12], 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 [2].

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 [1]. Obviously, if there were no recent detection, there is no face model active, and therefore the object-centered and local context detectors are applied sequentially to the whole image.

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 located too. 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.

The resulting system is able to manage in real-time complex scenes in which the human face experiences large scale, pose and appearance transformations. Each specialized detector is described in more detail below.

3 Experiments

The system here described has been applied to still images and video streams. For still images the lower boundary is the combination of both Viola-Jones based detectors performances [5,11]. If both detectors return a face in the same position, it is preferred the frontal face detector data 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.

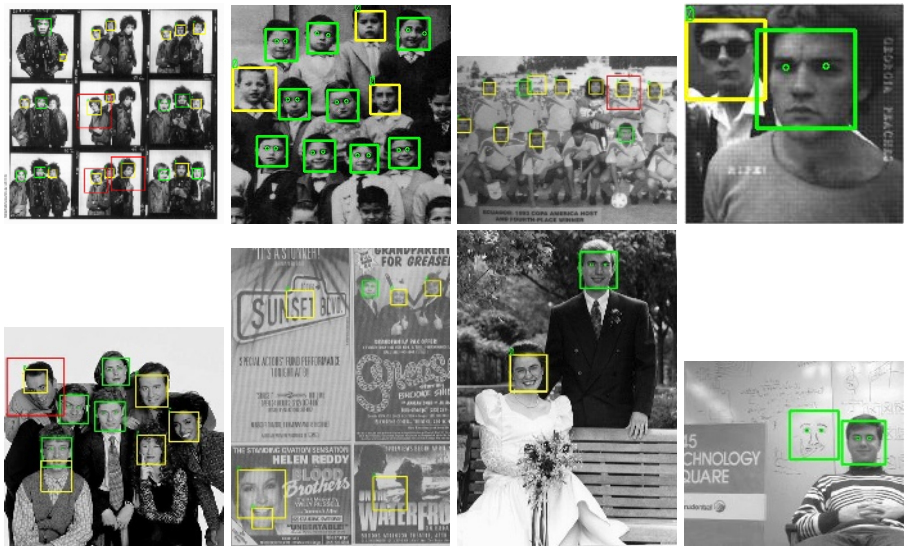

Fig. 1. Detection examples for some CMU database [9] samples. Color indicates the technique used: green means that the eyes were detected, yellow means that they were not detected, and red containing a yellow rectangle means that the local context detector was used. The images have been scaled down to fit the paper 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 skin color or tracking.

The benefits of our approach are evidenced in video stream processing. 70 desktop sequences containing more than 30000 images of different individuals for typical webcam resolutions 320 × 240, were processed at an average rate of 20 fps providing therefore multiple face detection in real-time. In Table 1, it is observed that the detection rates of the approach are slightly better than those achieved by the OpenCV implementation of the Viola frontal face detector [7] (3 percentage points greater). However, those results are obtained five times faster, and with the added value of correct eye detection for more than 60% of the faces detected. The approach is also suitable for sequences with resolution changes, see Figure 3.

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 analyzed 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.

Table 1. Results for face and eye detection. TD reflects correct detection ratio and FD means false detection ratio.

	Viola		Our approach	
	TD	FD	TD	FD
Faces	90.1%	8.2%	92.9%	8%
Eyes	0.0%	-	64.3%	3.7%
Proc. time	117.5 msecs.		21 msecs.	

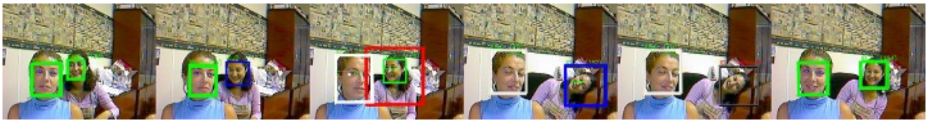

Fig. 2. From left to right: 1) Both faces are detected and their eyes, 2) the Viola based detectors failed detecting the right face, it is detected by tracking the face pattern, 3) the left face is detected using skin color and the right one by means of the local context face detector, 4) the same for the left face, the right one is found by tracking, 5) face pattern tracking is not allowed to be the only valid cue for many consecutive frames, so the right face detection thread is considered missed, and 6) the right face recover its vertical position and fused with the latent detection thread.

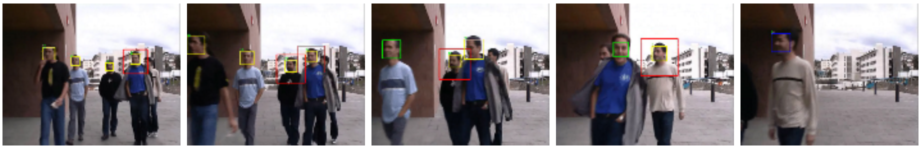

Fig. 3. Frames extracted from a video stream with 720 × 576 resolution. The color has the same meaning than in Figure 1, but observe that the last frame depicts a blue rectangle which means that tracking was used.

4 Conclusions and Future Work

We have developed a face detection system which provides robust multiple face detection at frame rate using a standard webcam. The system has been tested with 70 sequences containing around 30000 images achieving higher detection rate (aprox. five times faster) than the Viola-Jones based face detector, and providing additionally the location of the eyes for more than 60% of the images.

The results achieved processing video streams have been possible thanks to the integration of different cues and particularly the temporal coherence. The average processing time of 21 msecs. reported by the system, makes it suitable for further use in the field of perceptual user interfaces.

References

1. M. Castrillón Santana, F.M. Hernández Tejera, and J. Cabrera Gámez. Encara: real-time detection of frontal faces. In *International Conference on Image Processing*, Barcelona, Spain, September 2003.
2. Cayetano Guerra Artal. *Contribuciones al seguimiento visual precategórico.* PhD thesis, Universidad de Las Palmas de Gran Canaria, Octubre 2002.
3. Erik Hjelmas and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.
4. Intel. Intel Open Source Computer Vision Library, b4.0. www.intel.com/research/mrl/research/opencv, August 2004.
5. Hannes Kruppa, Modesto Castrillón Santana, and Bernt Schiele. 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)*, pages 157–164, October 2003.
6. Stan Z. Li, Long Zhu, ZhenQiu Zhang, Andrew Blake, HongJiag Zhang, and Harry Shum. Statistical learning of multi-view face detection. In *European Conference Computer Vision*, pages 67–81, 2002.
7. Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002*, volume 1, pages 900–903, September 2002.
8. Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 107–119, January 2000.
9. Henry Schneiderman and Takeo Kanade. A statistical method for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1759, 2000.
10. Moritz Storring, Hans J. Andersen, and Erik Granum. Physics-based modelling of human skin colour under mixed illuminants. *Robotics and Autonomous Systems*, 2001.
11. Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, pages 511–518, 2001.
12. Christopher Wren, Ali Azarrbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.
13. Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.