Multiple face detection at different resolutions for perceptual user interfaces

M. Castrillón-Santana, J. Lorenzo-Navarro, O. Déniz-Suárez, J. Isern-González, and A. Falcón-Martel

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

Abstract. This paper describes in detail a real-time multiple face detection system for video streams. The system adds to the good performance provided by a window shift approach, the combination of different cues available in video streams due to temporal coherence. The results achieved by this combined solution outperform the basic face detector obtaining a 98% success rate for around 27000 images, providing additionally eye detection and a relation between the successive detections in time by means of detection threads.

1 Introduction

People detection is a basic ability to be included in any Vision Based Interface [14] in order to use computer vision technology to perceive the user in a Human Computer Interaction (HCI) context. Among the different approaches for this purpose, face detection has been a revisited topic in the recent literature.

The face detection problem, defined as: *to determine any face -if any- in the image returning the location and extent of each* [18], seems to be solved, according to some recent works [9, 11, 16]. 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 presented in this paper describes a real-time vision system which goes beyond traditional still image face detectors. The resulting system is an approach for robust multiresolution real-time multiple face detection which combines different cues based on an obvious connection that exists between frames, i. e. temporal coherence. The resulting approach achieves better detection rates for video stream processing and cheaper processing costs than outstanding and public available face detection systems.

1.1 Previous work

Face detection methods are classified according to different criteria as recent face detection surveys do [5, 18]. In our opinion these techniques can be classified into two main families according to the information used to model faces:

– Pattern based or Implicit: These approaches work by searching exhaustively a previously learned pattern at every position and different scales of the whole input image.
– Knowledge based or Explicit: These approaches increase processing speed by taking into account face knowledge explicitly, exploiting and combining cues such as color, motion, face and facial features geometry, and appearance.

Recent window shift based approaches, i.e. pattern based, have achieved impressive results applied even to video streams [9, 11, 16]. However, the exclusive use of a monolithic approach has the disadvantage of despising a main cue useful for video processing: temporal coherence. Any face detected in a frame provides valid information which can be used to speed up the process in the next frames.

2 The Face Detection Approach

Our approach is related to both categories described in the previous section, as it makes use of both implicit and explicit knowledge to get the best of each one. The explicit knowledge is based on the face geometry and the descriptors extracted from a detection: color and appearance. On the other side, the implicit knowledge is integrated using the general object detection framework integrated in the Open Computer Vision Library (OpenCV) [6]. This framework is based on the idea of a boosted cascade classifier [16] but extends the original feature set and provides different boosting variants for learning [10]. The framework combines increasingly more complex classifiers in a cascade, allowing background regions of the image to be quickly discarded while spending more time on promising object-like regions.

The face detection approach here described has two different working modes depending on recent face detection events reported:

After no detection: This working mode takes place at the beginning of an interaction session, when all the individuals are gone from the field of view, or if nobody is detected for a while. The approach basically makes use of two window shift detectors based on the general object detection framework described in [16]. These two brute force detectors, integrated in the last OpenCV release [6], are the frontal face detector described in that paper, and the local context based face detector described in [8]. The last one achieves better recognition rates for low resolution images if the head and shoulders are visible. The respective minimum size searched are 24×24 and 20×20 pixels. In order not to waste processing time, the detectors are executed alternatively.

For any face detected, the system tries to detect its facial features assuming that it is a frontal face, and therefore its facial features would verify some geometric restrictions. The current implementation searches only the eyes, using a process similar to the one employed in [1] just for a single face detection approach. It was however improved by the addition of different alternatives for eye detection as described below:

1. *Skin blob detection:* Once a face is detected, its skin color is modelled using red-green normalized color space [17], considering just the center of the face container provided by any of the Viola-Jones based detectors. The system heuristically removes elements that are not part of the face, e.g. neck, and fits an ellipse to the blob in order to rotate it to a vertical position [12].
2. *Eyes location:* At this point, the approach searches eye candidates in the likely areas inside the skin blob considering that the face detected is a frontal face. Different candidate pairs are checked for their appearance until one of them, is accepted. The cues used for this purpose are:
   (a) *Dark areas:* Eyes are particularly darker than their surroundings [2].
   (b) *Viola-Jones based eye detector:* As the eye position can be roughly estimated and therefore restricted, a Viola-Jones based eye detector provides very fast results. The detector searches eyes with a minimum size of 16×12 pixels. For small faces, they are scaled up before performing the search.
   (c) *Viola-Jones based eye pair detector:* If other cues fail, the eye pair detection can provide another estimation for eye positions. The minimum pattern size searched is 34×8.
3. *Normalization:* Eye positions, if detected, provide a measure to normalize the frontal face candidate. The normalization step allows further face processing modules to reduce the problem dimensionality.
4. *Pattern Matching Confirmation:* Once the likely face has been normalized, its appearance is checked in two steps making use of Principal Component Analysis (PCA) spaces [7]. The PCA spaces were built using a face dataset of 4000 facial images extracted from internet and annotated by hand.
   (a) *Eye appearance test:* A certain area (11×11) around both eyes in the normalized image is projected to a PCA space and reconstructed. The reconstruction error [4] provides a measure of its eye appearance, and can be used to identify incorrect eye detections.
   (b) *Face appearance test:* A final appearance test applied to the whole normalized image. The image is first projected to a PCA space, and later its appearance is tested using a Support Vector Machine (SVM) classifier [15].

After recent detection(s): As briefly mentioned above, for each detected face, the system stores not only its position and size, but also its average color using red-green normalized color space [17], and the patterns of the eyes (if detected) and the whole face. Thus, a face is characterized by $f = \langle pos, size, red, green, leye_{pos}, leye_{pattern}, reye_{pos}, reye_{pattern}, face_{pattern} \rangle$. These features direct different cues in the next frames which are applied opportunistically in an order based on their computational cost and reliability.

- Eye tracking: A fast tracking algorithm [3] is applied in an area that surrounds previously detected eyes, if available. The tracker makes use of a fixed pattern size for both eyes, 24×24, and searches the minimum difference in the search area as follows:

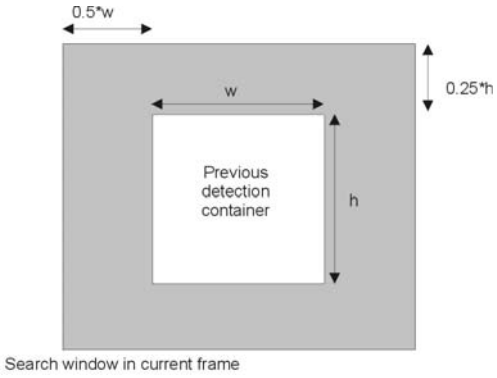

Fig. 1. The search area used for each detected face in the next frame is defined as an expansion of the previous face detection container.

$$D(u,v) = \sum_{Area} |I(u+i, v+j) - P(i,j)| \tag{1}$$

Eye patterns are previously saved with the first detection, and updated according to the strategies described in [3], i.e. only if there is a notorious change in relation to the original pattern, and this difference could confuse the tracker with any other pattern of the close context. If the difference reported is too big, the pattern will be considered lost.

- Basic face detector: The Viola-Jones face detector [16] searches faces but only in an area that covers the previous detection, see Figure 1. This strategy significant reduces processing time.
- Local context face detector: If previous techniques fail, the local context based face detector is applied in an area that includes the previous detection [8], see Figure 1.
- Skin color: The integration of other cues, likely weaker, help to improve the final system performance and robustness. Skin color based approaches for face detection have the 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 [13]. However, the skin color extracted from the face previously detected by the Viola detector can be used to estimate facial features position by means of the color blob, as described above. If previous cues fail, the modelled skin color is used to locate the face, and therefore it is searched in the window that contains the previous detection, see Figure 1. The new sizes and positions are coherently checked, due to the fact that the skin color container is not allowed to experiment large size changes just to avoid an incorrect color updating mechanism.

– Face tracking: If everything else fails, the prerecorded face pattern is searched in an area that covers previous detection [3], see Figure 1. The tracking pattern has a fixed size, for that reason the system scales down the face to fit it in the pattern size. The scale ratio is stored and later used if necessary to scale down the search area in the next frame. This action helps reducing the tracking shift problem. However, the tracking is not allowed to be the only valid cue for more than some consecutive frames in order to avoid tracking problems. Instead, the other cues should confirm the human presence, from time to time, or the person will be considered lost.

For each previous detection, these techniques are applied until one of them finds a new face coherent with the previous detection. Whenever a face is detected, and its eyes were not tracked, the skin color is used for facial features detection as explained above for the *After no detection* working mode. Also, every third frame one of the Viola-Jones based detectors is applied to the whole image in order to detect new faces. Those new faces are compared with those already detected by temporal coherence and those which are redundant removed. If no faces are detected for a while, the process switches to the default *After no detection* working mode.

The approach described considers the possibility of multiple face detection, as no restriction is imposed in that sense. It is interesting to relate the detection information achieved in the consecutive frames, especially when multiple individuals are present. 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. Thus, 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 Viola-Jones based detectors have some level of false detections. For that reason a new detection thread is created only if the eyes have been also detected. The use of color and tracking cues 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.

Ideally a detection thread contains samples detected from a single individual. However, different detection threads can correspond to the same individual, aspect which is not checked by the current implementation. Gaps are allowed during detection thread life, but a detection thread is considered lost if after a predefined number of frames it is not correctly associated to a new detection.

3 Performance Results

For static images the approach provides a performance which combines the results achieved for the standard Viola-Jones face detector [16] and the local con-

Fig. 2. Different samples of some sequences.

text based face detector [8]. We refer the reader to those works to get precise information for static images results.

The strength of our approach is exploited in video stream processing thanks to cue integration. 70 sequences, see Figure 2, corresponding to different individuals, cameras and environments with a resolution of 320×240 were recorded and processed. The total set contains 27271 images, presenting all of them a face easily detected by a human. The average processing time of 60 msecs. using a PIV 2.2Ghz, allowed the system to associate 26875 (98.5%) detections to a detection thread, see Figure 3. As described in that figure, some of those detections are not provided by the Viola-Jones based detectors, but by the cue integration approach. From those detections, their eyes were also located in 70% of them. It must be observed that eyes are located only for frontal poses in the current implementation.

At least 10 of those sequences reported detections which correspond to non face patterns. These detections were correctly not assigned to any detection thread as the eyes were not found and their position, color and size were not coherent with any active detection thread.

Only for 3 (4%) sequences with a single individual, the detection thread was not unique. In these sequences this was due to the fact that at a certain point a detection thread was incorrectly fused with an erroneous detection provided by the Viola-Jones based detectors. However, in all the cases the detection thread was shortly considered lost, and therefore some frames later the still present face was newly detected, and a new detection thread created.

For single individuals sequences this is an impressive result considering the large changes in pose experimented in many of the sequences. The processing rates achieved make the system suitable for further processing in the field of perceptual user interfaces.

For multiple individuals sequences, the system needs more time as more faces are tracked simultaneously, in our experiments the processing time is increased around 20 msecs. per. This effect can be reduced by decreasing the number of times per second that new faces are searched in the whole image, due to the fact that two faces cover more area and therefore it is less likely the presence of a new face. It must also be noticed that in these sequences as no appearance cue is used to relate a detection in the next frame with a previous one, the system is not currently able to manage coherently a situation when different detection threads can overlap, i.e., there is occlusion. It is not sure that after the occlusion

between two individuals, the detection threads will be properly assigned to the new detections.

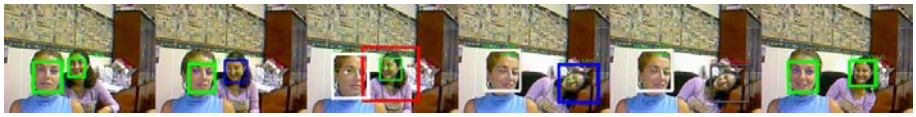

Fig. 3. From left to right: 1) Both faces and their eyes are detected, 2) the face on the right is detected by tracking the face pattern due to the Viola based detectors failure, 3) the left face is detected using skin color and the right one by means of the local context face detector, 4) the same for 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 recovers its vertical position and it is fused with the latent detection thread.

4 Conclusions

We have described a system which combines multiple cues taking into account their respective computational cost and reliability in the problem of face detection. The approach developed provides fast multiple face detection at different resolutions for standard webcam images, i.e. 320×240, suitable for perceptual user interfaces.

The system is also able to provide information about the relation of the detections in time, reporting good results in the experiments. Currently detection threads can contain among their samples some with bad eye detections, particularly when the face is not completely frontal. In this sense the appearance test must be improved. However, the system is always able to recover once a frontal face is present. Future work must cover the detection of other facial elements in order to have a more robust facial features detection for non frontal poses, as in the current implementation it is only based on eye detection.

Another interesting step to be done is the integration of additional descriptors, e.g. identity, t-shirts color, etc., in order to be able to manage situations with occlusions between individuals, which right now are not specifically analyzed.

Acknowledgments

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

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. Stefan Feyrer and Andreas Zell. Detection, tracking and pursuit of humans with autonomous mobile robot. In *Proc. of International Conference on Intelligent Robots and Systems, Kyongju, Korea*, pages 864–869, 1999.
3. Cayetano Guerra Artal. *Contribuciones al seguimiento visual precategórico*. PhD thesis, Universidad de Las Palmas de Gran Canaria, Octubre 2002.
4. Erik Hjelmas and Ivar Farup. Experimental comparison of face/non-face classifiers. In *Procs. of the Third International Conference on Audio- and Video-Based Person Authentication. Lecture Notes in Computer Science 2091*, 2001.
5. Erik Hjelmas and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3), 2001.
6. Intel. Intel open source computer vision library, b4.0. www.intel.com/research/mrl/research/opencv, August 2004.
7. Y. Kirby and L. Sirovich. Application of the karhunen-love procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1), July 1990.
8. 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)*, October 2003.
9. 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*, 2002.
10. Rainer Lienhart, Luhong Lian, and Alexander Kuranov. An extended set of haar-like features for rapid object detection. Technical report, Intel Research, June 2002.
11. 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*, 2000.
12. Karin Sobottka and Ioannis Pitas. A novel method for automatic face segmentation, face feature extraction and tracking. *Signal Processing: Image Communication*, 12(3), 1998.
13. Moritz Storring, Hans J. Andersen, and Erik Granum. Physics-based modelling of human skin colour under mixed illuminants. *Robotics and Autonomous Systems*, 2001.
14. M. Turk. Computer vision in the interface. *Communications of the ACM*, 47(1):61–67, January 2004.
15. V. Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995.
16. Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, 2001.
17. 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), July 1997.
18. 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.