Facial Recognition in Video

Gorodnichy, Dimitry

Proceedings of International Association for Pattern Recognition (IAPR) International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'03), LNCS 2688, 2003

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Facial Recognition in Video *

Gorodnichy, D.
June 2003

* published in the Proceedings of International Association for Pattern Recognition (IAPR) International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'03). LNCS 2688. pp. 505-514. Guildford, United Kingdom. June 9-11, 2003. NRC 47150.

Canada

Facial Recognition in Video*

Dmitry O. Gorodnichy

Computational Video Group, IIT, NRC-CNRC, Ottawa, Canada K1A 0R6
http://www.cv.iit.nrc.ca/~dmitry/face-in-video

Abstract. There is a physiological reason, backed up by the theory of visual attention in living organisms, why animals look into each others' eyes. This is to illustrate the main two properties in which recognizing of faces in video differs from its static counterpart – recognizing of faces in images. First, the lack of resolution in video is abundantly compensated by the information coming from the time dimension. Video data is inherently of a *dynamic* nature. Second, video processing is a phenomena occurring all the time around us - in *biological systems*, and many results unraveling the intricacies of biological vision already obtained. At the same time, as we examine the way the video-based face recognition is approached by computer scientists, we notice that up till now video information is often used partially and therefore not very efficiently. This work aims at bridging this gap. We develop a multi-channel framework for video-based face processing, which incorporates the dynamic component of video. The utility of the framework is shown on the example of detecting and recognizing faces from blinking. While doing that we derive a canonical representation of a face best suited for the task.

1 Face in video analysis

Video has become widely accepted as one of the most valuable sources of information. In the context of human-oriented applications such as security surveillance, immersive and collaborative environments, multi-media games, computer-human interactions, video-conferencing, video annotation and coding, video information is analyzed for the presence of information about faces. We refer to such an analysis as the *face in video analysis* and the problems tackled by this analysis as the *face processing problems.*

As opposed to off-line image processing, as in still images, video processing is bound to deal with low-quality images. The reason for this are the real-time and bandwidth constraints imposed on on-line video processing. Even with the image capture size of 640 x 480, the image processing has to be done on images of even lower resolution, on such as evenly sampled 160x120 images or mpeg-decoded unevenly sampled 352x240 images [1]. Such a seeming deficiency of video however should not be understood as the disadvantage of video processing – for let us not forget that in biological systems, which are very successful in performing face

* In Proc. of IAPR Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA'03), Guildford, UK, Springer-Verlag, 2003.

in video analysis, the quality of images is also very poor except for a very small area – but instead it should serve to demonstrate that approaches other than those developed for still imaginary should be used for video-based processing. In order to design such approaches, let us first focus on the hierarchy of face processing tasks performed in the face in video analysis and then see how these tasks are performed in biological systems.

Referring to Figure 1, we see that the low-level face segmentation task (FS), which deals with detecting parts in the image which potentially belong to faces, precedes the face detection task (FD), which puts together the pieces of information detected by FS to decide whether a face is seen or not. Should a face be detected, the processing switches to face tracking (FT) and face localization (FL). We deliberately make the distinction between FD and FT in that FT uses the past information about the face location, whereas FD does not; and between FT and FL in that FT detects an approximate face location, while FL provides the exact position of a face or facial feature(s). As such, FL can be used for hands-free control, which is well illustrated by the *Nouse 'Nose as Mouse'* technology [2], whereas FT is more suitable for video-surveillance applications where, for instance, a camera follows a face [3]. After a face is localized, high-level face processing tasks, such as face identification (FI), face memorization (FM), face classification (FC) and facial expression (or event) recognition (FER), can be performed.

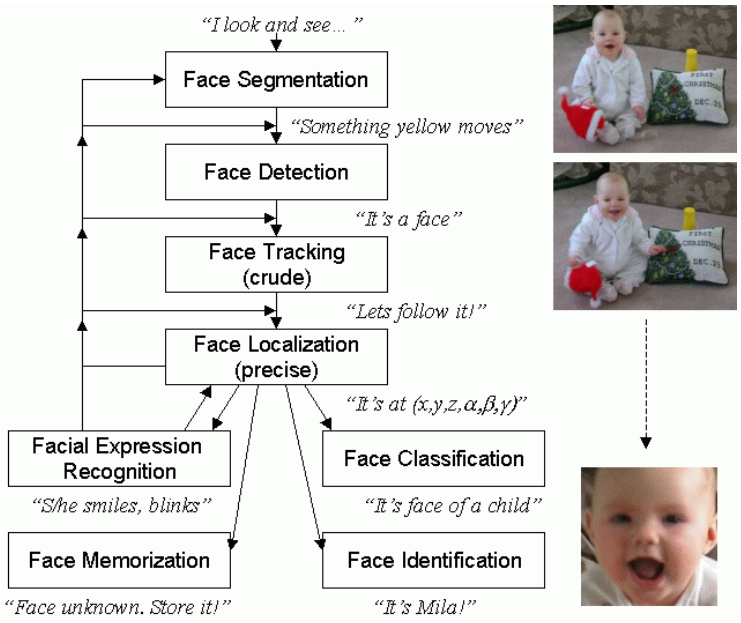

Fig. 1. The hierarchy of face processing tasks performed in face in video analysis.

1.1 The way nature does it

Following the work in neurobiology [4, 5] and empirical observations of human visual performance, we find the following results.
1 *Accumulation over time.* The resolution of video in biological vision systems is low everywhere except in the fovea, which is the neighborhood of the point of visual attention (fixation point). Excellent vision-based recognition is achieved therefore by, first, using an efficient selective visual attention mechanism and, second, by accumulating the information over time.
2. *Image-based attention visual saliency.* When viewing a scene, our eyes do not scan it pixel by pixel, but rather scan it in, what is known as saccadic motion, from one salient point to another.
3. *Visual saliency.* Salient points are those which correspond to maximums of discontinuities in a) motion, b) colour, c) intensity gradients (orientation and spatial frequency) and d) disparity (depth) values of the video data.
4. *Importance of eyes.* Eyes are high contrast moving parts on a face and are therefore very distinctive maximums on saliency maps built in the visual cortex, are attended by animals first. Figure 3.a, which shows commercially available pictures designed to capture the attention of infants, illustrates this phenomenon.
5. *Multi-channel nature of processing.* There are very efficient mechanisms in brain to process each of four mentioned types of video information, which most likely function in parallel and in different parts of the retina and the early visual cortical areas. This is illustrated by the examples of a frog catching a fly or a bull running on a torero.
6. *Intensities for recognition.* Humans make use of colour for the purpose of segmenting a face, rather than for the purpose of recognizing it. Recognition is thus performed on the intensity images. However, it is not the intensity values which are used but the gradient, orientation and frequency values of intensity.
7. *Non-linear colour perception.* Humans perceive colours discretely. Psychophysical thresholds separate colours which are perceived by humans as two different colours, which suggests that such a nonlinear colour representation helps humans to segment and memorize the colour.
8. *Goal- riven attention.* A high-level goal, such as "look for faces", also governs the order of scanning a scene; however goal driven deployment of attention is much slower than visual saliency driven one: 200-1000 ms vs 25-50 ms.
9. *Localization vs recognition.* Two separate parts of brain in the visual cortex are responsible for localization ("where") and identification ("what") tasks: in dorsal and ventral steams respectively. Recognition is therefore done after an object has been localized, which suggests that recognition is done in a canonical coordinate space used in memorizing the object.

As we examine the way facial recognition problems are approached by computer scientists – Table 1 shows the representative work done in the area — we notice that there is a big gap between the way the problems are resolved by biological systems and algorithmically. Very often high-level face processing tasks are performed regardless of the result of the low-level tasks. The video

information is often used partially (e.g. only the static part or only the intensity part of it is used), and therefore not very efficiently.

Table 1. Anthropometrics of face used to decide which facial recognition tasks can be performed in 160 x 120 video, and the types of video information applicable for each of these tasks.

face size	$\frac{1}{2}$ image	$\frac{1}{4}$ image	$\frac{1}{8}$ image	$\frac{1}{16}$ image				
in pixels	80x80	40x40	20x20	10x10				
between eyes	40	20	10	5				
eye size	20	10	5	2				
nose size	10	5	–	–	stereo	colour	motion	gradient
FS	X	X	X	m	+	+	+[6]	+
FD	X	X	m	–	+	+[7, 8]	+	+[9–12]
FT	X	X	m	–	+	+[13]	+	+[3]
FL	X	m	–	–	+[14]			+[15]
FER	X	X	m	–		+	+	+[16, 17]
FC	X	X	m	–				+[18]
FM/FI	X	X	–	–				+[19]

Seemingly simplified, the top left part of the table, in fact, very well describes the spatial relationship between the facial features. In the bottom left part of the table, 'X' indicates that for a given face size the task can be executed; 'm' signifies that the size is marginally acceptable for the task. For example, precise nose-based face tracking (FT) [15] is possible only if the face is close enough to the camera; or more specifically, when a face occupies at least $\frac{1}{4}$ of the image so that convex-shape of the nose is five pixels at least (see also Figure 2). The references are given as a guide only and are far from complete.

2 Multi-channel face processing vision system

Following the results presented in the previous section, we represent video data using the set of channels corresponding to motion, colour, intensity and depth[1] information of video. Each channel is responsible for the specific set of tasks for which the channel is the most suitable; the tasks are executed in the order of their complexity: from most low-level ones to most high-level ones, in accordance with Figure 1 .

The video is processed at the 160 x 120 resolution. This resolution allows one to perform video processing in real-time and, as mentioned above and shown in Table 1, is quite sufficient for many face processing tasks. In the following we describe how these tasks are performed by using the multi-channel video representation, the main challenges outlined. The binaries of our system are available from our website for public evaluation.

[1] The depth channel, which computes the disparity information and which is used in [14], is not used in the present paper.

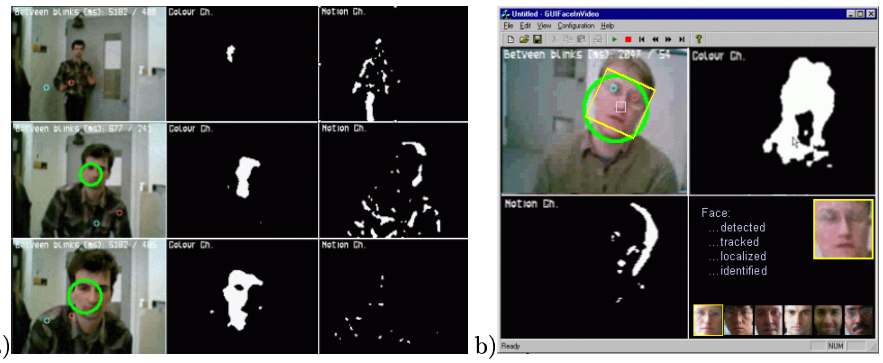

Fig. 2. Face segmentation, detection, tracking and recognition using the multi-channel (intensity, colour and motion) video representation: at three different ranges (a) – the last row shows the range at which blink detection and nose tracking become are triggered; at the close range after a person blinked (b) – after a face is localized, it is recognized using the canonical 24x24 face representation. See also Table 1 and Figure 3.

Colour channel for segmentation and detection. Colour provides the fastest and scale independent way of detecting a colour specific object. For some insects and animals, colour is also the most dominant source of information (see Section 1). In the case of faces, it is the skin colour that discriminates them. Thus we use the colour channel to provide the initial estimates for the face size and location. If the estimated face size is big enough for other channels to become applicable – refer to Table 1, then the face position is further refined by using the motion and intensity channels.

A critical step in this approach is finding a colour space in which a reliable colour skin model can be built. After examining several colour spaces, including HSV [13], YCbCr [7], we have found that the perceptually uniform colour system (UCS) [20] performs the best, which comes as an interesting result, should we recall that the UCS space is obtained from the RGB space by nonlinear mapping based on the empirical study on psychophysical thresholds of human colour perception (Section 1). For this colour space, we build the skin model by using the motion-based eye localization and convex-shape nose tracking described later in the paper. These techniques provide an efficient way of learning the skin model on-fly by computing the colour histogram around the eyes and the nose tip. This is especially helpful for cameras, such as USB webcams, which have automatic colour adjustment.

Typical results of skin segmentation and skin-based face detection (shown using a green circle) are given in Figure 3. It should be noted that there still might be objects present in the field of view such as walls, for instance, which have the same colour as skin. These objects however in most cases are static and can therefore be filtered out by using the motion channel.

Motion channel for motion detection. Motion information plays an extremely important role in video processing, which is seen from the very fact that certain amphibians do not see an object at all unless it moves (Section 1). The detection of motion in captured video is not simple however. Because of the noise present in the image, the amount of which is especially large in low-quality cameras, the simple frame difference commonly used to detect change in an image may not always be used. Another reason not to use frame subtraction for motion detection is that it does not distinguish the changes due to lighting changes (as when switching the lights on and of) from the changes due to the motion of objects. This is why in order to design a motion detector tolerant to noise and illumination changes, we use a non-linear change detection method, according to which a pixel is considered to have changed only if the area around the pixel, called the support area, has changed non-linearly. One way of detecting such a non-linear change is to compare vectors $\boldsymbol{x}_t = \{x_{i,t}\}$ and $\boldsymbol{x}_{t+1} = \{x_{i,t+1}\}$, $i = 1...n$ created from the pixel intensities in the support area of pixel x in frames I_t and I_{t+1}, respectively. If these vectors are collinear, meaning that the intensity change was linear, than there was no motion observed in pixel x. Otherwise, pixel x is considered to have changed due to motion. Other more robust ways of detecting a non-linear intensity change can be found in [6].

The motion channel is useful for detecting the static background, tracking the moving objects away from the camera. It is also used at close range for eye localization. Figures 2 shows the results.

Motion channel for eye localization. Humans have to blink in order to keep their eyes moist. This, as mentioned in Section 1, makes eyes the most salient features in a scene. Besides, the eyes blink simultaneously. This provides even an additional piece of information which makes eyes and, consequently, the face easy to localize at close range. Not much work however seems to be done on making use of such an advantage of eyes. Recent two surveys on face detection [9, 10] mention only one paper [21] which uses blinking for face localization. In that paper however, as well as in others [22], the face is assumed to be stationery, which is a trivial situation not suitable for most applications.

In order to detect eye blinks in moving heads, we have recently proposed the technique based on computing the second order change (i.e. the change of the change) in the image [23]. This technique allows one to discriminate the local (most recent) change in image, such as blink of the eyes, from the global (long lasting) change, such as the motion of head. This, as shown in [23], makes it possible to design robust hands-free blink-operated systems. As we show in the current paper, this also aids significantly face recognition.

Intensity channel for localization and recognition. When a face is close enough to capture the convexity of the nose surface (refer to Table 2), the sub-pixel accuracy nose tracking technique [15] is invoked. This technique uses the intensity gradient to track the convexity of the nose, which is orientation and scale invariant . As a person blinks, his nose and eyes positions are retrieved and the locations of these three points are used to decide whether the orientation of the face is suitable for face recognition. If they form an equilateral triangle,

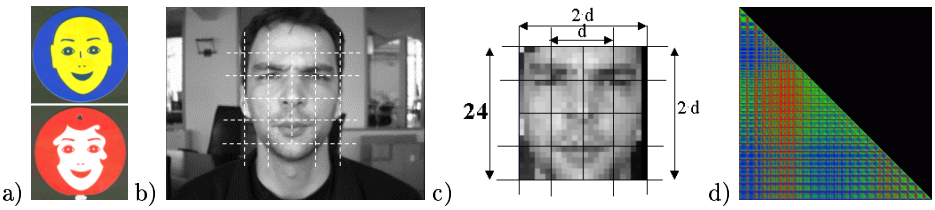

Fig. 3. Eyes in toy pictures (a), the canonical eye-centered 24x24 face representation (b-c) and the statistical relationship between the pixels of faces transformed to this representation as computed on 1500 faces from the BioID Face Database (d) ; each point in this 576x576 array shows, using the RGB colour, how frequently the two pixels of the transformed 24x24 face are darker one another (R), brighter one another (G) or are the same (B), within a certain boundary. The presence of pure RGB colours in the array image indicates the strong relationship between the face pixels. One of the faces from the database is shown.

meaning that the face is in the plane parallel to the image plane, than the high-level face recognition module is launched; otherwise the tracking continues until a better moment occurs.

Localizing a face prior to its recognition makes the recognition easier. This is because it allows one to perform the recognition in a canonical coordinate space, in which faces are aligned and rescaled to the same size and orientation used in face memorization.

3 Canonical face representation

With the location of the eyes known, it is convenient to transform faces to the eye-centered face representation, in which the eye positions are fixed. The main question then is what shape and size of the face should be used so that it is succinct and informative. To answer this question, we have studied the work of others [9–12, 8, 18, 19], especially that done in two major face detection schools, CMU [11] and MIT [12], and also conducted our own experiments using the BioID Face Database [24], which conveniently provides the location of the eyes in a face along with the face image.

As a result, we have constructed a canonical face representation of size 24x24 pixels as shown in Figure 3. The figure also shows the statistical relationship between the pixels in faces transformed to a 24x24 canonical form, as computed on 1500 faces from the database. It is fascinating to see that the size of face needed for face recognition purposes is twice the intra-ocular distance squared, with eyes being located exactly at the block intersections.

Since colour does not help recognition (Section 1), the grey-scale images are used. The size of 24x24 appears to be the most appropriate for two reasons. First, while being small, it is large enough to make images recognizable, which is

also supported by other work [18, 11, 12]. Second, it is a multiple of four, which facilitates positioning of the eyes in the image.

Canonical representation for recognition. After a face has been transformed to a canonical 24 x 24 form, it can be stored or recognized using generic pattern recognition techniques. For more reliable illumination independent recognition, the intensity derivatives, Gabor frequency filters and/or Haar-like wavelets [18] rather than the intensity values should be used as inputs to the recognizer.

In our case, we use *the pseudo-inverse neural network*[25], which we train on binarized pixel differences of faces transformed to the canonical 24x24 form. For the experiments presented in this paper, we have stored[2] only seven faces (seen in Figure 2.b and Figure 1) and use only 10000 neurons (instead of $(24 * 24)^2 = 331776$) manually selected according to the statistical relationship between the pixels shown in Figure 3.d. The result of on-line recognition with this network of a face, which has been detected, tracked, localized and transformed to the canonical form prior to recognition, as described above, is shown in Figure 2.

While we do not assert our recognition technique to be the most suitable for the task – there are better techniques such as, for instance, the cascaded Ada-boost multilayered perceptrons used in [18] – it serves to illustrate the convenience of storing and recognizing faces using the canonical face representation.

4 Conclusion

In this work we attempted to put together the most recent findings made in the areas of computer vision and biological vision in order to built a versatile and intelligent face processing system capable of detecting, tracking, localizing and recognizing faces in video. Face processing tasks are approached in hierarchical order based on the content of video using the multi-channel representation of video. We emphasized the importance of the dynamic component of video and showed how to use it to make facial recognition easier. To do that we localize a face by detecting the eye blinks, after which we transform the face to a canonical form convenient for recognition.

References

1. T. Ebrahimi and C. Horne, "Mpeg-4 natural video coding - an overview," in *In Signal Processing: Image Communication, vol. 15, pp. 365-385, Elsevier*, 2000.
2. NRC-CNRC, "*Nouse* (Nose as Mouse) perceptual interfaces technology," http://www.cv.iit.nrc.ca/Nouse, 2001.
3. M. Turk, C. Hu, R. Feris, F. Lashkari, and A. Beall, "TLA based face tracking," in *Proc. of Intern. Conf. on Vision Interface (VI'2002), online at www.visioninterface.org/vi2002*, Calgary, May 2002, pp. 229–235.
4. L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[2] The cpp code for storing and recognizing patterns with pseudo-inverse neural network is available at http://www.cv.iit.nrc.ca/~dmitry/pinn.

5. D. Walther, M. Riesenhuber, T. Poggio, L. Itti, and C. Koch, "Towards an integrated model of saliency-based attention and object recognition in the primate's visual system," in *Journal of Cognitive Neuroscience Vol. B14*, 2002.
6. E.Durucan and T.Ebrahimi, "Change detection and background extraction by linear algebra," in *IEEE Proc. on Video Communications and Processing for Third Generation Surveillance Systems, 89(10)*, 2001, pp. 1368–1381.
7. R-L Hsu, M Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Trans. on Pattern Anaylis and Machine Intelligence*, vol. 24, no. 5, pp. 696, 2002.
8. R.Feraund, O.J.Bernier andJ. E. Viallet, and M. Collobert, "A fast and accurate face detector based on neural network," *IEEE Transactions on Pattern Anaylis and Machine Intelligence*, vol. 23, no. 1, 2001.
9. E. Hjelmas and B. K. Low, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236 – 274, 2001.
10. M. Yang, N. Ahuja, and D. Kriegman, "Detecting faces in images: A survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.
11. H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Anaylis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.
12. Heisele, T. Poggio, and Pontil, "Face detection in still gray images," *ai.MIT.com tech report*, 2000.
13. G. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, , no. 2, 1998.
14. D.O. Gorodnichy, S. Malik, and G. Roth, "Affordable 3D face tracking using projective vision," in *Proc. Intern. Conf. on Vision Interface (VI'2002)*, Calgary, May 2002, pp. 383–390.
15. D.O. Gorodnichy, "On importance of nose for face tracking," in *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition (FG'2002)*, 2002.
16. Y.-L. Tian, T. Kanade, and J.F. Cohn, "Recognizing action units for facial expression analysis," *Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.
17. M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on PAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.
18. G Shakhnarovich, P. A. Viola, and B. Moghaddam, "A unified learning framework for realtime face detection and classification," in *Intern. Conf. on Automatic Face and Gesture Recognition, USA, pp 10-15*, 2002.
19. T. Fromherz, P. Stucki, and M. Bichsel, "A survey of face recognition," 1997.
20. H Wu, Q Chen, and M Yachida, "Face detection from color images using a fuzzi pattern matching method," *IEEE Transactions on Pattern Anaylis and Machine Intelligence*, vol. 21, no. 6, pp. 557, 1999.
21. J. L. Crowley and F. Berard, "Multi-modal tracking of faces for video communications," in *Proc. CVPR 97*, 1997, pp. 640–645.
22. K.Grauman, M. Betke, J.Gips, and G.Bradski, "Communication via eye blinks detection and duration analysis in real time," in *Proc. CVPR 01*, 2001.
23. D.O. Gorodnichy, "Second order change detection, and its application to blink-controlled perceptual interfaces," in *Proc. IASTED Conf. on Visualization, Imaging and Image Processing (VIIP 2003), Benalmdena, Spain, Sept.8-10*, 2003.
24. The BioID, "Face database," *http://www.bioid.com/downloads/facedb/facedatabase.html*, 2001.
25. D.O. Gorodnichy and A.M. Reznik, "Increasing attraction of pseudo-inverse autoassociative networks," *Neural Processing Letters*, vol. 5, no. 2, pp. 123–127, 1997.