FACE DETECTION IN THE COMPRESSED DOMAIN

Pedro Fonseca[1], Jan Nesvadba[2]

[1] Instituto Superior Técnico - Instituto de Telecomunicações, Lisboa - Portugal, *pmf@lx.it.pt*

[2] Philips Research, Eindhoven - The Netherlands, *jan.nesvadba@philips.com*

ABSTRACT

Face detection is important in many algorithms in the areas of machine object recognition and pattern recognition. The kaleidoscope of applications for face detection extends across automatic image and home video content annotation, face-image stabilisation and face recognition systems. By using information derived from colour, luminosity and frequency the face detection algorithm proposed in this paper aims to determine the location of multiple frontal and non-frontal faces in compressed MPEG-1, MPEG-2 video or JPEG image content. The described algorithm requires only low computational resources on CE devices and offers at one and the same time extremely high detection rates.

1. INTRODUCTION

Face detection may be defined as the identification of faces in image or video content. It may or may not encompass face localization, that is, the identification of the exact location of faces in the content. It may be seen as the initial step to, but should not be mistaken for, face recognition, i.e. the identification of particular persons and their expressions based on their facial parameters. When performed in the compressed domain, face detection uses data available of compressed images and video streams, needing only little decompression (the bare minimum necessary to retrieve necessary data from the compressed streams). Compared to pixel domain based algorithms this approach opens the way to faster and, in terms of processing power and computational complexity, cheaper solutions yielding equally good results.

Although the subject of face detection has been addressed intensively in the multimedia community, little work has been done to address the problem of face detection in the compressed domain. In [1], Wang and Chang proposed face region detection in MPEG video sequences using Discrete Cosine Transformation (DCT) coefficients of MPEG video as input. Faces are then detected based on their colour and shape properties. Frequency information provided by AC coefficients is then used to reduce the number of false detections. In [2], Luo and Eleftheriadis proposed face detection using DCT coefficients based on both colour and texture information. The processing of colour information is similar to that done in [1]. Statistical model training and detection formed the bases for texture analysis. In [3], Zhao et al. proposed a DCT-based system that allows the extraction, tracking and grouping of face sequences in MPEG video.

The face detection algorithm proposed in this paper performs the detection in the compressed domain and can be used to detect multiple faces in DCT compressed I-frames in MPEG-1 or MPEG-2 video streams and on DCT compressed JPEG images. Faces are detected based on colour, luminosity and frequency information. The proposed approach allows for an extremely robust solution regarding the detection of faces in the compressed domain, by combining techniques from existing pixel domain based algorithms and some original techniques. Besides, the algorithm's computational complexity is extremely low being adequate for implementation and usage in stationary and even mobile consumer electronic devices with low computational resources.

The organization of this paper is as follows: the next section describes the proposed face detection algorithm; Section 3 evaluates the algorithm's computational complexity; in Section 4 performance tests on the proposed algorithm are described along with their results; the paper concludes with Section 5.

2. THE FACE DETECTION ALGORITHM

To detect faces the algorithm uses a heuristic approach, i.e., uses knowledge about typical human faces, their facial features and the relationship between them. The block diagram in Figure 1 represents an overview of the face detector's architecture.

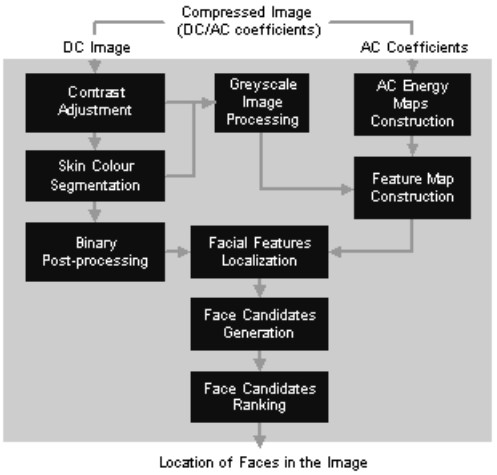


Figure 1 - Face detector algorithm architecture.

The face detector's input corresponds to some AC coefficients and all DC coefficients taken from the compressed image where detection is to be performed. A DC colour image is built, in which each three-component colour pixel corresponds to an entire block in the compressed image, represented by its DC value for each of the three colour components. This DC image's resolution is therefore reduced by 64 compared to the original image's resolution (8 times horizontally and vertically).

2.1 Contrast Adjustment

In order to make the face detection algorithm as robust and independent as possible of the image or video capturing conditions, automatic contrast adjustment is applied to each input DC image. This contrast adjustment process consists simply on a histogram equalization operation applied on the luminosity component of the image aiming to enhance the image's luminosity component's contrast dynamic range by flattening its intensity histogram. Figure 2 (a) and (b) illustrate an example of the input and output of the contrast adjustment operation.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

(a) (b) (c) (d) (e) (f)

Figure 2 - (a) Original image's luminance component, (b) after automatic contrast adjustment; (c) DC colour image after contrast adjustment (similar to that in (b) if print out in black and white) (d) after skin colour segmentation, (e) after binary closing and hole filling and (f) the two main regions identified in the binary image.

2.2 Skin Colour Regions Identification

Assuming that all potential faces are represented by patches of skin colour regions, image segmentation will immediately allow to discard most of the irrelevant information thus providing the first separation between regions of interest and those where no faces are expected to be found. Skin colour segmentation is applied on the DC colour image resulting from the contrast adjustment stage, by determining the Mahalanobis distance [3, 4] between each pixel's value in the DC image and a skin colour model built from the statistical properties of a large set of manually segmented faces. Pixels in the DC image for which the Mahalanobis distance is under the empirically determined threshold of 25 are declared to be skin colour pixels. The skin colour model is represented in the normalized RGB colour space since under this colour space skin colours form a compact cluster that can be approximated to a Gaussian distribution [6]. Besides, being a purely chromatic colour space, brightness will not influence the representation of colours [6]. This colour space was previously used on face detection and face tracking algorithms with good results (see [7] and [8]). Figure 2 (d) illustrates the result of the skin colour segmentation of the image in Figure 2 (c).

Afterwards, a binary closing operation with a 3×3 square structuring element followed by a hole filling operation are applied on the binary image resulting from the segmentation stage, ensuring that in most cases a face is completely covered by a binary mask, without any holes in it. The result of this processing stage is illustrated in Figure 2 (e). An algorithm for labelling of binary connected regions is then applied in order to identify connected regions in the image. The same technique was applied, for example, in [9] before searching for facial features. The result of this stage is illustrated in Figure 2 (f). Each one of the connected regions identified in the skin colour binary image forms the input for the next steps and is independently evaluated. In this way, faces present in each one of these regions can be located and thus, multiple faces can be detected in each image.

2.3 Feature Map Construction

It is known that among all facial features, the eyes/eyebrows and the mouth are the most prominent for facial detection, recognition and pose estimation [4, 8, 9]. Regions on and around the eyes, eyebrows and mouth are usually darker than surrounding areas and have a high vertical variance in their neighbourhood when compared to other parts of the face, e.g., cheeks or forehead. In order to explore these properties a feature map is built for each input image.

Greyscale Image Processing

Facial features' brightness properties are explored by computing greyscale dilated and eroded images from the contrast adjusted luminosity component of the input image. This stage corresponds to the greyscale processing stage indicated in the algorithm's architecture illustrated before. Greyscale dilation enhances the presence of bright regions in the image; the effect of this operator is the dilation of brighter regions that will grow in area over small darker regions. On the contrary, greyscale erosion enhances the presence of darker regions in the image; darker regions will grow or dilate while small bright regions will shrink. Using either of these two operators, large bright and dark regions remain approximately the same. Both operators are applied independently on the luminance component of the image that resulted from the contrast adjustment stage. The operators are applied only on the pixels identified as skin colour pixels in the skin colour segmentation stage. The result of these operators is illustrated in Figure 3 (b) and (c) for the example face image in Figure 3 (a) (note that this face image is not the result of any of the processing stages described, but rather an example that helps illustrate the effect of these operators).

AC Energy Map Construction

On the other hand, facial features' variance properties are explored using frequency information provided by AC coefficients. It is known that specific sets of DCT AC coefficients represent certain directional variations in the images. The sets of AC coefficients used in the face detection algorithm proposed were used in [11] to detect text captions in the compressed video domain. It can be shown that, for a given block in the image, the values of the set of AC coefficients represented in Figure 4 (a) will be high if there is a strong vertical variance in that block while the values in the set represented in Figure 4 (b) will be high if there is a strong horizontal and diagonal variance. An AC energy map consists of a matrix in which each position corresponds to a block in the compressed image and where each position's value corresponds to the squared sum of the specified AC coefficients' values. The input to this processing stage is a specific set of AC coefficients retrieved from the compressed image. The vertical AC energy map, illustrated in Figure 5 (a) is built from the AC coefficients indicated in Figure 4 (a) while the horizontal AC energy map, illustrated in Figure 5 (b) is built from the coefficients indicated in Figure 4 (b). In [3], AC coefficients are also used to build feature maps.

Feature Map Construction

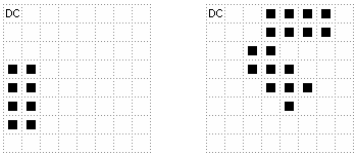
The feature map is built calculating for each position in the feature map matrix the following value

$$\text{featuremap}(x,y) = \frac{\text{dilated}(x,y)}{\text{eroded}(x,y)+1} \cdot \frac{\text{verticalenergymap}(x,y)+1}{\text{horizontalenergymap}(x,y)+1} \quad (1).$$

The first fraction in equation (1) enhances dark locations surrounded by bright areas (a similar equation is used in [5] to determine the location of the eyes in uncompressed images). The second fraction enhances regions with a high vertical variance and with a low horizontal variance thus highlighting the presence of facial features (e.g., eyes, eyebrows, mouths) and de-emphasizing locations like the sides of the face that may have high horizontal variance. The use of AC coefficients information to build feature maps where the location of facial features is enhanced is original. The '+1' factors in the denominator of each fraction prevent that a division by zero occurs. The '+1' factor in the numerator of the second fraction allows that, in case all AC coefficients for the considered sets are zero (due to a high quantization step used to compress the input image), a feature map can still be generated. Figure 6 illustrates the feature map built for the face image illustrated in Figure 3 (a).

(a) (b) (c)

Figure 3 - (a) Face image example, (b) after greyscale dilation and (c) erosion.



(a) (b)

Figure 4 - Set of AC coefficients for (a) vertical and (b) horizontal and diagonal variance.

(a) (b)

Figure 5 - (a) Vertical and (b) horizontal AC energy maps for the face image in Figure 3 (a).

2.4 Face Candidates Generation and Ranking

Facial Features Localization

In order to determine the location of faces that may exist in each skin colour region identified previously (output of Section 2.2), the location of facial features is first determined. Facial features are identified directly from the feature map simply by projecting its values on vertical and horizontal axis (this technique was also used in [9]). This projection is done within the bounding boxes of each skin colour region identified previously, e.g. Figure 2 (f). On the right of the feature map in Figure 6 (b), a plot illustrates the resulting horizontal projection on a vertical axis - as it can be easily seen, the most prominent maxima appear in the rows around the eyes and mouth and smaller local maxima appear in other locations of the face (eyebrows, nose and chin regions). After determining maxima in the horizontal projection, the vertical projection for each of the rows corresponding to these maxima is then determined. In the figure, the vertical projection for the row where the maximum corresponding to the eyes was located is illustrated below the feature map - most prominent maxima appear in the positions corresponding to the left and right eyes.

Face Candidates Generation

Face candidates representing possible locations of faces in the image can now be determined for each skin colour region. These candidates are generated based on the location of the features indicated in Table 1, according to a model of typical frontal and rotated (since vertical relations also apply) human faces illustrated in Figure 7. When generating face candidates based on the position of the mouth, an estimate of its centre must be provided. Multiple face candidates may therefore be generated for each independent skin colour region. It may happen that the face candidates are generated based on maxima locations that do not correspond, in the feature map, to the features being considered (eyes, eyebrows and mouth). In this case, it is likely that they do not represent well a face that may exist in the image. For that reason, face candidates are ranked by computing a relevance value to determine which best represents a face. The relevance value of face candidates of type A or B (Table 1), is determined as

$$r = p_a + p_s + p_e \quad (2)$$

The relevance value of each face candidate of type C or D, is determined as

$$r = p_a + p_s + p_m \quad (3)$$

In the two previous equations, p_a represents the face candidate's area, p_s represents the face candidate's skin colour percentage, p_e represents the face candidate's eyes/eyebrows intensity in the feature map and p_m represents the face candidate's mouth intensity; p_a, p_e and p_m are normalized in respect to the highest values found for each of these parameters in all face candidates. This face candidate generation and ranking procedure is original.

After computing the relevance of all generated face candidates for each individual skin colour region, the best face candidate is determined by choosing the face candidate with the highest relevance for each individual skin colour region - multiple faces can thus be detected.

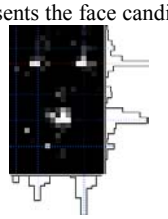


Figure 6 - Feature map for the image in Figure 3 (a) and horizontal and vertical projections.

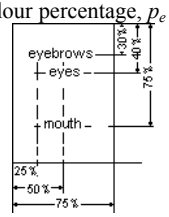


Figure 7 - Face model.

3. COMPUTATIONAL COMPLEXITY

In order to determine the algorithm's computational complexity, i.e., the required number of instructions per second, the face detection algorithm is considered to be applied on an ARM-9 based platform, performing real-time detection on an MPEG-2 video sequence, with a frame rate of 25 frames per second and a GOP size of 6; in this case, since face detection can only act on the I-frames, the actual frame rate where the algorithm will be applied corresponds to the I-frame rate (i.e., 4.17 frame per second). The algorithm's computational complexity, measured in millions of instructions per second (MIPS) and in millions of cycles per second (Mcycles/sec) is indicated in Table 2 - the algorithm uses less than 10% of an ARM-9 processor's capacity.

4. TESTS AND RESULTS

The face detection algorithm was tested on two home video sequences, one indoors and another outdoors. Both sequences were encoded in MPEG-2 with a GOP size of 6. The sequences were captured under difficult and heterogeneous light conditions. The subjects were recorded with frontal and non-frontal poses and with different facial expressions, a variety of hairstyles and some of the subjects had beard and glasses. A face is considered to be detectable if it is tilted below an angle of ±20 degrees, rotated below an angle of ±80 degrees and if most of the face's colour corresponds to skin colour. Detectable faces are further classified according to the following criteria:

- A **frontal face** (FF) is considered as such if it is in the upright position, tilted up to a maximum of ±10 degrees, with a maximum pose rotation of ±15 degrees, and all its facial features are completely visible in the image.
- **Non-frontal faces** (NFF) are all non frontal faces considered to be detectable;
- **Occluded faces** (OF) are any faces considered to be detectable, with more than 50% of occlusion or any faces where one or more facial features are not visible but would be if the face was not occluded.

Type	Facial features' locations
A	Left and right eyes' horizontal and vertical
B	Left and right eyebrows' horizontal and vertical
C	Eyes' vertical and mouth centroid's horizontal and vertical
D	Eyebrows' vertical and mouth centroid's horizontal and vertical

Table 1 - Types of face candidates and facial features from which they were generated.

In order to evaluate the algorithm's performance, face detection was performed on the 736 I-frames of the indoors video sequence and on the 541 I-frames of the outdoors video sequence. All detection results were manually annotated and classified among one of the following classes:

- A **correct detection** (CD) is considered if the bounding box encloses completely all visible facial features in a face: one eye if only one eye is visible or two eyes if both are visible and always the mouth (if visible); Figure 8 illustrates examples of correct detections;
- A **wrong detection** (WD) is considered when the bounding box does not satisfy the previous condition but covers approximately half of the face;
- A **false detection** (FD) is considered when the bounding box does not satisfy any of the two previous conditions;
- A **missed face** (MF) is considered when it appears in the image, is detectable and is not under the bounding box of a correct or wrong detection;

All the angles referred in the previous classifications, were empirically determined. To evaluate the algorithm's performance, the metric *recall* is defined, expressing the ratio of correct detections against the number of correct, wrong and missed detections,

$$recall = \frac{correct}{correct + wrong + missed} \quad (4)$$

Table 3 and Table 4 indicate the detection results for each type of face (frontal, non-frontal and occluded) and the number of false detections for the indoors and outdoors home video sequence, respectively. The test sequences used were considered to be representative of the application scenarios for which the proposed algorithms were developed. In order to assess the algorithm's performance under other application scenarios, further testing would naturally be necessary on representative video sequences (e.g., news videos).

5. CONCLUSIONS

A wide variety of techniques have been proposed among the vast number of publications in the subject of face detection. However, few techniques address the problem of real-time detection of faces in the compressed domain. The face detection algorithm proposed in this paper attempts to detect multiple frontal and non-frontal faces in images directly in the compressed domain. The algorithm's computational complexity was analysed and found to be extremely low being adequate to be implemented in devices with low computational resources - it uses less than 10% of an ARM-9 processor's capacity. Using a strict detection criterion when classifying correctly detected faces, the face detector achieved high detection rates detecting frontal faces in both indoors and outdoors video sequences (80% and 95%).

It was also able to detect non-frontal faces with a pose rotation of up to 90 degrees with a detection rate of 61% and 71%.

In conclusion, the novelty of the face detection algorithm here proposed lays in the modular combination of existing pixel domain techniques, adapted to the purpose of face detection in compressed domain, with some original techniques. Original techniques such as face candidates generation and ranking - allowing for a flexible and computationally cheap way to detect multiple faces - add up to an algorithm that is simultaneously modular, robust and computationally cheap, when compared to existing solutions.

REFERENCES

[1] H. Wang, S-F. Chang, *A Highly Efficient System for Automatic Face Region Detection in MPEG Video*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 7, no. 4, pp. 615-628, August 1997.

[2] H. Luo, A. Eleftheriadis, *On Face Detection in the Compressed Domain*, Proceedings of the Eighth ACM International Conference on Multimedia, pp. 382-384, Los Angeles - United States, October 2000.

[3] T.-S. Chua, Y. Zhao, M. S. Kankanhalli, *Detection of Human Faces in a Compressed Domain for Video Stratification*, The Visual Computer, pp. 121-133, vol. 18, no. 2, 2002.

[4] V. Vezhnevets, V. Sazonov, A. Andreeva, *A Survey on Pixel-based Skin Color Detection Techniques*, Graphicon-2003, Moscow - Russia, September 2003.

[5] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, *Face Detection in Color Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696-706, May 2002.

[6] J. Yang, W. Lu and A. Waibel, *Skin-color Modeling and Adaptation*, Proceedings of ACCV'98, vol. 2, pp. 687-694, January 1998.

[7] M-H. Yang, D. J. Kriegman, N. Ahuja, *Detecting Faces in Images: A Survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, January 2002.

[8] E. Hjelmås, B. K. Low, *Face Detection: A Survey*, Computer Vision and Image Understanding, vol. 83, no. 3, pp. 236-274, September 2001.

[9] K. Sobottka, I. Pitas, *A Novel Method for Automatic Face Segmentation, Facial Feature Extraction and Tracking*, Signal Processing: Image Communication, vol. 12, no. 3, pp. 263-281, June 1998.

[10] A. Nikolaidis, I. Pitas, *Facial Feature Extraction and Determination of Pose*, Proceedings of the 1998 NOBLESSE Workshop on Nonlinear Model Based Image Analysis, Glasgow - Scotland, July 1998.

[11] Y. Zhang, T. Chua, *Detection of Text Captions in Compressed Domain Video*, Proceedings of the 2000 ACM Workshops on Multimedia 2000, pp. 201-204, Los Angeles - United States, October 2000.

Resolution	MIPS	Mcycles/sec
PAL D1	**7.53**	**20.95**
VGA	**5.58**	**15.52**
SIF	**1.53**	**4.27**

Table 2 - Algorithm's complexity for different frame resolutions.

	CD	WD	MF	Total	Recall	FD
FF	147	35	1	183	**0.80**	
NFF	70	28	17	115	**0.61**	-
OF	12	19	6	37	**0.32**	
Total	229	82	24	335	**0.68**	**68**

Table 3 - Detection results for the indoors video sequence.

	CD	WD	MF	Total	Recall	FD
FF	113	4	2	119	**0.95**	
NFF	17	6	1	24	**0.71**	-
OF	4	5	7	16	**0.25**	
Total	134	15	10	159	**0.84**	**34**

Table 4 - Detection results for the outdoors video sequence.

Figure 8 - Examples of correctly detected faces.