Human face detection for automatic classification and annotation of personal photo collections

Bertrand Chupeau, Vincent Tollu, Jürgen Stauder and Izabela Grasland
Thomson
Research & Innovation, Corporate Research Rennes
1, Avenue de Belle-Fontaine, CS17616, 35576 Cesson-Sévigné Cedex, France

ABSTRACT

The recent proliferation of digital images captured by digital cameras and, as a result, the users' needs for automatic annotation tools to index huge multimedia databases arouse a renewed interest in face detection and recognition technologies. After a brief state-of-the-art, the paper details a model-based face detection algorithm for color images, based on skin color and face shape properties. We compare a stand-alone model-based approach with a hybrid approach in which this algorithm is used as a pre-processor to provide candidate faces to a supervised SVM classifier. Experimental results are presented and discussed on two databases of 250 and 689 pictures respectively. Application to a system to automatically annotate the photos of a personal collection is eventually discussed from the human factors point of view.

Keywords: face detection, skin color, segmentation, face model, supervised classification

1. INTRODUCTION

While a human being easily and immediately locates faces in a picture, the process is fairly more complex for a machine. This face detection problem has been tackled by the computer vision community for decades, as being of prime importance in many respects. For example, face detection lays the foundation of person tracking in a video. Allocating more bits to the face region in the encoding process improves quality in a videophone transmission. New interfaces, beyond the conventional keyboard and mouse, are enabled with a machine analyzing user's emotions from the captured face pictures. Furthermore, detection is a prerequisite to face recognition. The last decade having witnessed tremendous advances in face recognition technology, robust and efficient face detection pre-processing packages are more and more expected. In a multimedia indexing context, automatic annotation of films and videos on fine criteria such as characters appearance on a shot basis are enabled by automatic face recognition.

The aim of this paper is to describe a new algorithm for automatic face detection in color images, in the context of personal photo collections. Due to the proliferation of digital images, captured by digital cameras, automatic annotation tools to index huge multimedia databases are urgently required. Face detection and recognition is indeed a mandatory feature for any system aiming at automating the annotation of images, for example personal photos. Locating faces in such pictures of varying quality and unpredictable content remains a difficult task however.

In the paper, after a brief state-of-the-art section, we detail in section 3 a model-based face detection algorithm for color images. An alternative hybrid approach in which this algorithm is used as a pre-processor to provide candidate faces to a supervised SVM classifier is proposed. Experimental results are then presented and discussed in section 4 on two databases of 250 and 689 pictures respectively. Application to face recognition technology into future digital personal photo management systems is eventually discussed in section 5 from the human factors point of view.

2. STATE OF THE ART

Automatic face detection in images has been the subject of active research for decades. Recently two complementary reviews painted a panorama of that field [1][2]. Two broad families of techniques are identified: traditional "feature-based" approaches, using explicit knowledge through colorimetric or geometrical face models, and more recent "image-based" pattern recognition approaches, obtaining implicit knowledge by learning from examples.

The feature-based approaches make use of low-level features such as edges, gray-levels, color or motion. Local gray-level minima can indicate eyebrows, pupils and lips. In a normalized chrominance space, skin color (whatever the ethnic group) was shown to occupy a tight cluster that can be modeled with a single Gaussian distribution. Higher-level facial features can be sequentially searched for, beginning with eyes, or grouped into constellations and matched against a face template. Active shape models such as snakes or deformable templates are used to extract such non-rigid features as eye-pupil or lips, but are very sensitive to spatial initialization. Most of the above-listed model-based methods limit themselves to head-and-shoulders and quasi-frontal scenes.

In order to solve more difficult problems such as detecting multiple faces in cluttered background, pattern recognition algorithms, learning from examples without an explicit formulation of face knowledge, have been devised. Most of them require, as a preliminary step, an expensive multiresolution window scanning process (i.e. at different scales and positions). Sung and Poggio [3] model both the "face" and "non-face" classes with six clusters each, obtained by learning with a modified k-means, compute for each input image a 12-dimension distance vector to the class patterns that feed a MLP neural network. Their algorithm is reported, together with the advanced neural network systems by Rowley [4], to set the standards for research. The recent detector by Schneiderman [5] using wavelet coefficient as input to a Bayesian classifier also shows impressive results.

The feature-based approach is appropriate for real-time processing when color and motion are available. For static, gray-level pictures, the learning-based approach is the most effective one. Offline detection of unique faces in image with fair resolution is a nearly solved problem. Accurate location of eyes or mouth features remains however difficult. And devising algorithms robust to appearance changes in time and pose variations is still a research topic.

The idea of coupling a model-based approach as a pre-processor to a learning-based one is mentioned in the literature [1]. This would avoid the computationally demanding multiresolution window scanning by selecting a small subset of face candidates, through color and shape features. To our knowledge, no such implementation has been disclosed yet.

3. ALGORITHM DESCRIPTION

The algorithm consists of four steps: detection of skin-color pixels, region segmentation and selection of skin-color regions, shape-based merge of selected regions, discarding of false positives. They are detailed in the following subsections.

3.1 Skin color pixels detection

The starting hypothesis is that, whatever the ethnic group, skin color is localized in a precise subset of the chrominance space [1-2]. We can verify this assumption on the hereunder figure, in which a small portion only of the chrominance plane is depicted (the full scale ranges from –128 to 127) that contains the totality of skin pixels in the three faces. It is thus demonstrated that those three persons are not strictly speaking of different colors but more or less dark only. The *YCbCr* color space was chosen, first of all for being used in the popular JPEG image format, but also because it naturally separates the *Y* luminance component from the other two chrominance components.

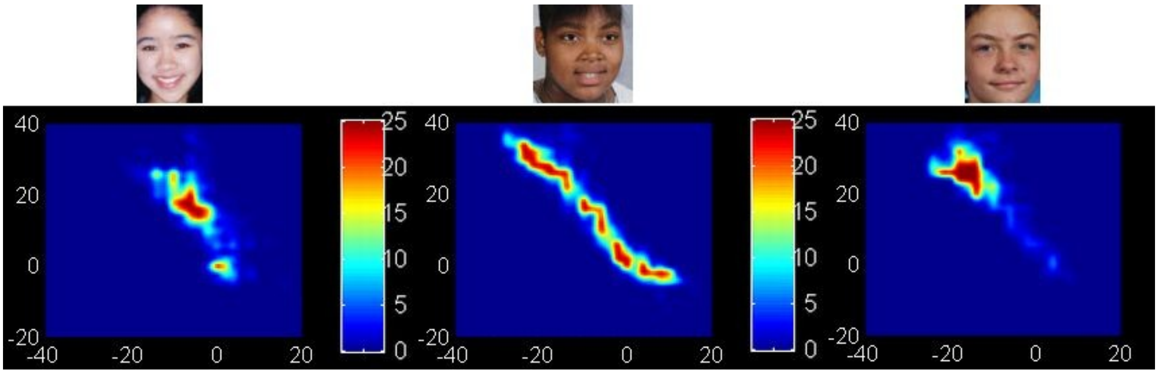

Figure 1. Pixel distribution in (Cb,Cr) plane

This separation is true in first approximation only, however. If at medium luminance levels most skin color *(Y,Cb,Cr)* pixels project onto the chrominance plane in a nearly fixed elliptical region, whatever the luminance, this assumption does not hold any more for very dark or bright luminance values. The ellipse diameter strongly reduces and its center is far away displaced. Hsu and Abdel-Mottaleb [6] proposed a non-linear transform of *YCbCr* to improve the chrominance separation, the effectiveness of which was clearly demonstrated in our experiments.

From those observations, a skin-color probability model can be built in the form of a bi-dimensional Gaussian function, the parameters of which are determined on a learning database. A threshold has then to be set on the probability values in order to reach a binary skin/non-skin decision for each pixel. Such a model is proposed by Gomila for example [7]. We decided to build our own learning bases of skin and non-skin pixels, with the aim to optimally set the Gaussian mean and variance, together with the threshold. Half million skin pixels and five million and a half non-skin pixels were thus manually segmented, such as illustrated in Figure 2.

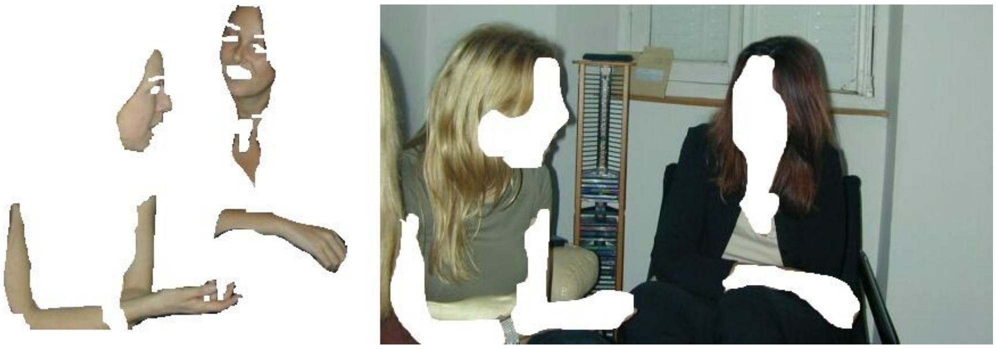

Figure 2. Manual segmentation of skin pixels

The learned mean vector and covariance matrix of the bi-dimensional Gaussian model (illustrated in Figure 3) are the following:

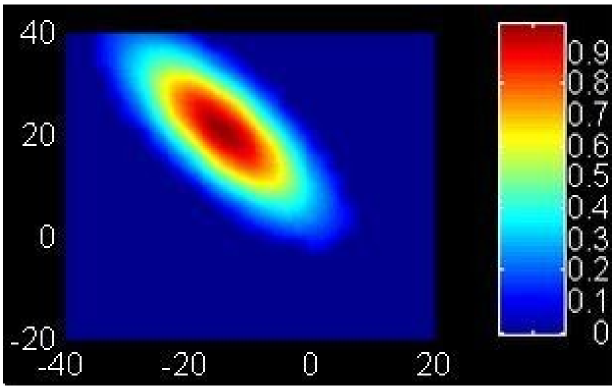

Figure 3. Skin-color probability model

$$\begin{pmatrix} \mu_{Cb} \\ \mu_{Cr} \end{pmatrix} = \begin{pmatrix} -17.32 \\ 24.06 \end{pmatrix}$$

$$\begin{pmatrix} \sigma_{CbCb} & \sigma_{CbCr} \\ \sigma_{CrCb} & \sigma_{CrCr} \end{pmatrix} = \begin{pmatrix} 93.22 & -77.91 \\ -77.91 & 118.62 \end{pmatrix}$$

The obtained skin-color decision map is shown on figure 5 on an example picture. The raw binary map was cleaned up by a morphological simplification processing, consisting in successive opening and closing by reconstruction [7].

3.2 Skin region selection

The next step to climb in the semantic ladder is from pixels to regions, i.e. connected groups of pixels sharing common color or texture properties. Our approach consists in segmenting color-homogeneous regions independently from the skin-color pixel detection, and further counting the number of skin-classified pixels per region. A region-merging algorithm referred to as “RSST” (for Recursive Shortest Spanning Tree [8]) is chosen. Despite its relative computational complexity, it is considered as one of the most powerful tools for image segmentation, compared to other techniques (including color clustering, pyramidal region growing and morphological watershed). Starting from one region per pixel, the two regions that minimize the fusion cost, among all pairs of connected regions (represented by links in the region adjacency graph), are merged iteratively, until mean color difference exceeds a predefined threshold. The fusion cost simply consists of the quadratic mean color differences, summed up on the three components, between the two neighboring regions under consideration, with a weighting coefficient to favor the merge of small regions first and their isotropic growing. Letting $\left(\overline{Y}_1,\overline{Cb}_1,\overline{Cr}_1\right)$ and $\left(\overline{Y}_2,\overline{Cb}_2,\overline{Cr}_2\right)$ be the average color vectors of two regions of size N_1 and N_2, the fusion cost is defined as follows in Equation (1):

$$C_{color} = \frac{N_1 * N_2}{N_1 + N_2}\left[\left(\overline{Y}_1 - \overline{Y}_2\right)^2 + \left(\overline{Cb}_1 - \overline{Cb}_2\right)^2 + \left(\overline{Cr}_1 - \overline{Cr}_2\right)^2\right] \qquad (1)$$

Prior to applying the above described RSST algorithm, the picture is simplified in order to avoid an over-segmentation due to noise. For that purpose we make use of image filters belonging to the family of morphological connected filters [7]. They remove small structures by erosion and dilation while recovering contour accuracy through reconstruction processes. More precisely an “opening by reconstruction” followed by “closing by reconstruction” is implemented.

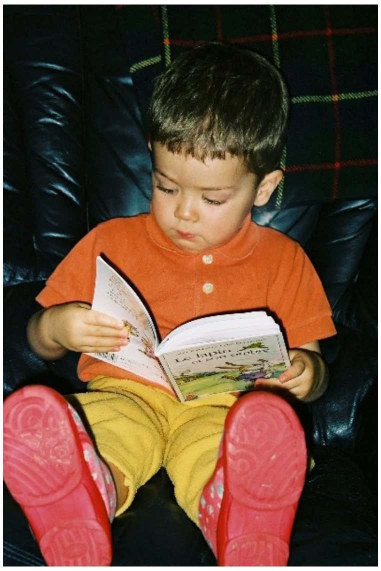

Figure 4. Original picture

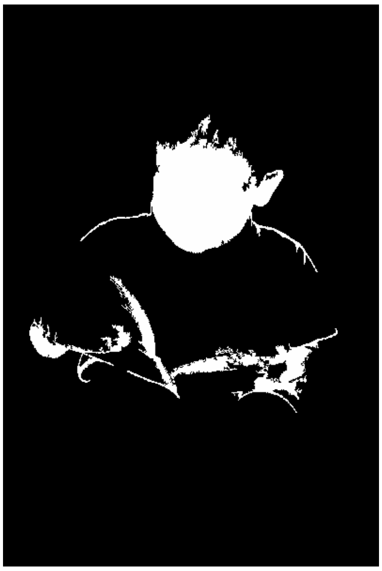

Figure 5. Skin-color decision map

On Figure 6, the color segmentation result is first shown, followed in Figure 7 by a partition map in which the regions with a majority (more than 50%) of skin-labeled pixels have been retained only. It clearly demonstrates that color properties are not enough to discriminate faces.

3.3 Shape-based segmentation

A face appearing to be composed of several regions according to a mere average color criterion, the region merging process has to be continued among the regions selected as skin-color ones, but now taking into account shape properties. The aim is to favor the merge of neighboring skin regions making up elliptical macro-regions. For that purpose the previous RSST algorithm is still used but with a modified fusion cost that combines color and shape properties. The global cost in Equation (2) consists of a weighted sum of a first term ensuring color homogeneity and a second term favoring elliptical shape. The color term is merely the (normalized in [0,1]) quadratic difference between average colors of the two neighboring regions r_1 and r_2 under consideration. The shape term is detailed in Equation (3) where $E\, r_1 \cup r_2$, $E\, r_1$ and $E\, r_2$ refer to a measure of the elliptical nature of r_1, r_2 and their union respectively.

$$C_{face} = C_{skin} + 2 * C_{shape} \tag{2}$$

$$C_{skin} = \frac{1}{3 * 255^2}\left[\left(\overline{Y}_1 - \overline{Y}_2\right)^2 + \left(\overline{C}b_1 - \overline{C}b_2\right)^2 + \left(\overline{C}r_1 - \overline{C}r_2\right)^2\right] \tag{3}$$

$$C_{shape} = \frac{1}{2}\left(1 + Max\left(Er_1, Er_2\right) - Er_1 \cup r_2\right) \tag{4}$$

In other words, the shape-based fusion cost is as smaller as the merged region is more elliptical than the most elliptical one of the two regions under consideration. In order to characterize the elliptical nature of a region, the best-fit ellipse – with same gravity center, same area and maximum overlap – is first computed [9]. Then the mutual overlap is assessed by counting the region pixels outside best-fit ellipse and vice versa the pixels within ellipse but not belonging to the region, and normalizing with respect to region area to get a figure in [0,1]. Results of shape-based fusion are given in Figure 8: the face was merged into a single region.

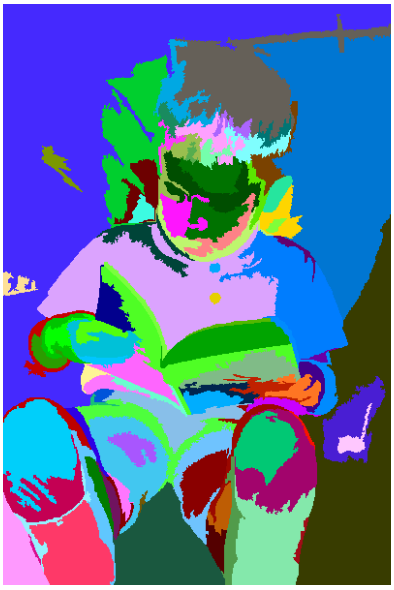

Figure 6. Color segmentation

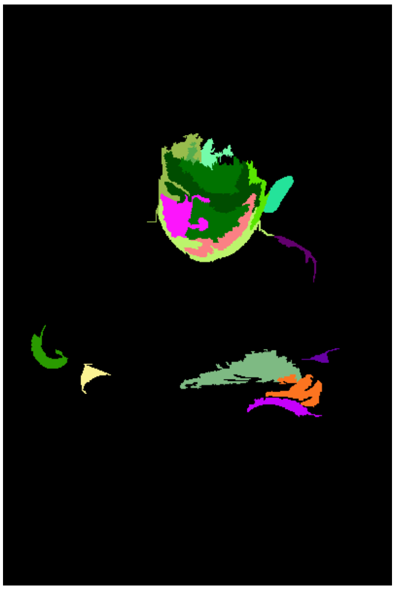

Figure 7. Skin-region selection

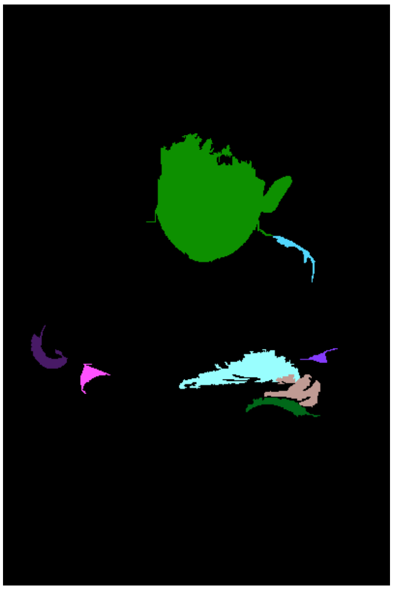

Figure 8. Shape-based fusion

By analyzing the detection results, we realize that defining a face as an elliptical skin-color region is not precise enough. This definition can apply without any problem to a hand, part of a leg and many other objects of very different nature, which explains the high number of false positives at this step.

3.4 Rule-based false positives discarding

The first chosen approach to discard false positives is to successively apply explicit rules derived from common knowledge of what a face looks like. Too large or too small candidate ellipses are discarded, as well as too eccentric ones (too high ratio between major and minor axis lengths). We also check whether the mean color of the grouped regions stays in the skin-color area. Filtering on face candidate orientation is further performed: ellipses the major axis orientation of which lies in $\pi/2 \pm \pi/4$ are selected only. Doing so we are aware we also discard some true positives such as the horizontal face of a person lying on bed for example.

A last and powerful discarding rule taken out from [10] is based on the analysis of the luminance variance. Based on the observation that a face is a complex 3D object reflecting light in a complex way, we expect to find a significant luminance variance on its surface, while the background will have a more uniform luminance distribution. A variant of this method is to focus the variance analysis on the nose axis separating the face in two equal parts: we expect to find along it a more significant luminance distribution as compared with the background. Profile faces that we detected until this last rule, are now discarded however.

In Figure 9, seven candidate ellipses have been identified, only one of which is eventually retained in Figure 10 after applying the above mentioned rules.

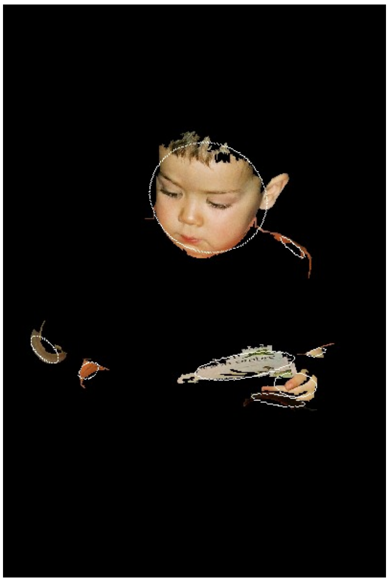

Figure 9. Candidate ellipses

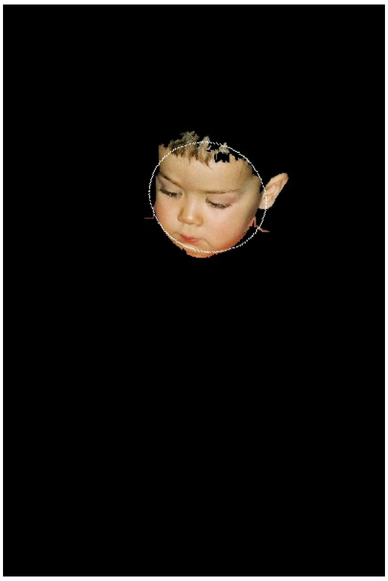

Figure 10. Result of rule-based filtering

3.5 Learning-based false positives discarding

An alternative to the previous rule-based approach is to feed a supervised classifier with the detected candidates. As mentioned in [2], such a hybrid approach alleviates the burden of multiresolution window scanning when systematically applying a computationally demanding classifier at any possible location. A support vector machine [11] was chosen, that determines the equation of the best separating surface in the space of descriptors between face and non-face objects.

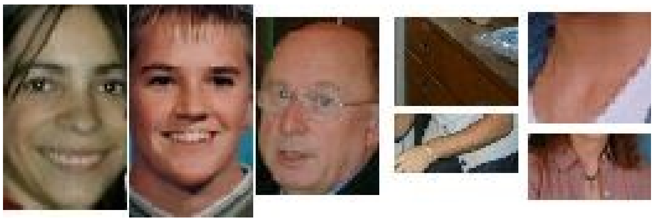

Figure 11. Some training (counter) examples

The classifier was trained with face and non-faces bounding boxes, some of which are illustrated in Figure 11 above. Manual labeling of true and false candidates at the end of the detection step yielded 2500 examples that represent our ground truth. Available descriptors developed for another classification problem (indoor Vs outdoor [12]) were re-used and adapted. They consist of the collection of energies in the high-frequency sub-bands of a hierarchical wavelet transform of the luminance component and are shown to implicitly capture the texture characteristics of faces. Profile faces were excluded from the learning bases.

4. EXPERIMENTAL RESULTS

The first model to adjust was the skin-pixel detector (see 3.1). In order to check and adapt the model proposed by Gomila [7], we created our own bases of skin and non-skin pixels. Hand segmentation of a number of pictures provided us with to 0.5 million skin pixels and 5.5 million non-skin ones. The optimal threshold on the Gaussian skin probability model was obtained by means of ROC (Receiving Operator Characteristic) curves analysis [13]. On those curves depicting false positive rate (1 – specificity) in horizontal and sensitivity (true positive rate) in vertical, the optimal setting corresponds to the point closest to (0,1). As illustrated in Figure 12, we chose to slightly favor over-detection (as a filtering step of false candidates follows). This lead to 82% of true positives and 17% of false negatives on our testing bases. Those figures should be compared to the results by Jones [14] announcing 80% of true positives for 8.5% of false negatives, with a much larger training population of 1 billion skin pixels (and a fairly more complex statistical model).

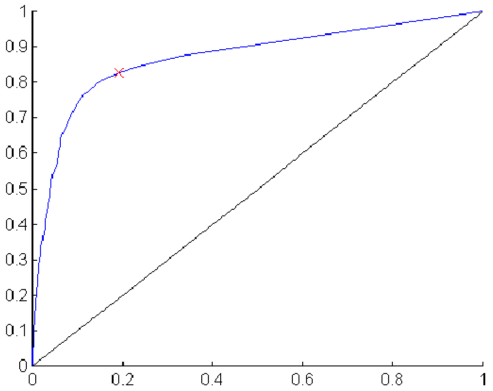

Figure 12. ROC curve to set the skin detection threshold

The whole detection algorithm chain was tested against two databases. The first one consists of 250 pictures collected from the web, containing 336 faces in frontal view in total (a number of pictures do not contain any face at all). They are mostly personal pictures with cluttered background, uncontrolled illumination and varying quality (e.g. under- or over-exposed), which makes face detection even more difficult. We also conducted experiments on an easier base made of 689 mugshot pictures from the Champion Database [15]. Those are on the contrary professional head-and-shoulder photos, with one and only one face per picture. They are relatively easy to analyze but contain a quite large spectrum of skin "colors", which is a good test for our skin-color model.

Before the last false candidates discarding step, after skin-color detection and elliptical shape segmentation, the attained detection rates are 71% on the personal photos from the web and 87% on the portrait database, but with much too high false positive rates of 8 to 1 and 1 to 1, respectively. Either the rule-based or the SVM-based approaches dramatically reduce this false detection rate, the first one providing best results on the head-and-shoulder portrait pictures, while the second one performed better on personal photos from the web. The ratio of false positive remains too high, however: 4 to 1 with the rule-based approach and 0.5 to 1 with the SVM-based one, on the web photos.

5. APPLICATIONS

A photo browsing system that uses image classification results in an error tolerant manner was presented in [12]. Images are hierarchically classified into indoor/outdoor and further into city/landscape. It was observed, however, that indoor/outdoor classification, based on global color histogram, wavelet sub-band energies and contour directions, did not perform satisfactorily on close-up, portrait pictures. A portrait classifier was consequently devised, based on size and location of detected face ellipses, and inserted at the first level of the classification hierarchy (before indoor/outdoor classification).

Besides automatic classification into such categories as "portrait", "people" or "group of people", face detection coupled with automatic recognition [16] enables a new exciting functionality in a digital personal photo collection management system: automatic annotation and search and browse by people's name in the relatives and friends face gallery. Several studies confirm indeed the users' expectation for a face recognition functionality in future photo asset management systems. In a study involving eleven families, Frohlich [17] reports users complaining about forgetting people's detail, and thus identifies the need to associate names with persons, in order to keep the memory. Another observed activity is the active selection among personal photo archives for particular social purposes or events, with a clear need of recognizing a given person. The emerging (digital) technology is used as a vehicle for duplicating and distributing family's precious memories to relatives and friends, sending photos being a remarkably common and consistent practice. Rodden [18] observed people using annotation to record names but long after photos were taken, when many of the details have already been forgotten. Once again, recognizing and indexing people help users preserving their memories. Three basic query types are identified: event, specific remembered photos, and set of photos corresponding to different events but sharing a common property such as the presence of a person.

Interesting findings from our own internal surveys highlight that digital pictures are grouped into categories based on time, events, and the people involved. Those systems overlap, allowing people to create non-linear stories with each viewing. We also found that search is the dominant activity in digital collections, because it is determined by precise user goals (send a photo by e-mail for instance). The most wanted automatic annotations are names and places. Regarding the relevance of querying modes, query by person's name is ranked 0.4, less than topic (0.6) but more than place or date (0.3). Those sample answers clearly motivate our current work to integrate face recognition in a content-based indexing and retrieval system for personal photo collections.

6. CONCLUSION AND FUTURE WORK

The algorithm presented in this paper provides promising results but needs more tuning. The achieved detection rate of 71% at the end of the first step on a difficult database compares with state-of-the-art results. It is difficult to benchmark with other approaches, however, as no commonly accepted color picture test database exist, to our knowledge.

In order to improve the robustness of the skin detector to illumination variations, the illumination correction or the skin color model could be improved. For example, Wong [17] recently presented an alternative to the non-linear transform of *YCbCr* by Hsu [6], which should be studied. Along the knowledge-based discarding path, specific detectors for such face features as eyes and mouth could be developed, based on the analysis of local minima in the skin probability map. Distance measurement to an average face model could also be implemented. Along the learning-based discarding path, specific descriptors should obviously be devised, and the learning base enriched. Also, rather than computing the descriptors in the rectangular face bounding box, which contains a number of outlier pixels (clothes, hair, etc.), the inner ellipse pixels only should be used instead.

The next step in our research will of course be the coupling of the detector with a recognizing module. Personal photos are characterized by their unpredictable content, cluttered background, varying quality and illumination. When applying well-known face recognition algorithm to that kind of material, rather than mugshot pictures, we expect to face new challenges.

REFERENCES

1. M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey", *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**(1), 2002, pp. 34-58.
2. E. Hjelmås, and B. K. Low, "Face detection, a survey", *Computer Vision and Image Understanding*, **83**(3), 2001, pp. 236-274.
3. K. K. Sung, and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, **20**(1), 1998, pp. 39-51.
4. H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, **20**(1), 1998, pp. 23-38.
5. H. Schneiderman, and T. Kanade, "A statistical model for 3D object recognition applied to faces and cars", *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.
6. R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images", *IEEE Trans. Pattern Analysis and Machine Intelligence,* **24**(5), 2002, pp. 696-706.
7. C. Gomila, and F. Meyer, "Automatic video object generation tool : segmentation and tracking of persons in real time", *Annales des Télécommunications*, **55**(3-4), 2000, p. 172-183.
8. E. Tuncel, and L. Onural, "Utilization of the recursive shortest spanning tree algorithm for video-object segmentation by 2-D affine modeling", *IEEE Trans. Circuits and Systems for Video Technology*, **10**, 2000.

9. A. K. Jain, *Fundamentals of digital image processing*, Prentice-Hall, Englewood Cliffs, 1989.
10. D. Petrescu, and M. Gelgon, “Face detection from complex scenes in color images”, *Proc. EURASIP European Signal Processing Conference*, Tampere, Finland, 2000, pp. 933-936.
11. V. N. Vapnik, *The nature of statistical learning theory*, Springs, New York, 1995.
12. J. Stauder, G. Gouzien, B. Chupeau, L. Nunez, J.-R. Vigouroux, and E. Kijak, “Semantic image browsing using hidden categories and confidence measures”, *Proc. SPIE Conf. on Storage and Retrieval for Media Databases*, Santa Clara, California, USA, 2003.
13. J. A. Swets, and R. M. Pickett, “Evaluation of diagnostic systems: methods from the signal detection theory”, Academic press, New York, 1982.
14. M. Jones, and J. Rehg, “Statistical color models with application to skin detection”, *Int. Journal of Computer Vision*, **46**(1), 2002, pp. 81-96.
15. The Champion Database, available at www.libfind.unl.edu/alumni/events/breakfast_for_champions.htm
16. R. Chellappa, C. L. Wilson, and S. Sirohey, “Human and machine recognition of faces: a survey”, *Proceedings of the IEEE*, **83**(5), 1995, pp. 705-740.
17. D. Frohlich, A. Kuchinsky, C. Pering, A. Don, and S. Ariss, “Requirements for Photoware”, *Proc. ACM Conf. on Computer Supported Cooperative Work*, New Orleans, Louisiana, USA, 2002, pp. 166-175.
18. K. Rodden, and K. Wood, “How do people manage their digital photographs ?”*, Proc. ACM Conf. on Human Factors in Computing Systems*, Fort Lauderdale, Florida, USA, 2003.
19. K.-W. Wong, K.-M. Lam, and W.-C. Siu, “A robust scheme for live detection of human faces in color images”, *Signal Processing: Image Communication*, **18**, 2003, pp. 103-114.