ENHANCING THE ROBUSTNESS OF SKIN-BASED FACE DETECTION SCHEMES THROUGH A VISUAL ATTENTION ARCHITECTURE

Konstantinos Rapantzikos[1] and Nicolas Tsapatsoulis[2]

[1]School of Electrical and Computer Engineering,
National Technical University of Athens
15780 Zografou, Greece
E-mail: rap@image.ntua.gr
Tel: +30 210 7724351, Fax: +30 210 7722492

[2]Department of Computer Science
University of Cyprus
75 Kallipoleos Str., P.O.Box 20537
CY-1678, Nicosia, Cyprus
Tel: +357 22 892746 Fax: +357 22 892701

ABSTRACT

Bottom up approaches to Visual Attention (VA) have been applied successfully in a variety of applications, where no domain information exists, e.g. general purpose image and video segmentation. In face detection, humans perform conscious search; therefore, bottom up approaches are not so efficient. In this paper we introduce the inclusion of two channels in the VA architecture proposed by Itti et al [8] to account for motion and conscious search in a scene. Increasing the channels in the architecture requires an efficient way of combining the various maps that are produced. We solve this problem by using an innovative committee machine scheme which allows for dynamically changing the committee members (maps) and weighting the maps according to the confidence level of their estimation. The overall VA architecture achieves significantly better results compared with the simple skin based face detection as shown in the experimental results.

1. INTRODUCTION

Color based face detection is one of the most popular approaches for face detection as it combines very fast implementation with rather accurate results. Among the color based algorithms that have been appeared in the corresponding literature the skin-color based are the ones that are more commonly used. The basic idea [1] is the use of color thresholding for face detection by means of a skin color model based on the chrominance components of the *YCrCb* color space and a suitable skin color distribution. However, most of the studies based on this idea reveal that considerable effort is required in post processing steps to achieve remarkable results [2][3]. Although the skin color subspace covers indeed a small area of the Cr-Cb chrominance plane, it cannot be modeled in such a general way to be efficient for all images that include faces since the influence of the luminance channel *Y* is not totally negligible. Moreover, false alarms are rather common since there is always the possibility in a scene to contain non-skin objects that have skin-like color. Finally, the compactness of the segmented skin objects is, in general, poor due to noise influence, illumination effects, and the objects' nature per se.

The negative influence of illumination has been tackled with techniques that dynamically update the skin color models through the use of mixture of Gaussians [4], Markov models [5], and histogram detectors [6]. Compactness enhancement is typically achieved through region growing techniques [2] in post processing steps. Discriminating between skin and skin color objects is not solved with the above methods. In face detection this is achieved [7] by combining skin color features with face outline and local symmetry information to locate prominent facial feature points.

In this study we propose a unified approach for enhancing the results of skin detection algorithms, concerning the above-mentioned problems, by using a visual attention architecture. The proposed scheme is based on the model of Itti et al [8] and combines five different channels: A skin channel that identifies skin-color objects based on the work presented in [9], a motion channel that identifies moving objects in a scene, an orientation channel which identifies objects with face-like structure through the use of multiscale orientation filters that are optimal for facial features detection, and the usual intensity and color channels. The use of the motion channel accounts for the problem of skin like objects; it is more likely for the skin objects to move since they are parts of the human body. The use of orientation filters accounts for the post processing task for discriminating between face and non face objects (although it should be mentioned that a more accurate technique for this purpose is still necessary as a post processing task). The aim here is mainly to enhance the skin detection methods. Illumination variations are handled both with the use of the illumination channel as well as within the skin detection channel [9]. Combination of the results of the various channels into a single saliency map is achieved through the use of a committee machine scheme which utilizes context based gating as well as confidence based weights.

2. SALIENCY-BASED VISUAL ATTENTION

The basis of many visual attention models proposed over the last two decades is the Feature Integration Theory of Treisman et al. [10] that was derived from visual search experiments. According to this theory, features are registered early, automatically and in parallel along a number of separable dimensions (e.g. intensity, color, orientation, size, shape etc).

One of the major saliency-based computational models of visual attention is presented in [8] and deals with static color images. Visual input is first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map, such that only locations that locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master saliency

map. Itti and Koch [8],[11] presented an implementation of the proposed saliency-based model. Low-level vision features (color channels tuned to red, green, blue and yellow hues, orientation and brightness) are extracted from the original color image at several spatial scales, using linear filtering. The different spatial scales are created using Gaussian pyramids, which consist of progressively low-pass filtering and sub-sampling the input image. Each feature is computed in a center-surround structure akin to visual receptive fields. Using this biological paradigm renders the system sensitive to local spatial contrast rather than to amplitude in that feature map. Center-surround operations are implemented in the model as differences between a fine and a coarse scale for a given feature. Seven types of features, for which evidence exists in mammalian visual systems, are computed in this manner from the low-level pyramids. The algorithm is summarized in Figure 1 (central part).

Motion is of fundamental importance in biological vision systems and contributes to visual attention as confirmed by Watanabe et al. in [12]. Despite the biological evidences, only few researchers [13],[14] studied the integration of motion into the saliency-based model. We consider the inclusion of motion essential in visual attention and we use a multiresolution gradient-based approach, [15], to estimate optical flow and generate a new conspicuity map in the same manner as with static maps (Figure 1-right part).

Face detection by humans is definitely a conscious process and is based on a prior model for the face built in the humans mind. The model of Itti even with the contribution of the motion channel remains simply a bottom up approach that lacks provision of conscious search. In the past it had been thought that bottom-up signals normally achieved attention capture; it is now appreciated that top-down control is usually in charge[16]. Towards this direction we attempt to integrate prior knowledge to the saliency-based model in order to draw the attention to regions with specific characteristics (Figure 1-left part). In our case we consider the face search (detection). We use a skin detector scheme presented in [9] to generate a skin map with possible face locations and link it with the other feature maps.

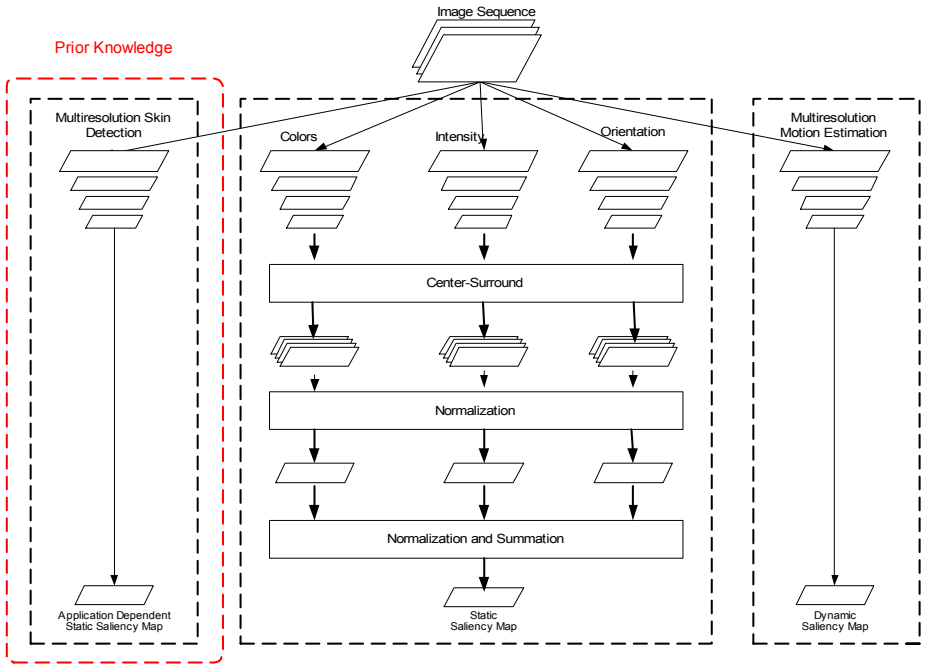

Figure 1: The proposed VA architecture. In the central part it is the model of Itti. In the two sides are our extensions for considering the influences of motion and conscious search (existence of prior knowledge - in the particular case detection of skin like objects)

3. MAP FUSION

Combining the various feature maps of the VA scheme is an important issue that affects the robustness of the skin/face detection. The lack of homogeneity of the information provided by each map requires a factual and efficient combination to make sense. In our method the combination of the feature maps is achieved through a committee machine scheme as shown in Figure 2. We should mention here that: (a) in addition to the corresponding saliency map, every channel of the VA architecture provides information about the confidence level for the estimation of this map (see Section 3.2), (b) the constituent maps of the static saliency map; that is the color, intensity, and orientation maps, are considered as independent channels.

As stated in [17], with a committee machine we can achieve a test set performance unobtainable from a single committee member and we can build a modular solution in a straightforward manner. In our case we expect a better performance for the skin-based face detection problem by combining the various channels. Furthermore, the complementary form of the feature maps requires, indeed, a modular solution.

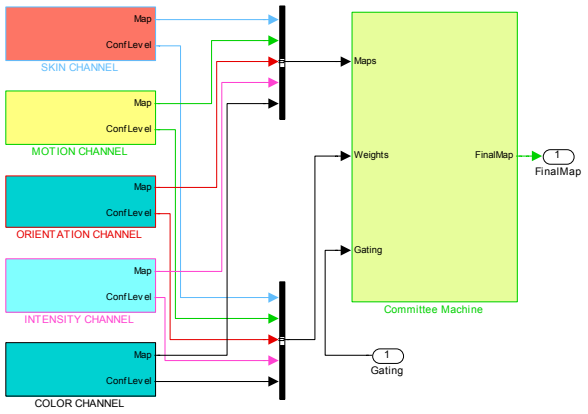

Figure 2: A committee machine for map fusion

3.1. The Committee Machine

The Committee Machine combines the individual maps taking into account the current context and the confidence levels provided by each channel. The *Gating* input models the current context in a simple way. In cases where prior information about the existence of humans in a scene is available, i.e., face databases, TV news, etc., this information is used to guide the Committee Machine to rely mainly on the skin and motion channels. Similarly, in cases where the scene is mainly static and it is not guaranteed that there are any humans in the scene the static channels (color, intensity, orientation) are given priority. In cases where no prior information about the scene exists (default case) then the *Gating* input is simply a vector consisting of some pre-calculated weights defining the importance of the various feature maps.

The confidence levels of the various channels are used to define which maps will, finally, attend the committee and with what weight. In cases where no gating information is available this works as follows: Only three maps are allowed to attend the Committee, the one being always the skin map. The maps from the other four channels that present the two lowest confidence levels are left out.

In case of a static context the motion map is always left out as well as the static channel (color, intensity, orientation) with the lowest confidence level. In scenes containing humans the two static channels with lowest confidence levels are left out. The functionality of the committee machine is summarized below.

Let I be the current video frame, g_i be the gating input for the i-th map, $y_i = f_i(I)$ be the i-th map and $c_i(I)$ be the corresponding confidence level. Then the final map is given by:

$$F(I) = \frac{\sum_{i=1}^{NumOfChannels} g_i \cdot c_i(I) \cdot f_i(I)}{\sum_{i=1}^{NumOfChannels} g_i \cdot c_i(I)} \qquad (1)$$

where by *NumOfChannels* we denote the available channels in the VA architecture (in our case 5 channels).

It is clear from the above description that the set of maps participating in the committee changes from frame to frame keeping pace with possible changes in the video sequence.

3.2. Confidence level estimation

It has been seen in the previous paragraph that the confidence levels for the estimation of the individual feature maps are multiplied with the gating weights of the corresponding maps in the committee (see Eq. 1). The confidence levels are estimated per channel and are based on inter-frame correlation of the maps. This aims at compensating for rapid or abrupt changes (lighting conditions, abrupt motion, degraded quality due to compression artifacts etc) from the previous to the current feature map. A simple correlation factor between the two temporally neighboring maps is calculated and serves as a way to determine the amount of contribution of each map to the saliency in terms of consistency and strength. If, for example, motion is consistent among frames the correlation factor will be high, but if a motion discontinuity occurs (e.g. due to scene change or abrupt motion) the correlation value will be low and the motion map will be less weighted than before. Formally, the confidence level of the i-th map in frame I is computed by:

$$c_i(I) = \frac{\sum \left(f_i(I-1) \cap f_i(I) \right)}{\sum \left(f_i(I-1) + f_i(I) - f_i(I-1) \cap f_i(I) \right)} \qquad (2)$$

where $\cap$ denotes a multilevel AND operation (i.e., minimum) and the summation being over all values of the i-th feature map.

4. EXPERIMENTAL RESULTS

Three representative video sequences were used for testing the performance of the VA architecture. Due to lack of space, the sequences, the results and ground truth for each frame (generated in a manual way) along with explanatory figures can be found on-line (http://www.image.ntua.gr/~rap/icip05/VA/face/). In order to present a fair comparison between VA-based and skin-based face detection we apply a minimum inter variance threshold value [18] so as to provide a final decision on skin and non-skin areas. The maps are then thresholded at this level and the average precision / recall is calculated for the whole clip. The results are summarized in Table 1.

The first sequence, named "*twoFaces*", is captured from a TV clip and is recorded from a static camera. The motion scenario is relative simple, but the quality of the sequence can pose difficulties to the skin detection algorithm due to several skin-like (in terms of color) regions. Additionally, the low frame rate of the clip produces unavoidable motion discontinuities, which may negatively affect the motion estimation algorithm. However, in the VA method this is compensated for in the committee machine where the motion channel is blocked from attending the committee due to low confidence level.

The second sequence, named "*myFace*" is recorded by a static personal video camera and shows a moving head under

non-uniform illumination conditions. The VA method is automatically adapted compensating for the illumination irregularity, performing significantly better than the simple skin detection scheme.

A hard case in terms of skin detection is examined in the third video sequence (named "*grandma*"). This clip exhibits global (camera) and local (face-arms) motion and degraded quality (noise). Precision and recall values are similar for both algorithms tested.

Video Seq.	***Method***	***Mean Precision (%)***	***Mean Recall (%)***
twoFaces	VA	78.9	80.5
	Simple Skin	71.6	73.4
myFace	VA	80.4	79.8
	Simple Skin	73.3	73.0
grandma	VA	55.9	56.3
	Simple Skin	49.5	51.5

Table 1: Comparison of VA and skin-based face detection in three video sequences

5. CONCLUSIONS AND FURTHER WORK

In this paper we elaborate on the combination of saliency-based visual attention with a committee machine in order to enhance robustness against common pitfalls in skin-based face detection. We have proposed an extension to the VA model of Itti *et al* [8] by including two more maps; the first dealing with motion and the second introducing top-down information that accounts for the conscious search performed by humans when looking for faces in a scene.

The feature maps are fused in dynamically updated manner with the use of a Committee Machine scheme. In this way, and by considering the scene context as gating information and the confidence levels of feature maps estimation as weights enable adaptation to various conditions that decrease the performance of skin-based face detection techniques.

The experimental results are promising. The VA method outperforms the simple skin-detection method in a variety of common situations. However, we plan to undertake more experiments on challenging sequences and against third party implementations of skin detection. Furthermore, we currently work towards an online learning implementation of the Committee Machine to examine the importance of supervision in map fusion.

Acknowledgments

This research work was supported (in part) by OPTOPOIHSH (Development of knowledge-based Visual Attention models for Perceptual Video Coding, 1104/01, funded by the Cyprus Research Promotion Foundation) and the European network of Excellence MUSCLE.

6. REFERENCES

[1]. H. Wang H, and S-F.Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video," *IEEE Transactions Circuits and Systems for Video Technology*, vol. 7(4), pp. 615-628, 1997.

[2]. C. Garcia, and G. Tziritas, "Face Detection Using Quantized Skin color Regions Merging and Wavelet Packet Analysis," *IEEE Transactions on Multimedia*, vol. 1(3), pp. 264-277, 1999.

[3]. Y. Avrithis, N. Tsapatsoulis N, and S. Kollias, "Color-Based Retrieval of Facial Images," *Proceedings of EUSIPCO*, Tampere, Finland, September 2000.

[4]. Y. Raja, S. J. McKenna, and S. Gong, "Tracking and Segmenting People in Varying Lighting Conditions Using Color," *Proceedings of 3rd Int. Conference on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998.

[5]. L. Sigal, S. Sclaroff, and V. Athitsos, "Estimation and Prediction of Evolving colour: Distributions for Skin Segmentation Under Varying Illumination," *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, Hilton Head Island, SC, June 2000.

[6]. M. J. Jones, M. R. Regh, "Statistical colour Models with Application to Skin Detection," *Compaq Cambridge Research Lab Technical Report CRL* 98/11, 1998.

[7]. Q. B. Sun, W. M. Huang, and J. K. Wu, "Face Detection Based on colour and Local Symmetry Information," Proceedings of 3rd International Conference on Automatic Face and Gesture Recognition, Nara, Japan, April 1998.

[8]. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(11), pp. 1254-1259, 1998.

[9]. N. Tsapatsoulis N, Y. Avrithis, and S. Kollias, "Facial Image Indexing in Multimedia Databases," *Pattern Analysis and Applications: Special Issue on Image Indexation*; vol. 4(2/3), pp 93-107, 2001.

[10] A. M. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12(1), pp. 97-136, 1980.

[11] L. Itti, and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489-1506, 2000.

[12] T. Watanabe T, Y. Sasaki, S. Miyauchi, B. Putz, N. Fujimaki, M. Nielsen, R. Takino, and S. Miyakawa, "Attention-regulated activity in human primary visual cortex," *Journal of Neurophysiology*, vol. 79, pp. 2218-2221, 1998.

[13]. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modelling visual attention via selective tuning," *Artificial Intelligence*, vol. 78 (1-2), pp. 507-545, 1995.

[14]. A. Maki, P. Nordlund, and J. O. Eklundh, "Attentional scene segmentation: Integrating depth and motion from phase," *Computer Vision and Image Understanding*, vol. 78, pp. 351-373, 2000.

[15]. M. J. Black, and P. Anandan "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields," *Computer Vision and Image Understanding*, vol. 63(1), pp. 75-104, 1996.

[16]. H. Pashler, "Attention and performance," *Ann. Rev. Psych*, vol. 52, pp. 629-651, 2001.

[17] V. Tresp, "Committee Machines", *Handbook for Neural Network Signal Processing*, Yu Hen Hu and Jeng-Neng Hwang (eds.), CRC Press, 2001.

[18] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 62-66, 1979