A committee machine scheme for feature map fusion under uncertainty: the face detection case

Konstantinos E. Rapantzikos*
School of Electrical and Computer Engineering,
National Technical University of Athens,
15780 Zografou, Greece
Fax: +30 210 7722492 E-mail: rap@image.ntua.gr
*Corresponding author

Nicolas Tsapatsoulis
Department of Computer Science,
University of Cyprus,
75 Kallipoleos Str., P.O. Box 20537,
CY 1678, Nicosia, Cyprus, Greece
Fax: +357 22 892701 E-mail: nicolast@ucy.ac.cy

Abstract: Feature map fusion in Visual Attention (VA) models is by definition an uncertain procedure. One of the major impediments in extending the static VA architecture proposed by Itti et al. (2000) to account for motion or other information is the lack of justification on how to integrate the various channels. We propose an innovative committee machine scheme that allows for dynamically changing the committee members (maps) and weighting them according to the confidence level of their estimation. Through this machine we handle the extensions on Itti’s model; we add a motion channel and a prior knowledge channel which accounts for the conscious search performed by humans when looking for faces in a scene. The experimental results, obtained when considering face detection, show that the map fusion, through the proposed committee machine, leads to significantly better statistical results when compared with the simple skin-based face detection method.

Keywords: committee machine; map fusion; visual attention; face detection.


Biographical notes: Konstantinos E. Rapantzikos received the Diploma and the MS Degree in Electronic and Computer Engineering from the Technical University of Crete, Greece in 2000 and 2002 respectively. He is currently working towards the PhD Degree in Electrical Engineering at the National Technical University of Athens, Greece. He has published five articles in international journals and books and 12 in proceedings of international conferences. His current research interests include computational modelling of human vision, spatiotemporal segmentation, motion estimation in compressed and uncompressed video and biomedical imaging.
A committee machine scheme for feature map fusion

Nicolas Tsapatsoulis graduated from the Department of Electrical and Computer Engineering, the National Technical University of Athens in 1994 and received his PhD Degree in 2000 from the National Technical University of Athens. In December 2000 he joined the Image, Video and Multimedia Systems Lab (IVML) of NTUA, as a Post Doctoral Researcher. Since September 2003 he has the post of Visiting Assistant Professor at the same Department. His publication record includes more than 55 papers in scientific journals, book chapters conference proceedings and at least 130 citations. His current research interests lie in the areas of visual attention, facial image processing, and machine vision.

1 Introduction

Committee machines or mixtures of experts are well-known methods for combining learning machines. Given that these machines can be used in both a supervised and an unsupervised manner, we consider them as an appropriate solution for information fusion from different sources. In general, there are three reasons for using committee machines (Tresp, 2001):

- a committee can achieve a test set performance unobtainable from a single committee member
- through committee machines, modular solutions can be obtained in a straightforward manner
- reduction of computational complexity can be achieved by partitioning a given training set to smaller datasets and allowing the committee members to be trained on different sets.

In our case, we test the potential of such a fusion scheme by applying it to the intermediate outputs of a saliency-based visual attention model biased by a skin detection module. Face detection is the goal of the algorithm, and searching for face-like objects is considered prior knowledge. The proposed scheme is based on the model of Itti et al. (2000) and combines five different channels (Rapantzikos and Tsapatsoulis, 2003): A skin channel that identifies salient skin objects based on the work presented in Tsapatsoulis et al. (2001), a motion channel that identifies salient moving objects in a scene, an orientation channel which identifies objects with face-like structure through the use of multi-scale orientation filters that are optimal for facial features detection and the usual intensity and colour channels.

We propose a committee machine as a fusion operator, since all the three reasons above hold: We need a better performance for the skin-based face detection problem by combining the various channels and a modular solution for the complementary form of the feature maps. Skin-based face detection using colour is one of the most popular approaches for face detection as it combines very fast implementation with rather accurate results. The work presented in Wang and Chang (1997) received considerable attention and inspired many researchers for a variety of implementations. The basic idea of Wang and Chang (1997) is the use of colour thresholding for face detection by means of a skin colour model based on the chrominance components of the YCrCb colour space and a suitable skin colour distribution. However, most of the studies based on this idea
reveal that considerable effort is required in post processing steps to achieve remarkable results (Garcia and Tziritas, 1999; Avrithis et al., 2003). Although the skin colour subspace covers indeed a small area of the Cr-Cb chrominance plane, it cannot be modelled 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 of having non-skin objects in a scene that have skin like colour. Finally, the compactness of the segmented skin objects is, in general, poor owing to noise influence, illumination effects and the objects’ nature per se (for example the presence of eyes, eyebrows and teeth in a face).

The influence of illumination has been tackled with techniques that dynamically update the skin colour models through the use of a mixture of Gaussians (Raja et al., 1998), Markov models (Sigal et al., 2000) and histogram detectors (Jones and Regh, 2002). Compactness enhancement is typically achieved through region growing techniques (Garcia and Tziritas, 1999) in post processing steps. The problem of discriminating between skin and skin colour objects is not solved with the above methods. In the particular case of face detection, this is achieved (Sun et al., 1998) by combining skin colour features with face outline and local symmetry information to locate potential facial feature points. A general-purpose colour segmentation approach, combined with a Gaussian skin colour model and shape filtering is also proposed in Tsapatsoulis et al. (2000) for face detection.

In this study, we propose a unified approach for enhancing the results of skin detection algorithms, as far as the above-mentioned problems are concerned, by using the extended version of visual attention model that splits the input into five different 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. Nevertheless, it should be mentioned that a more accurate technique for this purpose is still necessary. The aim here is only to enhance the skin detection methods. Illumination variations are handled both with the use of the illumination channel and within the skin detection channel as will be explained in Section 3. Combination of the results of the various channels into a single saliency map is achieved through the use of a committee machine scheme which utilises gating as well as confidence based weights. Details for these will be given in Section 4.

2 Saliency-based visual attention schemes

The basis of many visual attention models proposed over the last two decades can be found in the Feature Integration Theory of Treisman and Gelade (1980) 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, colour, orientation, size, shape, etc). Koch and Ullman (1985) have suggested a model based on this theory that leads to the generation of a saliency map. The saliency map consists of a retinotopic grid of neurons, which accumulate bottom-up information from various feature maps. Activity of a neuron in this map is proportional to the relevance of the corresponding location. In primates, such a map is believed to be located in the posterior parietal cortex (Gottlieb et al., 1988) as well as in the various visual maps in the pulvinar nuclei of the thalamus (Robinson and Peterson, 1992). Meaningful objects
A committee machine scheme for feature map fusion

(conjunction of features) are identified at a second stage, which requires focused attention. At the interface between the first and second stages, there is a bottleneck that functions as a gate allowing only part of the visual information to proceed to the second stage.

One of the major saliency-based computational models of visual attention is presented in Itti et al. (1998) and deals with static colour 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 et al. (1998) and Itti and Koch (2000) presented an implementation of the proposed saliency-based model. Low-level vision features (colour channels tuned to red, green, blue and yellow hues, orientation and brightness) are extracted from the original colour 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 centre-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. Centre-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 summarised 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. (1998). Despite the biological evidence, only few researchers studied the integration of motion into the saliency-based model. Several authors Tsotsos et al. (1995), Maki et al. (2000) and Milanese et al. (1995) attempted the integration of dynamic features, but their methods lack the interaction between the two different feature classes (static-dynamic) to build a global attention map. We use a multi-resolution gradient-based approach, (Black and Anandan, 1996), 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 of the face built in the human mind. The model of Itti et al. even with the contribution of the motion channel is still 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. Involuntary attention capture by distracting inputs occurs only if they have a property that a person is using to find a target (Pashler, 2001). 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) as the prior knowledge. We use a skin detector scheme to generate a skin map with possible face locations and link it with the other feature maps.
3 Skin detection

This section describes the method followed within the skin channel to locate the probable skin segments, i.e., image/frame areas that contain objects with colour similar to the one of the human skin, and produce the skin map. It is stated in some classic studies (Rzeszewski, 1975; Harwood, 1976) that skin-tone colours are spread over a small area of the \( Cr-Cb \) chrominance plane of the \( YCrCb \) colour model. Wang and Chang (1997), based on that idea, presented a fast face detection algorithm that inspired many researchers. In a similar way, we approximated skin-tone colours distribution using a two-dimensional Gaussian density function. Expanding the idea of Wang, we use a feedback model for reestimating skin colours based on the current image or the previous video frame to account for illumination variations (Tsapatsoulis et al., 2001).

The mean vector \( \mu_0 \) and the covariance matrix \( \Sigma \) of the human skin chrominance components are initially estimated from training data containing facial areas of different races, obtained from regular TV clips, colour images and personal video cameras. Then, according to the Gaussian distribution:

\[
P(x \mid \mu_0, \Sigma) = \frac{\exp \left\{ -\frac{1}{2} (x - \mu_0)^T \Sigma^{-1} (x - \mu_0) \right\}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}}
\]  

(1)
where \( k = 2 \) is the number of chrominance components; the likelihood of an input
pattern \( x \) (chrominance components of a particular pixel or average chrominance
components of an image block) can be approximated by the quantity:

\[
P(x) = \exp \left\{ -\frac{1}{2} (x - \mu_0)^T C^{-1} (x - \mu_0) \right\}.
\]  

Although using the above Gaussian model is rather efficient for the classification of
pixels as skin and non-skin ones, especially in a controlled illumination environment,
better performance is obtained by reestimating the mean vector \( \mu_0 \) based on the current
image/previous video frame. In particular, the initial Gaussian distribution is used for a
first pass classification, employing a threshold value, which is based on the statistics of
the likelihood on the current image, so that it adapts to varying lighting conditions.

Let \( I \) be the current frame of dimensions \( M \times N \), and \( I_C \) and \( I_Cb \) its \( Cr \) and \( Cb \)
components, respectively. Let also \( p(x(s)) \) be the likelihood of pixel \( s \) with chrominance
components \( x(s) = [I_C(s) \ I_Cb(s)] \). Then the threshold value is selected as \( T_I = \mu_I + \sigma_I \),
where

\[
\mu_I = \frac{1}{MN} \sum_{s=1}^{MN} p(x(s))
\]

is the average likelihood of image \( I \) and

\[
\sigma_I = \frac{1}{MN-1} \left\{ \sum_{s=1}^{MN} (p(x(s)) - \mu_I)^2 \right\}
\]

is the corresponding standard deviation. The pixels that are classified as skin ones are
used for the reestimation of \( \mu_0 \) according to the following equation:

\[
\mu_0 = (1 - m) \cdot \mu + m \cdot \mu_I
\]

where \( \mu \) is the mean chrominance vector of image segments classified as skin ones,
estimated from of the current image, and \( m \) is a memory tuning constant. The value of \( m \)
defines the amount of model’s readjustment. A high value prevents the initial model from
being significantly altered while a small one quickly adapts the model to the current data.
It can be estimated by measuring the false alarm/dismissal rates or the determinant of the
confusion matrix as a function of \( m \) and selecting a suitable point on the corresponding
graph (ROC curve) (Sigal et al., 2000). From our experiments, \( m = 0.7 \) seems to be a
good compromise. Note also that the adaptation of the covariance matrix \( C \) is also
possible; however, its effect on classification performance turns out to be insignificant.

Equation (3) is also used to adapt the model in a fully dynamic environment. In video
sequences, the model is reestimated in a per frame basis giving a robust framework for
skin segment tracking. In particular, the first-pass classification is performed in the first
video frame, using threshold \( T_I \). This scheme guaranties that there will always be some
image parts classified as skin ones, and their average chrominance vector \( \mu \) is used for
model adaptation. In subsequent frames, only the final classification is employed, using
the minimum risk threshold; the adaptation of the model is performed based on the skin
areas of the previous frame. If there is at least one skin segment detected, its average
chrominance vector is used for adaptation; if, however, no skin segments were found in
the previous frame, the first-pass classification is also performed in the current frame, the
model is adapted and final classification is employed again. If the final binary mask produced after adaptation still contains no skin segments, the previous value of $\mu$ is restored as in the case of static images.

The overall scheme manages to keep track of illumination changes in video sequences and at the same time handles the appearance of skin segments with new colour characteristics without deviating from the generic skin colour model, thus risking the possibility of a high false alarm rate.

4 MAP fusion

The combination of the feature maps is achieved through a committee machine scheme as shown in Figure 2. Before proceeding with the discussion, we should mention that

- 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 (more on this is in Subsection 4.1)
- the constituent maps of the static saliency map, that is the colour, intensity and orientation maps, are considered as independent channels.

Figure 2 A committee machine for 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. The Committee Machine combines the individual maps taking into account the current context and the confidence levels provided by each channel. The input, named Gating, models the current context in a very 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 (colour, 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 precalculated 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 (colour, intensity and orientation) with the lowest confidence level. In case of scenes containing humans, the two static channels with lowest confidence levels are left out. The functionality of the committee machine is summarised below. Let $I$ be the current video frame of, $g_i$ 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:

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

(4)

where by $\text{NumOfChannels}$ we denote the available channels in the VA architecture (in our case $\text{NumOfChannels} = 5$).

It becomes evident that the set of maps attending the committee, changes from frame to frame keeping pace with corresponding changes in the video sequence.

4.1 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. The confidence levels are estimated per channel and are based on an inter-frame correlation of the maps. This aims at compensating for rapid or abrupt changes (lighting conditions, abrupt motion, degraded quality owing to compression artefacts, etc.) from the previous to the current feature map. A simple correlation factor between the two temporally neighbouring 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., owing 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:

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

(5)

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

5 Experimental results

We tested the visual attention-based face detector using a set of video clips obtained from regular TV clips and personal video cameras. The ground truth for each frame is generated in a manual way. Figure 3 illustrates an example of the results obtained by applying the proposed VA architecture on the first two frames of a video sequence called ‘twoFaces’. In Figure 4, the detection potential of the proposed algorithm on three representative cases from the available set of video clips is presented. The first two rows show the original frames and the corresponding ground truth masks, while the last two rows show the thresholded results of both the simple skin detection (use of skin channel only) and the VA-based enhancement. We use the same parameters for the skin detection algorithm throughout the video clips.

In order to present a (as far as possible) fair comparison between VA-based and skin-based face detection, we apply a minimum inter variance threshold value (Otsu, 1979) 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 (Table 1).

<table>
<thead>
<tr>
<th>Video seq.</th>
<th>Method</th>
<th>Mean precision (%)</th>
<th>Mean recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TwoFaces</td>
<td>VA</td>
<td>78.9</td>
<td>80.5</td>
</tr>
<tr>
<td></td>
<td>Simple skin</td>
<td>71.6</td>
<td>73.4</td>
</tr>
<tr>
<td>myFace</td>
<td>VA</td>
<td>80.4</td>
<td>79.8</td>
</tr>
<tr>
<td></td>
<td>Simple skin</td>
<td>73.3</td>
<td>73.0</td>
</tr>
<tr>
<td>grandma</td>
<td>VA</td>
<td>55.9</td>
<td>56.3</td>
</tr>
<tr>
<td></td>
<td>Simple skin</td>
<td>49.5</td>
<td>51.5</td>
</tr>
</tbody>
</table>

Table 1 Summary of results in three video sequences

Three video sequences were used for testing the performance of the VA architecture. 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 owing to several skin-like (in terms of colour) regions. Additionally, although the motion is simple, the low frame rate of the clip, which produces unavoidable motion discontinuities, 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 owing 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-based technique is automatically adapted and compensated for the illumination irregularity and performs 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’) corresponding to the last column of Figure 4. This clip exhibits global (camera) and local (face-arms) motion and degraded quality (noise). Precision and recall values are similar for both algorithms tested.

Figure 3  An example of saliency map extraction using the VA architecture
Figure 4  Characteristic examples of face detection in three video sequences. In the first row are shown the original frames, in the second row the manually extracted ground truth, in the third row we show the results of the skin detection channel and in the last row we present the detection obtained through the VA architecture. The same threshold values are used both for skin- and VA- detection in order to provide a fair comparison.

6 Conclusions and further work

In this paper, we elaborate on the use of a committee machine as an alternative way to combine (fuse) the intermediate maps of an extended saliency-based visual attention scheme in order to enhance robustness against common pitfalls in skin-based face detection. The proposed extension on the visual attention modules the inclusion of two more maps; the first dealing with motion and the second introducing top-down information related to the task at hand. In our case, the last map accounts for the conscious search performed by humans when looking for faces in a scene.

The feature maps are fused in a dynamically updated manner with the use of a Committee Machine scheme and therefore adapt to various conditions that decrease the performance of skin-based face detection techniques. The dynamically changing
parameters of the scene (e.g., illumination) are incorporated into a scene context model that is used as gating information in the committee machine. This gating combined with the weighted feature maps, obtained through a confidence-based scheme, enables the adaptation to the scene.

The experimental results are quite promising, since the VA method outperforms the simple skin-detection method in a series of natural videos. However, more experiments are required to prove the efficiency of the proposed method and to test its performance 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 and simulate better the ability of the human visual system to recognise easier/faster the familiar faces.

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.

References


