An Embedded Saliency Map Estimator Scheme: Application to Video Encoding

Nicolas Tsapatsoulis[1], Konstantinos Rapantzikos[2] and Constantinos Pattichis[1]

[1] Department of Computer Science, University of Cyprus, CY 1678, Cyprus (phone: +357-2289-2697; fax: +357-2289-2701; email: {nicolast,pattichi}@ucy.ac.cy)

[2] School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str.,15780, Zografou, Greece (phone: +30-210-7724351; fax: +30-210-7722492; email: rap@image.ntua.gr)

Abstract. In this paper we propose a novel saliency-based computational model for visual attention. This model processes both top-down (goal directed) and bottom-up information. Processing in the top-down channel creates the so called skin conspicuity map and emulates the visual search for human faces performed by humans. This is clearly a goal directed task but is generic enough to be context independent. Processing in the bottom-up information channel follows the principles set by Itti *et al* but it deviates from them by computing the orientation, intensity and color conspicuity maps within a unified multi-resolution framework based on wavelet subband analysis. In particular, we apply a wavelet based approach for efficient computation of the topographic feature maps. Given that wavelets and multiresolution theory are naturally connected the usage of wavelet decomposition for mimicking the center surround process in humans is an obvious choice. However, our implementation goes further. We utilize the wavelet decomposition for inline computation of the features (such as orientation angles) that are used to create the topographic feature maps. The bottom-up topographic feature maps and the top-down skin conspicuity map are then combined through a sigmoid function to produce the final saliency map. A prototype of the proposed model was realized through the TMDSDMK642-0E DSP platform as an embedded system allowing real-time operation. For evaluation purposes, in terms of perceived visual quality and video compression improvement, a ROI-based video compression setup was followed. Extended experiments concerning both MPEG-1 as well as low bit-rate MPEG-4 video encoding were conducted showing significant improvement in video compression efficiency without perceived deterioration in visual quality.

1 Introduction

A popular approach to reduce the size of compressed video streams is to select a small number of interesting regions in each frame and to encode them in priority. This is often referred to as RegionOf-Interest (ROI) coding [24]. The rationale behind ROI-based video coding relies on the highly non-uniform distribution of photoreceptors on the human retina, by which only a small region of visual angle (the fovea) around the center of gaze is captured at high resolution, with logarithmic resolution falloff with eccentricity [31]. Thus, it may not be necessary or useful to encode each video frame with

uniform quality, since human observers will crisply perceive only a very small fraction of each frame, dependent upon their current point of fixation.
A variety of approaches have been proposed in the literature for ROI estimation [24]. In most of them the definition of ROI is highly subjective; that is, they lack scientific evidence in supporting their claim that the areas defined as ROIs are indeed regions of interest for the most of human beings. In this paper we attempt to model ROIs as the visually attended areas indicated by a saliency map [15] in order to lower as much as possible the subjectivity in selecting ROIs. Estimation of the saliency map is performed though a novel model in which both top-down (goal oriented) and bottom-up information is utilized. Furthermore, emphasis was put to the minimization of computational complexity through inline computation of the features used for the computation of the various topographic maps.
In saliency-based visual attention algorithms efficient computation of the saliency map is critical for several reasons. First, the algorithm itself should model appropriately the visual attention process in humans..Visual attention theory has been constructed mainly by neuroscientists without taking into account computational modelling difficulties. On the other hand, computational models have been developed mainly by engineers and computer scientists which in several cases compromise theory in favor of implementation efficiency. Second, algorithm's implementation should conform to real life situations and settings. Perceptual based video coding is one of the areas that visual attention fits well. However, in applications like video-telephony real-time video encoding is required. Therefore, if a computational model of visual attention is to be used, then its implementation should be both fast and effective. Finally, integration of the topographic feature maps into the overall saliency map should be performed in a reasonable way and not ad hoc as it happens in most existing models where normalization and additions is the combination method of preference.
In the proposed model for saliency map estimation, all the above mentioned issues were taken into account. A computationally efficient way for identifying salient regions in images, based on bottom-up information, by utilizing wavelets and multiresolution theory is employed. Furthermore, a top-down channel, emulating the visual search for human faces performed by humans has been also added. This goal oriented information is justified by the fact that in several applications like visual-telephony and teleconferencing the existence of, at least, one human face in every video frame is almost guaranteed. Therefore, it is anticipated that the first area to receive the human attention is the face area. However, bottom-up channels remain in process modelling sub-conscious visual attention attraction. Finally, the overall saliency map estimator is implemented as an embedded system to allow real-time video encoding.

2 Saliency-based Visual Attention

2.1 Existing computational models

The idea of attention deployment dates back to the pioneering work of James [14], the father of American psychology. Several theoretical models have been proposed in the past using the two component attention framework, consisting of a top-down and a

bottom-up component, as proposed by James. Computational modelling of this theoretical framework has been an important challenge for researchers both with a neuroscience and engineering background. It is widely accepted that one of the first neurally plausible computational architecture for controlling visual attention was proposed by Koch & Ullman [15]. The core idea is the existence of a saliency map that combines information from several feature maps into a global measure where points corresponding to one location in each feature map project to single units in the saliency map. Attention bias is then reduced to drawing attention towards high activity locations of this map.

Influenced by the work of Koch & Ullman several successful computational models have been built around the notion of a saliency map. The Guided Search model of Wolfe [33] hypothesizes that attentional selection is the net effect of feature maps weighted by the task at hand (requires statistical knowledge) and bottom-up activation based on local feature differences. Hence, local activity of the saliency map is based both on feature contrast and top-down feature weight. The FeatureGate model of Cave [1] provides a full neural network implementation of a similar system that combines top-down with bottom-up mechanisms. The more recent work of Torralba [27], who relates attention to visual context of a scene, is also in a similar vein. One of the most successful saliency-based models was proposed by Itti *et al.* [13][12]. It is based on the principles governing the bottom-up component of the Koch & Ullman's scheme. Visual input is first decomposed into a set of topographic feature maps. Different locations then compete for saliency independently within each map, so that only locations that locally stand out from their surround persist. This competition is based on center-surround-differences akin to human visual receptive fields. These center-surround operations are implemented in the model as differences between a fine and coarse scale for a given feature. Finally, all feature maps are linearly combined to generate the overall saliency map. Itti's model has been widely used in the computer vision community, since it provides a complete front-end for the analysis of natural color images. Applications include experimental proof of model's relation to human eye fixations, target detection [12]and video compression [11].

Tsotsos *et al.* [30] proposed a different view of attentional selection. Their selective tuning model applies attentional selection recursively during a top-down process. This means that attentional selection does not occur either at the top of the processing hierarchy or at several levels during the bottom-up process. First the strongest item is selected at the top level of the processing hierarchy using a Winner-Take-All (WTA) process (the equivalent of a saliency map). Second, the hierarchy of a WTA process is activated to detect and localize the strongest item in each layer of representation, pruning parts of the pyramid that do not contribute to the most salient item and continuously propagating changes upwards. Overall, saliency is calculated on the basis of feed-forward activation and possibly additional top-down biasing for certain locations of features. It is important to note that the saliency map is not the only computational alternative for the bottom-up guidance of attention. Desimone & Duncan [2], Hamker [7] and several other researchers argue that salience is not explicitly represented by specific neurons (saliency map), but instead is implicitly coded in a distributed manner across the various feature maps. Nevertheless, these models are not directly related to the proposed method and are not further examined

In the majority of the studies mentioned above the term top-down refers to the way computation of the saliency maps is performed. That is, in contrary to bottom-up approaches in the top-down ones attention is considered to be directed in a rather large scene area and then within this area the attentional selections are identified. There is, however, another view of top-down attention selection. In this alternative view, frequently referred to as 'goal directed attention', attentional selection is based on the principle that humans are directing their attention on known specific targets (i.e., searching for green cars). Unfortunately, it is impossible to model every possible target for every human and under any context. Therefore, the majority goal directed approaches simply outline a general framework [4], [5], [20] on how to use goal knowledge for identifying visually salient areas.

2.2 The proposed Visual Attention Model

The proposed Visual Attention model is illustrated by the architectural diagram of Figure 1. Saliency map computation is based both on bottom-up and top-down (goal directed) information. The input sequence is supposed to contain regions of interest and non important distractors or background areas. The role of the top-down component, depicted on the left, is to bias the attention system towards these regions of interest that can be statistically modelled using prior knowledge. Any region of interest may be modelled using statistical methods and inserted to this component. Nevertheless, since statistical modelling is not the focus of the proposed method, we only use a previously developed [29] statistical model for human skin representation, used for face detection, to test the proposed scheme. Faces are probably the only kind of objects in which attention to them is natural to be drawn independently of context. In the proposed model, another conspicuity map, the skin map, is computed based on the color similarity of objects with human-skin. The skin map is modulated through multiplication with a texture map so as to emphasize on structured skin areas which have a high probability to correspond to human faces. The modulating texture map is created through range filtering [6] of the intensity channel (see also Figure 3).
Interaction between top-down and bottom-up sub-systems is performed through modulation and takes place both within feature integration (feature level) and across feature integration (result level). The feature level is related to the bias applied to specific features in order to enhance regions similar to the prior model. For the sake of clarity there is an intentional redundancy in Figure 1: skin areas may be enhanced using a combination of *Cb* and *Cr* channels as indicated by the corresponding arrow; nevertheless, we add an independent skin detection module to show the creation of skin map as well as the combination across the conspicuity maps at the result level.
Let us consider a video-telephony scenario where the video stream contains one or more faces. The proposed scheme is, then, activated as follows:

(a) The skin model is selected as the one to bias further analysis and the input sequence is decomposed into different feature dimensions;
(b) Each of the feature maps is transformed in the wavelet domain and center-surround differences are independently applied. The center-surround operator is applied between a coarse and a finer scale and aims at enhancing areas that pop-out from their surroundings;

(c) the intermediate results (conspicuity maps) are modulated by the top-down gains computed using the selected prior model (this is not necessary for the intensity channel, since only a single conspicuity map exists)

(d) all conspicuity maps are again weighted using top-down gains and finally fused to generate the saliency map.

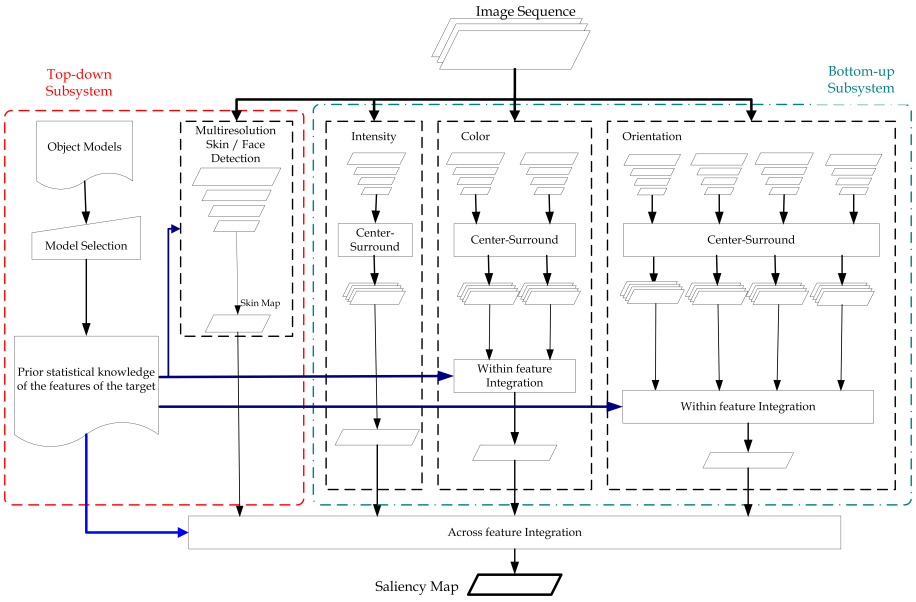


Fig. 1. The overall architecture of the Wavelet-based VA model for saliency map estimation

2.3 Bottom up Saliency-map computation

Multiscale analysis based on wavelets [16] was adopted for computing the center-surround structure in the bottom-up conspicuity maps. The center-surround scheme of Itti *et al.* [12] includes differences of Gaussian filtered versions of the original image and is in a way similar to a Laplacian pyramid that is constructed by computing the difference between a Gaussian pyramid image at a level L and the coarser level $L+1$ after it has been expanded (interpolated) to the size of level L. The Laplacian pyramid, therefore, represents differences between consecutive resolution levels of a Gaussian pyramid, similar to how the wavelet coefficients represent differences. Two main characteristics differentiate Laplacian from wavelet pyramids: (a) A Laplacian pyramid is overcomplete and has 4/3 as many coefficients as pixels, while a complete wavelet transform has the same number of coefficients and, (b) the Laplacian pyramid localizes signals in space, but not in frequency as the wavelet decomposition does.

The YCbCr color space (instead of the RGB used by Itti [13], was selected, first to keep conformance with the face detection scheme [29], and second to use the decorrelated illumination channel *Y* for the intensity and orientation conspicuity maps derivation and the *Cb, Cr* channels for the color conspicuity map. In this way parallel construction of the conspicuity maps is allowed and the required transformations of one color model to another is minimized.

Let's consider a color image *f*, transformed into the YCrCb color space. Channel *Y* corresponds to the illumination, and can be used for identifying outstanding regions according to illumination and orientation, while *Cr* (Chrominance Red) and *Cb* (Chrominance Blue) correspond to the chrominance components and can be used to identify outstanding color regions. In the proposed implementation salient areas based on intensity, orientation, and color are computed in several scales. In this way, outstanding areas of different sizes are localized. Combining the results of intensity, orientation, and color feature maps at various scales provide the intensity (C_I), orientation (C_O) and color (C_C) conspicuity maps. The motivation for the creation of the separate conspicuity maps is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency map. Hence, intra-feature competition is followed by competition among the three conspicuity maps to provide the bottom-up part of the saliency map.

In order for multiscale analysis to be performed a pair of low-pass $h_\phi(\cdot)$ and high-pass filter $h_\psi(\cdot)$ are applied to each one of the image's color channels *Y, Cr, Cb*, in both the horizontal and vertical directions. The filter outputs are then sub-sampled by a factor of two, generating the high-pass bands *H* (horizontal detail coefficients), *V* (vertical detail coefficients), *D* (diagonal detail coefficients) and a low-pass subband *A* (approximation coefficients). The process is then repeated to the *A* band to generate the next level of the decomposition. The following equations describe mathematically the above process for the illumination channel *Y*. It is obvious that the same process applies also to *Cr* and *Cb* chromaticity channels:

$$Y_A^{-(j+1)}(m,n) = \{h_\phi(-m) * \{Y_A^{-j}(m,n) * h_\phi(-n)\} \downarrow^{2n}\} \downarrow^{2m} \quad (1)$$

$$Y_H^{-(j+1)}(m,n) = \{h_\psi(-m) * \{Y_H^{-j}(m,n) * h_\phi(-n)\} \downarrow^{2n}\} \downarrow^{2m} \quad (2)$$

$$Y_V^{-(j+1)}(m,n) = \{h_\phi(-m) * \{Y_V^{-j}(m,n) * h_\psi(-n)\} \downarrow^{2n}\} \downarrow^{2m} \quad (3)$$

$$Y_D^{-(j+1)}(m,n) = \{h_\psi(-m) * \{Y_D^{-j}(m,n) * h_\psi(-n)\} \downarrow^{2n}\} \downarrow^{2m} \quad (4)$$

where $*$ denotes convolution, $Y_A^{-j}(m,n)$ is the approximation of *Y* channel at *j*-th level (note that $Y_A^{-0}(m,n) = Y$), and $\downarrow^{2m}$ and $\downarrow^{2n}$ denote down-sampling by a factor of two along rows and columns respectively.

Following the decomposition of each color channel at specific depth we use center-surround differences to enhance regions that locally stand-out from the surround. Center-surround operations resemble the preferred stimuli of cells found in some parts of the visual pathway (lateral geniculate nucleus-LGN) [28]. Center-surround differences are computed in a particular scale (level *j*) as the point-by-point substraction of the interpolated approximation at the next coarser scale (level *j+1*) from the approximation at this scale (level *j*). The following equations describe the center-surround operations, at

level *j*, for the various bottom-up feature maps:

$$I^{-j} = |Y_A^{-j}(m,n) - ((Y_A^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n)| \tag{5}$$

$$C^{-j} = w_{C_r} C_r^{-j} + w_{C_b} C_b^{-j} \tag{6}$$

$$C_r^{-j} = |C_{r_A}^{-j}(m,n) - ((C_{r_A}^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n)| \tag{7}$$

$$C_b^{-j} = |C_{b_A}^{-j}(m,n) - ((C_{b_A}^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n)| \tag{8}$$

$$O^{-j} = w_{Y_D}|Y_D^{-j} - \hat{Y}_D^{-j}| + w_{Y_V}|Y_V^{-j} - \hat{Y}_V^{-j}| + w_{Y_H}|Y_H^{-j} - \hat{Y}_H^{-j}| \tag{9}$$

$$\hat{Y}_D^{-j} = ((Y_D^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n)| \tag{10}$$

$$\hat{Y}_V^{-j} = ((Y_V^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n)| \tag{11}$$

$$\hat{Y}_H^{-j} = ((Y_H^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n)| \tag{12}$$

In the above equations by I^{-j}, O^{-j}, C^{-j}, we denote the intensity, orientation and colour feature maps computed at scale (level) *j*; $C_{r_A}^{-j}$, $C_{b_A}^{-j}$, are the approximations of chromaticity channels *Cr* and *Cb* at *j*-th scale; $\uparrow^{2m}$ and $\uparrow^{2n}$ denote up-sampling along rows and columns respectively, while $\hat{Y}_A^{-j}$, $\hat{Y}_D^{-j}$, $\hat{Y}_H^{-j}$, $\hat{Y}_V^{-j}$ are the upsampled approximations of $Y_A^{-(j+1)}$, $Y_D^{-(j+1)}$, $Y_V^{-(j+1)}$, and $Y_H^{-(j+1)}$.
The weights $\{w_{C_r}, w_{C_b}, w_{Y_A}, w_{Y_D}, w_{Y_V}, w_{Y_D}\}$ correspond to the within feature modulating gains obtained from the top-down subsystem as illustrated in Figure 1. In the absence of any top-down information, modulating gains can be set to the same value. However, they must be non-negative and their sum must equals one.
The conspicuity maps for intensity (C_I), orientation (C_O) and color (C_C) are computed by combining the features maps at various scales in order to identify both small and large pop-out regions. Combination of the feature maps at different scales is achieved by interpolation to the finer scale, point-by-point addition and application of a saturation (e.g., sigmoid) function to the final result. The following equations describe mathematically the creation of the intensity conspicuity map. The same process applies also to orientation and color conspicuity maps:

$$C_I = \frac{2}{1 + e^{-(\sum_{j=-J_{max}}^{-1} C_I^j)}} - 1 \tag{13}$$

$$C_I^{-j} = I^{-j}(m,n) - ((C_I^{-(j+1)}(m,n) \uparrow^{2m}) * h_\phi(m)) \uparrow^{2n} * h_\phi(n) \tag{14}$$

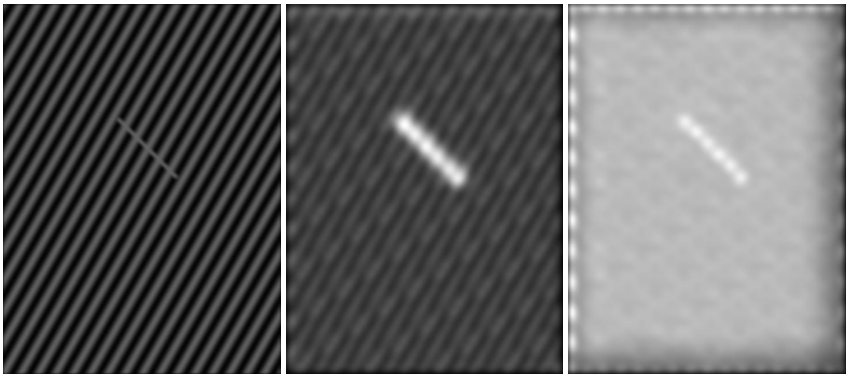
$$C_I^{-J_{max}} = I^{-J_{max}} \tag{15}$$

The maximum analysis depth J_{max} is computed as follows:

$$J_{max} = \lfloor \frac{log_2 N}{2} \rfloor, \;\; N = min(R, C) \tag{16}$$

where in $y = \lfloor x \rfloor$, y is the highest integer value for which $x \geq y$, and R, C are the number of rows and columns of input image respectively.

In Figure 2(a) an image with an object differing from the surround due to orientation is shown. As expected, the orientation conspicuity map, illustrated in Figure 2(b), captures this difference accurately. In contrast, the intensity map, shown in Figure 2(c) is rather noisy because there are no areas that clearly stand-out from their surround due to intensity.



(a) Original image (b) Orientation map (c) Intensity map

Fig. 2. The importance of the orientation channel. Compare the results shown in (b) and (c)

2.4 Combination of Conspicuity Maps

The overall saliency map is computed once the conspicuity maps per channel have been computed. Apart from the bottom-up conspicuity maps the top-down map (in the particular case the skin map) is also used in this stage. As in the previous case a saturation function is used to combine, the modulated by the across feature gains, conspicuity maps into a single saliency map. Normalization and summation, which is the simplest way of combining the conspicuity maps (as followed by Itti [13]), may create inaccurate results in cases where regions that stand-out from their surround in a single modality exist. For example in the case of Figure 2(a) it is expected that only the orientation channel would produce a salient region. Averaging the results of orientation, intensity and color maps (not to mention skin map) will weaken the importance of the orientation map in the total (saliency) map. Therefore, the saturate function is applied so as to preserve the independency and added value of the particular conspicuity maps as shown in eq. 17:

$$C_S = \frac{2}{1 + e^{-(C_I + C_O + C_C + C_F)}} - 1 \tag{17}$$

where C_I, C_O, C_C and C_F are intensity, orientation, colour and skin conspicuity maps respectively, while S is the combined saliency map.

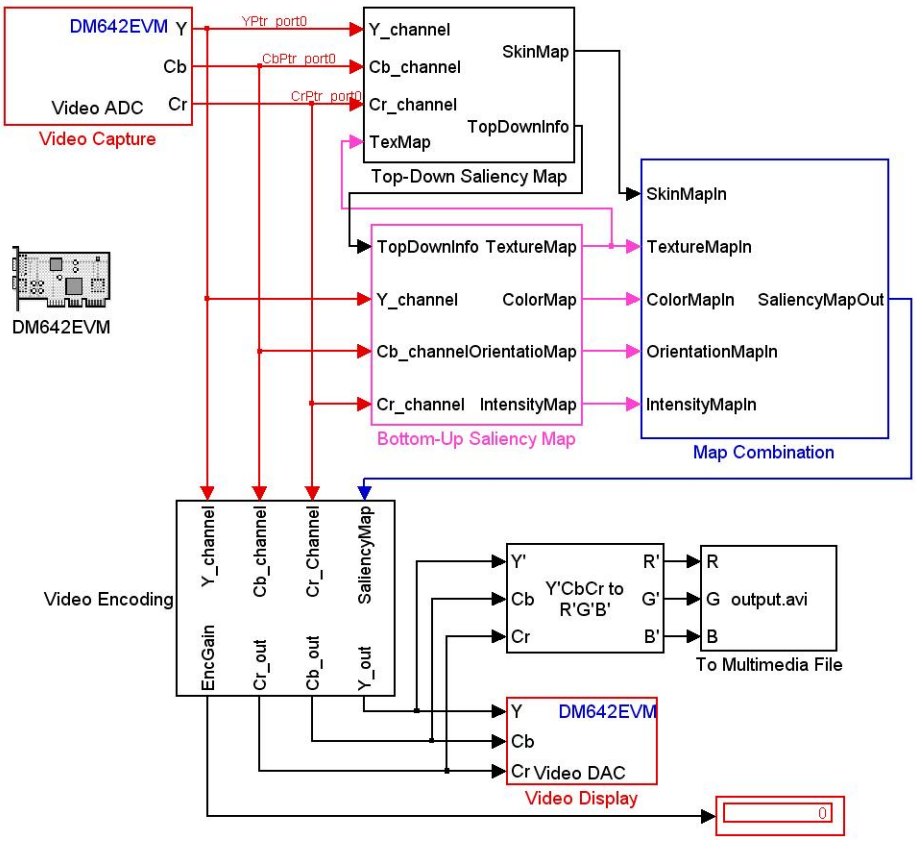


Fig. 3. The saliency map estimator as an embedded system

3 The Saliency map estimator as an embedded system

The saliency map estimator described in the previous section has been implemented as an embedded system with the use of the TMDSDMK642-0E DSP platform [26]. The main architectural components are shown in Figure 3. The corresponding model was developed using the SIMULINK model libraries [19] and the Embedded Target for TI C6000^{TM} blockset [18]. The application code was optimized for speed and deployed to the TMDSDMK642-0E DSP platform using the Texas Instruments' Code Composer Studio^{TM} [10]. A simplified version of the SIMULINK model can be found at [22]. In this emulated version all hardware requirements have been removed to allow for anyone who wishes to test it.
In Figure 3 the blocks named 'TopDown Saliency Map' and 'Bottom-up Saliency Map' indicate the embedded subsystems for the computation of the skin map and bottom-up conspicuity maps respectively. In the 'Map Combination' block the across feature combination of Maps is implemented. Finally, in the 'Video Encoding' block the original video frames are encoded as ROI-based MPEG video streams using the process described in the next section.

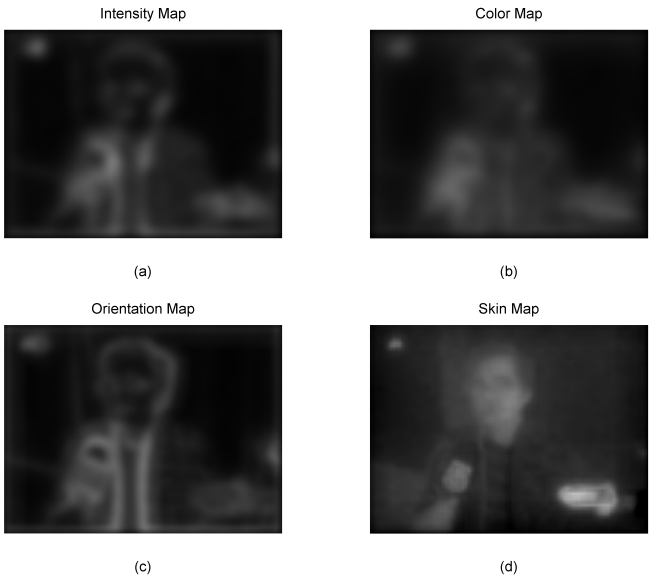


Fig. 4. Conspicuity Maps of an example image

An example of the application of the proposed visual attention model in a still image is shown in Figures 4 and 5. In Figure 4 (a)-(d) the intensity, orientation, color, and skin conspicuity maps are shown respectively. It can be seen that regions that stand out from

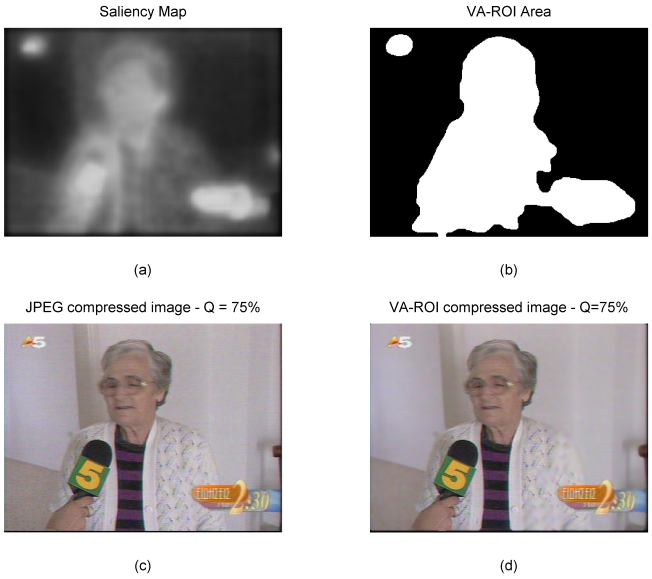


Fig. 5. An example of the visually salient areas identified using the proposed algorithm

their surround in each one of the feature channels (intensity, orientation, color, skin) differ significantly and a linear combination of the corresponding conspicuity maps would result in a noisy saliency map. The use of the saturation function for map combination smooths this effect and maintains the individual value of each feature channel as shown in Figure 5(a). In this Figure the combined saliency map of all feature maps is depicted. In Figure 5(b) the actual ROI area created by thresholding (using Otsu's method [23]) the saliency map is illustrated. Finally, in Figures 5(c) and 5(d) the streamline and ROI-based JPEG encoded images are shown. Non-ROI areas in Figure 5(d) are smoothed, by using a low pass filter, before passed to the JPEG encoder. The compression ratio achieved in this particular case, compared to standard (streamline) JPEG, is about 1.2:1.

4 ROI-based Image and Video Encoding using Saliency Maps

The visually attended areas indicated by a saliency map and computed through the proposed model, can be considered as ROIs in video entertainment movies as well as in a variety of applications such as teleconferencing and video surveillance. In the first case bottom-up channels are mainly engaged to model sub-conscious visual attention attraction while in visual-telephony applications the existence of human faces in every video frame is almost guaranteed, and therefore, it is anticipated that the first area to receive the human attention is the face area. However, we cannot identify as ROIs only the face like areas [17], [32] because there is always the possibility, even in a video-telephony

setting, that other objects in the scene attract the human interest in a sub-conscious manner.

For ROI-based video encoding we consider as ROI an area created by thresholding the saliency map. The latter is obtained by applying the proposed model to every video frame. Once ROI areas are identified the non-ROI areas in the video frames or images are blurred via a smoothing filter. It is well-known that in smooth areas a higher compression ratio than textural ones can be achieved due to the spatial decorrelation obtained by applying either the DCT transform (JPEG, MPEG-1) or wavelet decomposition (JPEG 2000, MPEG-4). The assumption made for smoothing non-ROI areas is that in the limited time in which a frame is presented to an observer the latter will concentrate on visually salient areas and will not perceive deterioration in non-visually important areas. Smoothing non-ROI areas is not optimal in terms of expected encoding gain. However, it has the advantage of producing compressed streams that are compatible with existing decoders.

The quality of the VA-ROI based encoded videos and images is evaluated through a set of visual trial tests. These tests were conducted based on ten short video clips, namely: *eye_witness, fashion, grandma, justice, lecturer, news_cast1, news_cast2, night_interview, old_man, soldier* (see [21]). All video clips were chosen to have a reasonably varied content, whilst still containing humans and other objects that could be considered to be more important (visually interesting) than the background. They contain both indoor and outdoor scenes and can be considered as typical cases of news reports based on 3G video telephony. However, it should be noted that the selected video clips were chosen solely to judge the efficacy of VA-ROI coding in MPEG-1 and MPEG-4 and are not actual video-telephony clips. In the MPEG-1 case variable bit rate (VBR) encoding was performed with a frame resolution of 288x352 pixels, frame rate of 25 fps, and GOP structure: IBBPBBPBBPBB. In the MPEG-4 case, VBR encoding was also adopted but the basic aim was the low bit rate. Therefore, a frame resolution of 144x176 pixels and a frame rate of 15 fps was chosen so as to conform to the constraints imposed by 3G video telephony. The *ImTOO MPEG Encoder* plugin [9] was applied to uncompressed *avi* files, generated by using the *avifile* function of Matlab$^{(R)}$, to create three MPEG-1 and three MPEG-4 video-clips for each case. The first one corresponds to the proposed VA based encoding (named VA-ROI), the second corresponds to VA based coding proposed by Itti [11] (named IttiROI), and the third corresponds to standard MPEG (MPEG-1 and MPEG-4) video coding. In both VA methods (the proposed and Itti's) non-ROI areas in each frame are smoothed before communicated to the encoder.

5 Visual trial tests and experimental results

The purpose of the visual trial test was to directly compare the subjective visual quality of VA-ROI based, IttiROI based, and streamline MPEG-1 and MPEG-4 video encoding. ROI s were determined using the proposed embedded saliency map estimator for the VA-ROI method and the Neuromorphic Vision Toolkit [8] for Itti's method. In both cases saliency maps were thresholded using Otsu's method [23] to create the binary masks that correspond to ROI areas.

A three alternative forced choice (3AFC) methodology was selected because of its sim-

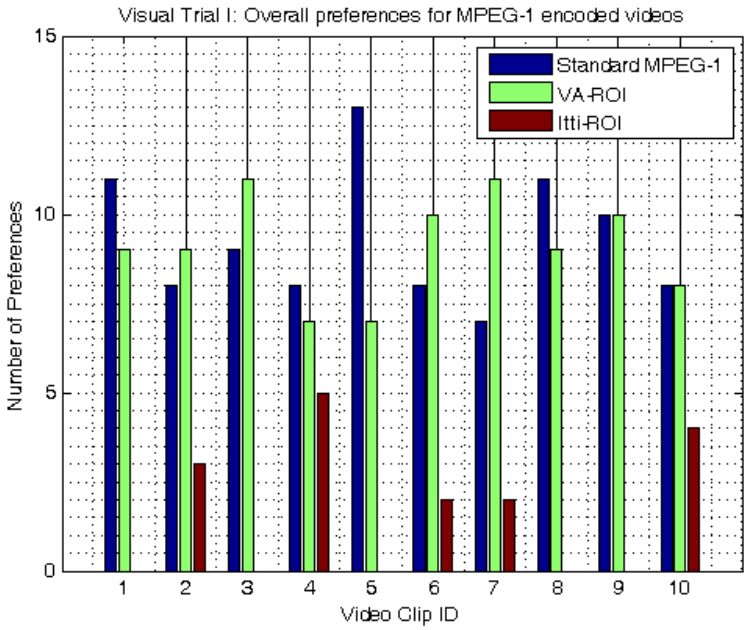


Fig. 6. VA ROI-based encoding (green), Itti-ROI (red) and standard MPEG-1 encoding (blue) preferences on the eye_witness (1), fashion (2), grandma (3), justice(4), lecturer (5), news_cast1(6), news_cast2 (7), night_interview (8), old_man (9) and soldier (10) video clips.

Table 1. Overall preferences (independent of video clip) - MPEG-1 case

Encoding method	**Preferences**	**Average Bit Rate** (Kbps)
VA-ROI	91	1125
Itti-ROI	16	1081
Standard MPEG-1	93	1527

plicity, i.e., the observer watches the three differently encoded video clips and then selects the one preferred, and so there are no issues with scaling opinion scores between different observers [3]. There were ten observers, (five male and five female) with good, or corrected, vision, all being non-experts in image compression (university students). The viewing distance was approximately 20 cm (i.e., a normal PDA / mobile phone viewing distance) for the MPEG-4 videos and 50 cm (typical PC screen viewing distance) for the MPEG-1 video. The video clip triples were viewed one at a time in a random order. The observer was free to view the video clip triples multiple times before making a decision within a time framework of 60 seconds. Each video clip triple was viewed twice, giving (10x10x2) 200 comparisons. Video-clips were viewed on a Smartphone (Nokia^{TM} N90) display in the case of the MPEG-4 videos and on a typical PC monitor in a darkened room (i.e., daylight with drawn curtains) in the case of the MPEG-1 videos. Prior to the start of the visual trial all observers were given a short period of training on the experiment and they were told to select the video clip they preferred. Both the MPEG-1 and the MPEG-4 encoded videos were tested through a visual trial.

Table 1 shows the overall preferences, i.e., independent of video clip, for the standard MPEG-1, the Itti-ROI and the proposed VA-ROI-based method. It can be seen that there is slight preference to standard MPEG-1 which is selected at 46.5% of the time as being of better quality. The difference in selections, between VA-ROI based (selected at 45.5% of the time as being of better quality) and standard MPEG-1 encoding, is actually too small to indicate that the VA-ROI based encoding deteriorates significantly the quality of the produced video stream. At the same time the bit rate gain, which is about 36% on average (see also Table 3), shows clearly the efficiency of VA-ROI based encoding. IttiROI encoded videos were selected as few as 8% of the time as being of better quality. This fact indicates a clear visual deterioration. The slightly increased encoding gain (41% - see also Table 3), compared to the VA-ROI method, does not trade off this lowering in perceived visual quality.

In Figure 6, the selections made per video clip are shown. In one of them (*justice*) there is a clear preference to standard MPEG-1, while in *news_cast2* there is a clear preference to VA-ROI. The latter is somehow strange because the encoded quality of individual frames in VA-ROI based encoding is, at best, the same as standard MPEG-1 (in the ROI areas). Therefore, preference to VA-ROI based encoding may be assigned to denoising, performed on non-ROI areas by the smoothing filter. In the remaining eight video clips the difference in preferences between VA-ROI and standard MPEG-1 may be assigned to statistical error. On the other hand, it is important to note that Itti's ROI-based encoding method is in all cases least preferable than the proposed VA-ROI method. This may be assigned to the fact that Itti's saliency map estimation is optimized for identifying rather large objects that stand out from their surround. In this way small areas, such as channel logos, are not recognized as ROI's though they attract human attention. In order to be fair we should mention, however, that in typical 3G video telephony circumstances the existence of TV channel logos in a scene is rather unusual. On the other hand, existence of other small but visually salient objects is possible.

Table 2 shows the overall preferences of the visual trial test for the MPEG-4 case. Standard MPEG-4 and VA-ROI were both selected at 39.5% of the time as being of better

Table 2. Overall preferences (independent of video clip) - MPEG-4 case

Encoding method	**Preferences**	**Average Bit Rate** (Kbps)
VA-ROI	79	197.5
Itti-ROI	42	194.0
Standard MPEG-4	79	224.6

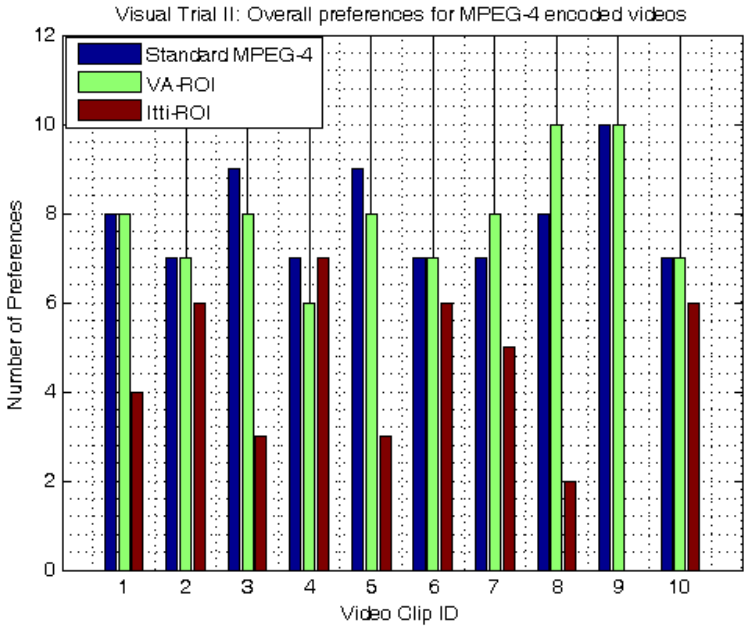


Fig. 7. VA ROI-based encoding (green), Itti-ROI (red) and standard MPEG-4 encoding (blue) preferences on the eye_witness (1), fashion (2), grandma (3), justice(4), lecturer (5), news_cast1(6), news_cast2 (7), night_interview (8), old_man (9) and soldier (10) video clips.

quality. As in the MPEG-1 case there is no indication that VA-ROI based encoding deteriorates the visual quality of the MPEG-4 video stream. On the other hand, the bit rate gain achieved by the VA-ROI method is about 14% on average (see also Table 4), which is rather high if we take into account that encoding gain is difficult to obtain for very low bit rates. MPEG-4 encoded videos with ROIs identified based on Itti's visual attention scheme were selected 21% of the time indicating a clear deterioration in subjective visual quality compared to both standard MPEG-4 and VA-ROI. It should be noted, however, that IttiROI MPEG-4 videos were selected significantly more times than their IttiROI MPEG-1 counterparts in the MPEG-1 visual trial test. This fact, as well as the lower encoding gain achieved in the MPEG-4 case for both VA-ROI and IttiROI videos (13.7% and 15.7% respectively) is due to the reduced frame resolution selected for the MPEG-4 frames (176x144 pixels compared to 288x352 pixels of the MPEG-1 frames). Downsampling video frames to 176x144 pixels leads to an overall frame smoothing which results in lower difference in quality between ROI and non-ROI areas in both VA-ROI and IttiROI MPEG-4 videos. As a consequence a lower encoding gain is achieved (remember that, practically, in standard MPEG-4 encoding the whole frame is considered as ROI). The main conclusion of the previous discussion is that smoothing of non-ROI areas is not the appropriate way to apply ROI-based encoding. Modifications of the corresponding encoders and their encoding parameters such as the quality factor for ROI and non-ROI macroblocks may be necessary in order to take advantage of the existence of ROI areas. Another option is to encode the disconnected ROI areas as video objects keeping their shapes with the aid of the Shape Adaptive DCT transform [25].
In Figure 7, the selections made per video clip, for the MPEG-4 visual trial, are shown. In several cases (*fashion*, *justice*, *news_cast1* and *soldier*) the perceived visual quality for the three encoding modes is the same. In other cases, such as the *old_man* video clip, the quality of the IttiROI video is much lower than the corresponding standard MPEG-4 and VA-ROI videos. In addition to missing some small visually salient objects such as TV channel logos, Itti's method fails in some cases to identify foreground human faces as ROI areas. This is, for example, the case for the *old_man* video clip. One the other hand in all video clips there is no distinguishable difference in visual quality between standard MPEG-4 and VA-ROI. Therefore, the lower encoding gain achieved by VA-ROI compared to IttiROI is exchanged by perceived visual quality.

5.1 Bit rate gain

Table 3 presents the bit-rates achieved by VA-ROI based, IttiROI based and standard MPEG-1 encoding in each one of ten video clips. It is clear that the bit rate gain obtained by the VA-ROI method is significant, ranging from 15% (*grandma* video clip) to 64% (*soldier* video clip). Furthermore, it can be seen from the results obtained in the *soldier* and *news_cast2* video sequences, that increased bit-rate gain does not necessarily mean worse quality of the VA-ROI encoded video. The encoding gain in the IttiROI encoded videos is, in general, similar to that of VA-ROI. In the four video sequences (*grandma*, *lecturer*, *soldier* and *news_cast2*) where IttiROI clearly outperforms VA-ROI in terms of encoding gain this gain is non-gracefully exchanged with visual quality as it can be seen in Figure 6 (0, 0, 3, and 0 selections respectively in the visual trial test). In

Table 3. Comparison of VA-ROI based, IttiROI based and Standard MPEG-1 encoding in ten video sequences

Video Clip	**Encoding method**	**Bit Rate** (Kbps)	**Bit Rate Gain**
Eye_witness	VA-ROI	1610	30.5%
	Itti-ROI	1585	32.6%
	Standard MPEG-1	2101	
fashion	VA-ROI	1200	31.1%
	Itti-ROI	1188	32.4%
	Standard MPEG-1	1573	
grandma	VA-ROI	1507	15.2%
	Itti-ROI	1300	33.6%
	Standard MPEG-1	1737	
justice	VA-ROI	1468	30.5%
	Itti-ROI	1606	19.3%
	Standard MPEG-1	1916	
lecturer	VA-ROI	950	57.3%
	Itti-ROI	848	76.3%
	Standard MPEG-1	1495	
news_cast1	VA-ROI	991	39.7%
	Itti-ROI	999	38.5%
	Standard MPEG-1	1384	
news_cast2	VA-ROI	930	41.9%
	Itti-ROI	836	57.8%
	Standard MPEG-1	1319	
night_interview	VA-ROI	790	50.8%
	Itti-ROI	750	59.0%
	Standard MPEG-1	1192	
old_man	VA-ROI	1307	32.9%
	Itti-ROI	1085	60.0%
	Standard MPEG-1	1737	
soldier	VA-ROI	500	63.9%
	Itti-ROI	614	33.3%
	Standard MPEG-1	819	
Average	VA-ROI	1125	35.7%
	Itti-ROI	1081	41.2%
	Standard MPEG-1	1527	

Table 4. Comparison of VA-ROI based, IttiROI based and Standard MPEG-4 encoding in ten video sequences

Video Clip	**Encoding method**	**Bit Rate** (Kbps)	**Bit Rate Gain**
Eye_witness	VA-ROI	392	12.2%
	Itti-ROI	381	15.3%
	Standard MPEG-4	439	
fashion	VA-ROI	288	10.4%
	Itti-ROI	285	11.7%
	Standard MPEG-4	318	
grandma	VA-ROI	264	11.9%
	Itti-ROI	247	19.5%
	Standard MPEG-4	296	
justice	VA-ROI	227	11.5%
	Itti-ROI	236	6.9%
	Standard MPEG-4	253	
lecturer	VA-ROI	107	28.3%
	Itti-ROI	110	25.3%
	Standard MPEG-4	138	
news_cast1	VA-ROI	164	13.5%
	Itti-ROI	170	9.6%
	Standard MPEG-4	186	
news_cast2	VA-ROI	118	14.9%
	Itti-ROI	115	17.8%
	Standard MPEG-4	136	
night_interview	VA-ROI	127	15.7%
	Itti-ROI	128	14.8%
	Standard MPEG-4	147	
old_man	VA-ROI	217	13.3%
	Itti-ROI	189	30.3%
	Standard MPEG-4	246	
soldier	VA-ROI	71	22.5%
	Itti-ROI	79	10.4%
	Standard MPEG-4	87	
Average	VA-ROI	197.5	13.7%
	Itti-ROI	194.0	15.7%
	Standard MPEG-4	224.6	

contrary, in the two cases (*soldier*, *justice*) where VA-ROI clearly outperforms IttiROI in terms of encoding gain, VA-ROI encoded videos also outperform IttiROI videos in terms of visual quality.
In Table 4 the bit-rates achieved by VA-ROI based, IttiROI based and standard MPEG-4 encoding are also presented. The bit rate gain obtained by the VA-ROI method ranges from 10.4% (*fashion* video clip) to 28.3% (*lecturer* video clip) and is important if we take into account that improvement in video compression at low bit-rates is more difficult than improvement in intermediate (MPEG-1) and high (MPEG-2) bit rates. The encoding gain in the IttiROI encoded videos presents higher variance across the various video clips since it ranges from 6.9% (*fashion*) to 30.3% (*old_man*). In general, however, similar encoding gains are obtained by VA-ROI and IttiROI. In the two video sequences (*grandma* and *old_man*) where IttiROI clearly outperforms VA-ROI in terms of encoding gain the visual quality of the VA-ROI encoded videos is significantly higher (see also Figure 7). In contrary, in the *soldier* video clip where VA-ROI clearly outperforms IttiROI in terms of encoding gain, it has similar visual quality with the IttiROI encoded video.

6 Conclusions and future work

In this paper we have presented a detailed implementation of a visual attention model that can be used to identify regions of interest for ROI-based video coding. The proposed model operates along two separate information channels; a high-level one which models conscious search for human faces and a low-level one which models sub-conscious attention attraction. Processing within the low-level information channel is basically implemented in a bottom-up manner with the aid of wavelet decomposition for multiscale analysis. The high-level information channel models in a probabilistic manner the skin color and in combination with a bottom-up created texture map generates a skin conspicuity map. The latter corresponds to the likelihood of image pixels to belong to a human face. Multiresolution analysis was also adopted in the creation of skin conspicuity map in order to allow both small and large faces to be detected. The skin conspicuity map and the bottom-up conspicuity maps corresponding to intensity, orientation and color are combined with the aid of a saturation function to create the final saliency map. Through the saturation function strong stimuli in a particular conspicuity map (intensity, color, orientation or skin) are not smoothed (as it could happen in a typical linear combination) and they are propagated to the final saliency map. A prototype of the porposed model has been implemented as an embedded system with the aid of the TMDSDMK642-0E DSP platform while a fully software version is available online [22].
In order to apply the proposed model for ROI-based video encoding, ROI areas are generated by thesholding the computed saliency maps using an image-adaptive threshold. ROI based encoding is then applied by smoothing the non-ROI areas with a low-pass filter. Coding efficiency was examined based on both visual trial tests and encoding gain. The results presented indicate that: (a) Significant bit-rate gain, compared to streamline MPEG-1 and MPEG-4, can be achieved using the VA-ROI based video encoding, (b) the areas identified as visually important by the proposed VA algorithm are in confor-

mance with the ones identified by the human subjects, as it can be deducted by the visual trial tests, and (c) VA-ROI outperforms the corresponding method proposed by Itti [11] in terms of visual quality but achieves slightly lower encoding gains.
Further work includes conducting experiments in an object basis framework where as objects will be considered the disjoint ROI areas. Furthermore, the effect of incorporating priority encoding by varying the quality factor of the DCT quantization table across VA-ROI and non-ROI frame blocks will be examined.

Acknowledgement: The study presented in this paper was supported (in part) by the research project "OPTOPOIHSH: Development of knowledge-based Visual Attention models for Perceptual Video Coding", PLHRO 1104/01 (`http://www.optopiisi.signalgenerix.com`) funded by the Cyprus Research Promotion Foundation (`http://www.research.org.cy/`)

References

1. K.R. Cave. The feature gate model of visual selection. *Psychol. Res.*, 62(2).
2. R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.*, 18:193–222, 1995.
3. M. Eckert and A. Bradley. Perceptual models applied to still image compression. *Signal Processing*, 70(3):177–200, 1998.
4. S. Frintrop. Vocus: A visual attention system for object detection and goal-directed search. 2006.
5. R. S. Gaborski, V. S. Vaingankar, and R. L. Canosa. Goal directed visual search based on color cues: Co-operative effects of top-down & bottom-up visual attention. In *Proceedings of the Artificial Neural Networks in Engineering - ANNIE2003*, November 2003.
6. R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall Inc, 2^{nd} edition, 2002.
7. F.H. Hamker. A dynamic model of how feature cues guide spatial attention. *Vision Research*, 44:501–521, 2004.
8. iLab. Neurmomorphic Vision C++ Toolkit (iNVT). Univ. Of Southern California, online at: http://ilab.usc.edu/toolkit/, last visited: January 2007.
9. ImTOO^{TM}. ImTOO MPEG Encoder. online: http://www.imtoo.com/mpeg-encoder.html, last visited: January 2007.
10. Texas Instruments. Code Composer Studio User's Guide. Texas Instruments Literature Number SPRU328B, 2000.
11. L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. on Image Processing*, 13:1304–1318, 2004.
12. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.
13. 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*, 20(11):1254–1259, 1998.
14. W. James. *The Principles of Psychology*. Cambridge, MA: Harvard University Press, 1890/1981.
15. C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.
16. S. Mallat. A theory of multiresolution signal decomposition: The wavelet model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

17. E. Mandel and P. Penev. Facial feature tracking and pose estimation in video sequences by factorial coding of the low-dimensional entropy manifolds due to the partial symmetries of faces. In *Proceedings of the 25$^t h$ IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume IV, pages 2345–2348, June 2000.
18. The MathWorks. Embedded target for TI C6000^{TM} DSP 3.1. Online at: http://www.mathworks.com/products/tic6000/, last visited: January 2007.
19. The MathWorks. Simulink$^{(R)}$ - simulation and model-based design. Online at: http://www.mathworks.com/products/simulink/, last visited: January 2007.
20. V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 453–461, London, UK, 2002. Springer-Verlag.
21. [Online]. Original video frames. http://www.cs.ucy.ac.cy/∼nicolast/research/frames.rar, last visited: January 2007.
22. [Online]. Visual Attention Model Code. http://www.cs.ucy.ac.cy/∼nicolast/research/VAmodel.zip, last visited: January 2007.
23. N. Otsu. A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9:62–66, 1979.
24. C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.
25. T. Sikora and B. Makai. Shape-Adaptive DCT for Generic Coding of Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1):59–62, 1995.
26. DSP Development Systems. Tms320dm642 evaluation module technical reference. Technical report, Spectrum Digital Inc., 2003.
27. A. Torralba. Modeling global scene factors in attention. *J. Opt. Soc. Am.-A*, 20(7):1407–1418, 2003.
28. A. M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.
29. N. Tsapatsoulis, Y. Avrithis, and S. Kollias. Facial image indexing in multimedia databases. *Pattern Analysis and Applications*, 4(2).
30. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.
31. B. A. Wandell. *Foundations of Vision*. Sinauer Associates, Sunderland, MA 01375, 1995.
32. Z. Wang, L. Lu, and A. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Transactions on Image Processing*, 12(2):243–254, 2003.
33. J.M. Wolfe. Visual search in continuous, naturalistic stimuli. *Vision Research*, 34(9):1187–95, 1994.