Priority Coding for Video-telephony Applications based on Visual Attention

Nicolas Tsapatsoulis
Department of Computer Science
University of Cyprus
Cyprus
nicolast@ucy.ac.cy

Konstantinos Rapantzikos
School of Electrical and Computer Engineering
National Technical University of Athens
Greece
rap@image.ntua.gr

Yannis Avrithis
School of Electrical and Computer Engineering
National Technical University of Athens
Greece
iavr@image.ntua.gr

***Abstract*— In this paper we investigate the utilization of visual saliency maps for ROI-based video coding of video-telephony applications. Visually salient areas indicated in the saliency map are considered as ROIs. These areas are automatically detected using an algorithm for visual attention (VA) which builds on the bottom-up approach proposed by Itti *et al*. A top-down channel emulating the visual search for human faces performed by humans has been added, while orientation, intensity and color conspicuity maps are computed within a unified multi-resolution framework based on wavelet subband analysis. Priority encoding, for experimentation purposes, is utilized in a simple manner: Frame areas outside the priority regions are blurred using a smoothing filter and then passed to the video encoder. This leads to better compression of both Intra-coded (I) frames (more DCT coefficients are zeroed in the DCT-quantization step) and Inter coded (P,B) frames (lower prediction error). In more sophisticated approaches, priority encoding could be incorporated by varying the quality factor of the DCT quantization table. Extended experiments concerning both static images as well as low-quality video show the compression efficiency of the proposed method. The comparisons are made against standard JPEG and MPEG-1 encoding respectively.**

Keywords - visual attention, perceptional video coding, saliency map, video telephony

I. 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 region of interest (ROI) coding [1]. 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 2–5 of visual angle (the fovea) around the center of gaze is captured at high resolution, with logarithmic resolution falloff with eccentricity [2]. 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 [1]. 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 visually attended areas. For this purpose we make use of the Feature Integration Theory (FIT) of Treisman *et al* [3] that was derived from visual search experiments. According to this theory, visual features are registered early, automatically and in parallel along a number of separable dimensions (e.g. intensity, color, orientation, size, shape etc). The FIT theory was the basis of several visual attention algorithms and computational models that have been developed over the last two decades [4]-[8]. Among them the computational model of Itti and Koch [9], [10] is probably the most representative and well documented. It deals mainly with static color images and combines several topographic maps computed independently along different feature channels into a final saliency map. 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. 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.

In this study we present an alternative, computationally efficient way of saliency map estimation utilizing wavelets and multiresolution theory. In addition a top-down channel, emulating the visual search for human faces performed by humans has also been added. In visual-telephony applications 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 modeling sub-conscious visual attention attraction.

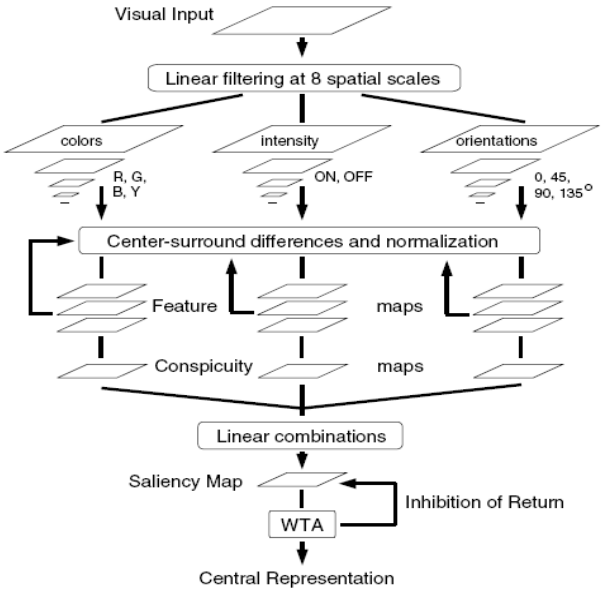


Figure 1. The Visual Attention model of Itti *et al*. Copied from [9].

The organization of the paper is as follows: In Section II we describe the Visual Attention method that is used to identify the visually salient regions on a per frame basis. Experimental results and the visual trial tests that were used to justify that the VA ROI-based encoding is of similar quality compared to the standard MPEG-1 and outperforms the method proposed by Itti [20] are presented in Section III. Finally, further work is proposed and conclusions are drawn in Section IV.

II. SALIENCY-BASED ESTIMATION OF ROIS

A. A VA model enhanced with top-down information

Bottom-up approaches to visual attention lack provision of conscious search. In the past it had been thought that bottom-up signals normally achieved attention capture; it is considered that top-down control is usually in charge [11]. Towards this direction we integrate prior knowledge to the saliency-based model in order to draw the attention to regions with specific characteristics (Figure 2-left 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. It is reasonable to consider that in typical video-telephony settings humans almost always focus on human objects. Thus, a model for deriving another conspicuity map based on the color similarity of objects with human-skin is reasonable. We use a multiresolution skin detector scheme presented in [12] to generate a skin map with possible face locations. These locations are highlighted by multiplying the skin map with a texture map created through range filtering [13] of the intensity channel (see Figure 3). It should be noted, however, that we cannot identify as ROIs only the face like areas [14], [15] because there is always the possibility, even in a video-telephony setting, that other objects in the scene attract the human interest in a subconscious manner. The face map is, therefore, combined with the bottom-up conspicuity maps to produce the final saliency map.

B. Wavelet –based estimation of conspicuity maps

In addition, to adding a top-down branch to Itti's model, a multiscale analysis based on wavelets [16] was adopted for computing the center-surround structure in the bottom-up feature maps. Furthermore, the *YCbCr* color space (instead of the *RGB* used by Itti) was selected, first to keep conformance with the face detection scheme [12], 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.

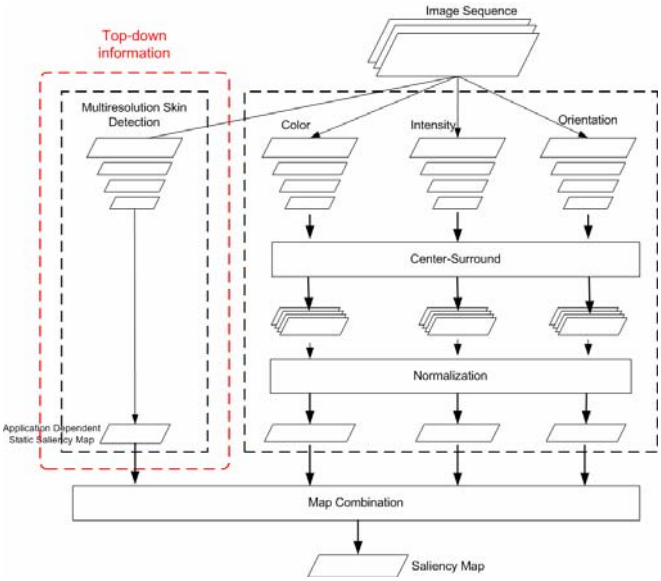


Figure 2. The proposed Visual Attention model. In the left side is the top-down channel addition for considering the influences conscious search (existence of prior knowledge – e.g., skin like objects)

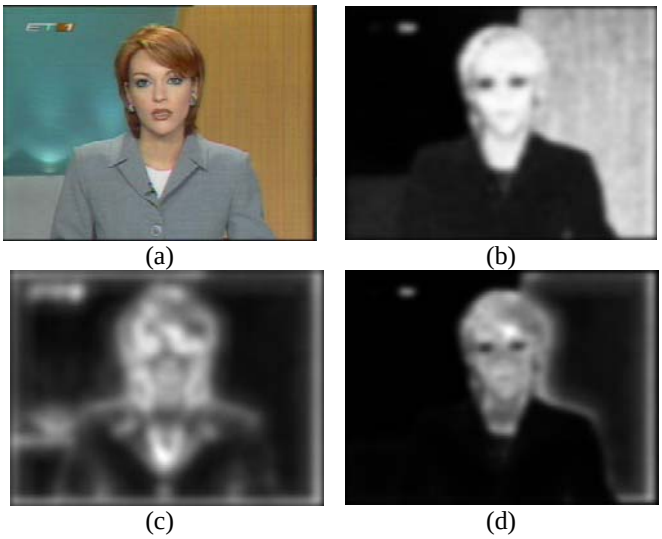


Figure 3. (a) Original video frame, (b) Skin map (see the influence of the skin like color in the background), (c) texture map, (d) multiplication of texture and skin maps. It can be seen that face area is now more prominent compared to that in (b)

In order of 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 colour 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:

$$\begin{aligned} Y_A^{-(j+1)}(m,n) &= \left(h_\phi(-m) * \left(Y_A^{-j}(m,n) * h_\phi(-n)\right)\downarrow^{2n}\right)\downarrow^{2m} \\ Y_H^{-(j+1)}(m,n) &= \left(h_\psi(-m) * \left(Y_A^{-j}(m,n) * h_\phi(-n)\right)\downarrow^{2n}\right)\downarrow^{2m} \\ Y_V^{-(j+1)}(m,n) &= \left(h_\phi(-m) * \left(Y_A^{-j}(m,n) * h_\psi(-n)\right)\downarrow^{2n}\right)\downarrow^{2m} \\ Y_D^{-(j+1)}(m,n) &= \left(h_\psi(-m) * \left(Y_A^{-j}(m,n) * h_\psi(-n)\right)\downarrow^{2n}\right)\downarrow^{2m} \end{aligned} \quad (1)$$

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.

For the wavelet decomposition we use Daubechie's (length four-eight coefficients) wavelet (filter h_ψ) and approximation coefficients (filter h_ϕ).

The center-surround scheme of Itti *et al.* 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 following equations describe the way intensity, color and orientation conspicuity maps are computed at level *j:*

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

$$C^{-j} = C_r^{-j} + C_b^{-j} \quad (3)$$

$$C_b^{-j} = \left| C_{b_A}^{-j} - \left(\left(C_{b_A}^{-(j+1)}(m,n)\uparrow^{2m}\right) * h_\phi(m)\right)\uparrow^{2n} * h_\phi(n) \right| \quad (3.1)$$

$$C_r^{-j} = \left| C_{r_A}^{-j} - \left(\left(C_{r_A}^{-(j+1)}(m,n)\uparrow^{2m}\right) * h_\phi(m)\right)\uparrow^{2n} * h_\phi(n) \right| \quad (3.2)$$

$$O^{-j} = \left| Y_D^{-j} - \hat{Y}_D^{-j} \right| + \left| Y_V^{-j} - \hat{Y}_V^{-j} \right| + \left| Y_H^{-j} - \hat{Y}_H^{-j} \right| \quad (4)$$

$$\hat{Y}_D^{-j} = \left(\left(Y_D^{-(j+1)}(m,n)\uparrow^{2m}\right) * h_\phi(m)\right)\uparrow^{2n} * h_\phi(n) \quad (4.1)$$

$$\hat{Y}_V^{-j} = \left(\left(Y_V^{-(j+1)}(m,n)\uparrow^{2m}\right) * h_\phi(m)\right)\uparrow^{2n} * h_\phi(n) \quad (4.2)$$

$$\hat{Y}_H^{-j} = \left(\left(Y_H^{-(j+1)}(m,n)\uparrow^{2m}\right) * h_\phi(m)\right)\uparrow^{2n} * h_\phi(n) \quad (4.3)$$

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*, $C_{b_A}^{-j}$, $C_{r_A}^{-j}$ 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 $\hat{Y}_A^{-j}$, $\hat{Y}_D^{-j}$, $\hat{Y}_V^{-j}$, $\hat{Y}_H^{-j}$ are the upsampled approximations of $Y_A^{-(j+1)}$, $Y_D^{-(j+1)}$, $Y_V^{-(j+1)}$, and $Y_H^{-(j+1)}$.

The maximum analysis depth *Jmax* is computed as follows:

$$J\max = \left\lfloor \frac{1}{2}\log_2 N \right\rfloor, \; N = \min(R,C) \quad (5)$$

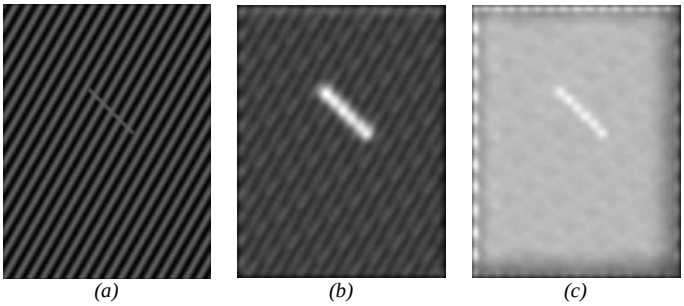
where in $y = \lfloor x \rfloor$ y is the highest integer value for which $x \geq y$, and R, C are the number of rows and columns of input image respectively.

Combination of the features maps at the various scales (conspicuity maps) is needed to cover both small and large stand-out objects. Combination of different scales is achieved by interpolation to the finer scale, point-by-point subtraction and application of a saturate function to the final result. The following equations describe mathematically the process of combining the results of two successive scales for the intensity conspicuity map. The same process applies also to intensity and colour conspicuity maps:

$$C_I^{-j} = I^{-j} + \left(C_I^{-(j+1)}(m,n)\uparrow^{2m} * h_\phi(m)\right)\uparrow^{2n} * h_\phi(n) \quad (6)$$

$$C_I^{-J\max} = I^{-J\max} \quad (6.1)$$

where C_I^{-j} is the intermediate, at scale *j*, intensity conspicuity map.

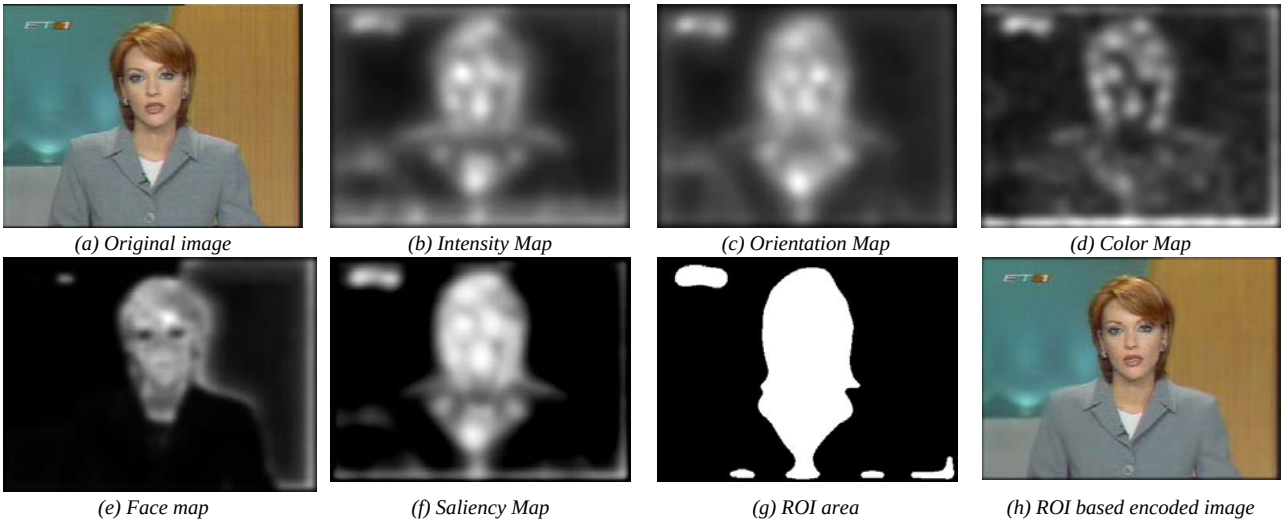


(a) *(b)* *(c)*

Figure 4. (a) Original image, (b) orientation map, (c) intensity map. Results shown in (b) and (c) indicate the importance of the orientation channel.

In Figure 4(a) an image with an object differing from the surround due to orientation is shown. As expected, the orientation conspicuity map, illustrated in Figure 4(b), captures this difference accurately. In contrast, the intensity map, shown in Figure 4(c) is rather noisy because there are no areas that clearly stand-out from their surround due to intensity.

An example of the visually salient areas identified using the proposed algorithm is shown in Figure 5(g). In Figures 5(b)-5(e) the (normalized) intensity, orientation, color, and skin maps are shown respectively. In Figure 5(f) the combined saliency map of all feature maps is depicted, while in Figure 5(g) the actual ROI area created by thresholding (using Otsu's method [17]) the saliency map is presented. Finally, in Figure 5(h) the ROI-based JPEG encoded image is shown. In this Figure non-ROI areas are smoothed before passed to JPEG encoder. The compression ratio achieved in this particular case, *compared to standard JPEG*, is about 1.44:1.



(a) Original image *(b) Intensity Map* *(c) Orientation Map* *(d) Color Map*

(e) Face map *(f) Saliency Map* *(g) ROI area* *(h) ROI based encoded image*

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

C. Saliency Map computation

The overall saliency map is computed once the conspicuity maps per channel have been computed. A saturate (sigmoid) function is used to combine the conspicuity maps into a single saliency map. Normalization and summation, which is the simplest way of combining the conspicuity maps, 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 4(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 face 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 concpicuity maps as shown in eq. (7).

$$S = \frac{2}{1 + e^{\left(-\left(C_I^{-0} + C_O^{-0} + C_C^{-0} + C_F^{-0}\right)\right)}} - 1 \qquad (7)$$

where C_I^{-0}, C_O^{-0}, C_C^{-0} and C_F^{-0} are the final (computed at scale *0*) intensity, orientation, colour and face conspicuity maps respectively, while *S* is the combined saliency map.

III. Visual Trial Tests and Experimental Results

To evaluate the algorithm, we simply use it as a front end; that is, once the VA-ROI areas identified the non-ROI areas in the video frames are blurred. Although not optimal in terms of expected encoding gain, this approach has the advantage of producing compressed streams that are compatible with existing decoders. In order to examine the quality of the VA-ROI based encoded videos a set of visual trial 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 [18]). 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 are not actual video- telephony clips.

For each video clip encoding aiming at low-bit rate (frame resolution of 144x192, frame rate 24 fps, GOP structure: IBBPBBPBBPBB) has been taken place so as to conform to the constraints imposed by 3G video telephony. Three low-resolution video-clips were created for each case, one corresponding to the proposed VA based coding, the second corresponding to VA based coding proposed by Itti [20]- [21], and the third corresponding to standard MPEG-1 video coding. In both VA methods (the proposed and Itti's) non-ROI areas in each frame are smoothed before communicated to the encoder

A. Experimental methodology

The purpose of the visual trial test was to directly compare VA ROI based and standard MPEG-1 encoded video where the ROI is determined using the proposed VA algorithm. A two alternative forced choice (2AFC) methodology was selected

because of its simplicity, i.e., the observer views the video clips and then selects the one preferred, and so there are no issues with scaling opinion scores between different observers [19]. There were ten observers, (5 male and 5 female) all with good, or corrected, vision and all observers were non-experts in image compression (students). The viewing distance was approximately 20 cm (i.e., a normal PDA / mobile phone viewing distance) and the video clip pairs were viewed one at a time in a random order. The observer was free to view the video clips multiple times before making a decision within a time framework of 60 seconds. Each video pair was viewed twice, giving (10x10x2) 200 comparisons. Video-clips were viewed on a typical PDA display in a darkened room (i.e., daylight with drawn curtains).

B. Results

Table I 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 44% of the time as being of better quality, compared to 37.5% of the proposed VA-ROI method. In contrary, Itti's method has been selected significantly fewer times (18.5%). The difference in selections, between VA ROI-based and standard MPEG-1 encoding, is actually too small to indicate that the VA ROI-based encoding deteriorates significantly the quality of produced video. At the same time the bit rate gain, which is about 28% on average (see also Table II), shows clearly the efficiency of VA ROI based encoding. At the same time, VA-ROI outperforms Itti's ROI-based encoding method both in compression ratio (220.9 vs 224.1 kbps) and quality (75 vs 37 selections)..

TABLE I

OVERALL PREFERENCES (INDEPENDENT OF VIDEO CLIP)

Encoding Method	***Preferences***	***Average Bit Rate (Kbps)***
VA-ROI	75	220.9
Itti-ROI	37	224.1
Standard MPEG -1	88	308.1

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_cast1* 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 but two 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.

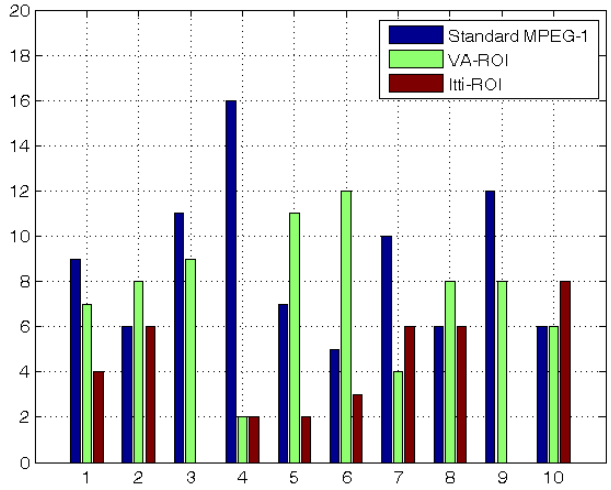


Figure 6. VA ROI-based encoding (greent), Itti-ROI (red) and standard MPEG-1 encoding (blue) preferences on the *eye_witness (1), fashion (2), grandma (3), justice(4), lecturer (5), news_cast1(6), news_cast2 (7), night_interview (8), old_man (9) and soldier (10)* video-clips.

TABLE II

COMPARISON OF VA-ROI, ITTI AND STANDARD MPEG-1 ENCODING IN TEN VIDEO SEQUENCES

Video Clip	***Enc. Method***	***Bit Rate (Kbps)***	***Bit Rate Gain***
eye_witness,	VA-ROI	289	25 (%)
	Itti-ROI	291	25 (%)
	Standard	386	-
fashion	VA-ROI	237	33 (%)
	Itti-ROI	253	29 (%)
	Standard	354	-
grandma	VA-ROI	195	24 (%)
	Itti-ROI	192	25 (%)
	Standard	256	-
justice	VA-ROI	238	25 (%)
	Itti-ROI	274	14 (%)
	Standard	318	-
lecturer	VA-ROI	197	28 (%)
	Itti-ROI	176	36 (%)
	Standard	274	-
news_cast1	VA-ROI	211	29 (%)
	Itti-ROI	229	23 (%)
	Standard	297	-
news_cast2	VA-ROI	181	33 (%)
	Itti-ROI	183	32 (%)
	Standard	270	-
night_interview	VA-ROI	231	31 (%)
	Itti-ROI	214	36 (%)
	Standard	335	-
old_man	VA-ROI	251	22 (%)
	Itti-ROI	220	32 (%)
	Standard	321	-
soldier	VA-ROI	179	34 (%)
	Itti-ROI	209	23 (%)
	Standard	270	-
Average	VA-ROI	220.9	28.3 (%)
	Itti-ROI	224.1	27.3 (%)
	Standard	308.1	-

IV. Conclusions and Further Work

In this paper we have examined the efficiency of VA-ROI encoding for video telephony applications. The algorithm that was involved for identifying the visually salient areas is based on a modification of the Itti's model [9] in which an additional map that accounts for the conscious search performed by humans when looking for faces in a scene, has been incorporated Furthermore, center surround differences are computed using multiscale analysis based on wavelets instead of Gaussian pyramids. Finally, combination of conspicuity maps is obtained through a sigmoid function which saturates the result (final saliency map) in cases of strong stimuli in a particular feature map (intensity, color, orientation or skin).

The results presented indicate that: (a) Significant bit-rate gain, compared to MPEG-1, can be achieved using the VA-ROI based video encoding, (b) the areas identified as visually important by the VA algorithm are in conformance 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 [20] in both encoding gain and video quality.

Further work includes conducting experiments to test the efficiency of the proposed method in the MPEG-4 framework. Furthermore, it is useful to examine the effect of incorporating priority encoding by varying the quality factor of the DCT quantization table across VA-ROI and non-ROI frame blocks.

Acknowledgments

The majority of 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 funded by the Cyprus Research Promotion Foundation [22] and the European Commission under contract FP6-027026-K-SPACE [23]. This publication represents the view of the authors but not necessarily the view of the community.

References

[1]. C. M. Privitera, L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22 (9), pp. 970–982, 2000.

[2]. B. Wandell. Foundations of Vision. Sunderland, MA: Sinauer, 1995.

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

[4]. F.H. Hamker, "Modeling Attention: From Computational Neuroscience to Computer Vision," *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos (editors), Academic Press, 2005.

[5]. R. P. N. Rao, D.H. Ballard, "Probabilistic Models of Attention based on Iconic Representations and Predictive Coding," *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos (editors), Academic Press, 2005.

[6]. K. Cave, "The feature gate model of visual selection". *Psychological Research*, vol. 62, pp.182-194, 1999.

[7]. U. Rutishauser, D. Walther, C. Koch, P. Perona, "Is bottom-up attention useful for object recognition?" *Proceedings of CVPR'04*, pp. 37-44, Jul 2004.

[8]. K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, C. Zetzsche, "Scene analysis with saccadic eye movements: top-down and bottom-up modeling", *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 152-160, 2001.

[9]. 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.

[10]. 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.

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

[12]. N. Tsapatsoulis, 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.

[13]. R. C. Gonzalez, R. E. Woods, *Digital Image Processing*, 2nd edition, Prentice Hall Inc, NJ, 2002, ISBN: 0-13-094650-8.

[14]. Z. Wang, L. G. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," IEEE Transactions on Image Processing, vol. 12, pp. 243–254, 2003.

[15]. 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 *Proc. 25th IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. IV, June 2000, pp. 2345–2348.

[16]. S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 11, pp 674-693, 1989.

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

[18]. [Online] http://www.cs.ucy.ac.cy/~nicolast/research/VAclips.zip

[19]. M. P. Eckert and A.P. Bradley, "Perceptual models applied to still image compression," Signal Processing, 70 (3), pp. 177–200, 1998.

[20]. L. Itti, "Automatic Foveation for Video Compression Using a Neurobiological Model for Visual Attention," *IEEE Transactions on Image Processing*, vol. 13 (10), pp. 1304-1318, 2004.

[21]. Neurmomorphic Vision C++ Toolkit (iNVT), iLab, Univ. Of Southern California, http://ilab.usc.edu/toolkit/.

[22]. [Online] http://www.optopiisi.com

[23]. K-SPACE: Knowledge Space of Semantic Inference for Automatic Annotation and Retrieval of Multimedia Content. [Online]: http://kspace.qmul.net:8080/kspace/