Visual Attention based Region of Interest Coding for Video-telephony Applications

Nicolas Tsapatsoulis[(1)], Constantinos Pattichis[(1)], Anastasios Kounoudes[(2)], Christos Loizou [(3)], Anthony Constantinides[(4)], John G. Taylor[(5)]

[(1)] Department of Computer Science, University of Cyprus
75 Kallipoleos Str., CY-1678, Nicosia, Cyprus

[(2)] Philips College, 4 Lamias Str., CY-2001 Nicosia, Cyprus

[(3)] IMCS Intercollege, 92 Agias Fylaxeos Str., CY-3025, Limassol, Cyprus

[(4)] Dept. of Electrical and Electronic Engineering, Imperial College,
South Kensington Campus, London SW7 2AZ, UK

[(5)] Dept. of Mathematics, King's College London,
Strand Str., WCR2R2LS, London, UK

***Abstract*-Bottom up approaches to Visual Attention (VA) have been applied successfully in a variety of applications, where no domain information exists, e.g. general purpose image and video segmentation. On the other hand, when humans are looking for faces in a scene they perform an implicit conscious search. Therefore, using simple bottom up approaches for identifying visually salient areas in scenes containing humans are not so efficient. In this paper we introduce the inclusion of a top-down channel in the VA architecture proposed in the past (i.e., Itti *et al*) to account for conscious search in video telephony applications. In such kind of applications the existence of human faces is almost always guaranteed. The regions, in the video-telephony stream, identified by the proposed algorithm as being visually salient are encoded with higher precision compared to the remaining ones. This procedure leads to a significant bit-rate reduction while the visual quality of the VA based encoded video stream is only slightly deteriorated, as the visual trial tests show. Furthermore, extended experiments concerning both static images as well as low-quality video show the efficiency of the proposed method, as far as the compression ratios achieved is concerned. The comparisons are made against standard JPEG and MPEG-1 encoding respectively.**

I. INTRODUCTION

The advent of third generation (3G) mobile phones increased the demand for efficient transmission of multimedia data, such as speech, audio, text, images, and video. Of these multimedia data types video data impose the toughest challenges because of its high bandwidth and user expectations in terms of high quality of service. In order to enable the successful adoption of 3G applications, the transmission of multimedia data must be at high compression ratios and be of a perceptually high quality. One 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. Points / areas of fixation are, in several cases, estimated by Visual Attention (VA) models.

In this paper we investigate ROI-based video coding for video-telephony applications. As ROIs we consider the visually salient areas. These areas are automatically detected using an algorithm for visual attention (VA). The proposed algorithm is based on the bottom-up approach proposed by Itti *et al* [3] but is enhanced with a top-down channel emulating the visual search for human faces performed by humans. 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.

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, are presented in Section III. Finally, further work is proposed and conclusions are drawn in Section IV.

II. Saliency-based Visual Attention

The basis of many visual attention models proposed over the last two decades is the Feature Integration Theory of Treisman *et al* [4] that was derived from visual search experiments. According to this theory, features are registered early, automatically and in parallel along a number of separable dimensions (e.g. intensity, color, orientation, size, shape etc).

One of the major saliency-based computational models of visual attention is presented in [3] and deals with static color images. Visual input is first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map, such that only locations that locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master saliency map. Itti and Koch [3],[5] presented an implementation of the proposed saliency-based model. Low-level vision features (color channels tuned to red, green, blue and yellow hues, orientation and brightness) are extracted from the original color image at several spatial scales, using linear filtering. The different spatial scales are created using Gaussian pyramids, which consist of progressively low-pass filtering and sub-sampling the input image. Each feature is computed in a center-surround structure akin to visual receptive fields. Using this biological paradigm renders the system sensitive to local spatial contrast rather than to amplitude in that feature map. Center-surround operations are implemented in the model as differences between a fine and a coarse scale for a given feature. Seven types of features, for which evidence exists in mammalian visual systems, are computed in this manner from the low-level pyramids. The algorithm is summarized in Figure 1 (central part).

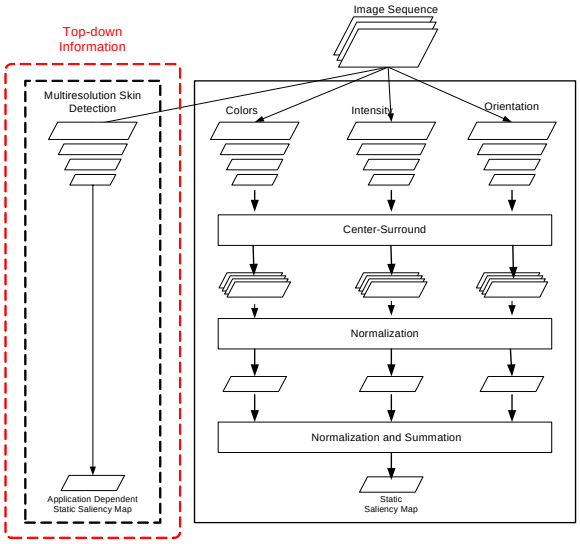

Figure 1: The modified Itti's model for Visual Attention. In the left side is our primary extension for considering the influences conscious search (existence of prior knowledge - in the particular case detection of skin like objects)

Itti's model is 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 [6].Towards this direction we integrate prior knowledge to the saliency-based model in order to draw the attention to regions with specific characteristics (Figure 1-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. In our case we 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 skin detector scheme presented in [7] to generate a skin map with possible face locations and link it with the other feature maps. On the other hand, we cannot identify as ROI only the skin like areas [8],[9] because there is always the possibility, even in a video-telephony setting, that other objects in the scene attract the human interest.

In addition, to adding a top-down branch to Itti's model, we have made several other modifications for better performance and simplicity of implementation. In particular we use the *YCbCr* color space (instead of the *RGB* used by Itti) first to keep conformance with the face detection scheme [7], and second to use the *Y* channel for the intensity and orientation conspicuity maps derivation and the *Cb* ,*Cr* channels for the color conspicuity map. Finally, we use wavelet decomposition instead of Gaussian filtered pyramids for computing the center-surround structure of the feature maps. The overall algorithm is summarized in Appendix A. The full software package (Matlab m-files) can be downloaded from [10].

An example of the visually salient areas identified using the proposed algorithm is shown in Figure 2(g). In Figures 2(c)-2(f) are shown the (normalized) skin, color, intensity, and orientation maps respectively. In Figure 2(b) it is depicted the combined saliency map of all feature maps, while in Figure 2(g) it is presented the actual ROI area created by combining the thresholded (using Otsu's method [11]) bottom-up (color, intensity, and orientation) and top-down (skin) maps. Finally, in Figure 2(h) it is shown the ROI-based JPEG encoded image. 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 4:3.

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 this approach is not optimal in terms of expected file size gains, it has the advantage of producing compressed streams that are compatible with existing decoders. However, this method should be regarded as a worst-case scenario because during motion estimation smoothed or partially smoothed macroblocks may be compared with a non-blurred ones leading to poor motion compensation in the encoder. Video codecs have been proposed to address this problem inherent to any foveated video compression

technique (e.g., encode high-priority regions first, then lower-priority regions, in a continuously variable bit-rate encoding scheme [8]). To simplify the visual evaluation of our algorithm and to evaluate whether the proposed technique might prove useful even with standard codecs, however, we use standard MPEG-1 encoding and simple spatially-variable blur of non-ROI areas prior to compression. Any file size gain obtained despite these limitations would, hence, represent the promise that even better size gains should be obtained with a video codec that would truly use the algorithm's visually salient maps to prioritize encoding.

(a) Original video frame *(b) Overall saliency map* *(c) Skin map*

(d) Color map *(e) Intensity map* *(f) Orientation map*

(g) ROI areas (after post-processing of saliency map) *(h) ROI based encoded image*

Figure 2: Intermediate results of the VA based coding algorithm applied to a single image

Visual trial tests were conducted to examine the quality of the VA-ROI based encoded videos. These tests are based upon ten short video clips, namely: *eye_witness, fashion, grandma, justice, lecturer, news_cast1, news_cast2, night_interview, old_man, soldier* (see [12]). 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 is important to note 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 with the constraints imposed by 3G video telephony. Two low resolution video-clips were created for each case, one corresponding to VA based coding and the other to standard MPEG-1 video coding.

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 [13]. 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). 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 clips they preferred assuming that it had been downloaded over a 3G mobile / wireless network.

B. Results

Table I shows the overall preferences, i.e., independent of (summed over) video clips for standard MPEG-1 and VA ROI-based encoded MPEG-1. It can be seen in Table 1 that there is slight preference to standard MPEG-1 which is selected at 52.5% of the time as being of better quality. However, 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 27% on average (see also Table II), shows clearly the efficiency of VA ROI based encoding.

TABLE I

OVERALL PREFERENCES (INDEPENDENT OF VIDEO CLIP)

Encoding Method	*Preferences*	*Average Bit Rate (Kbps)*
VA-ROI	95	224.4
Standard MPEG -1	105	308.1

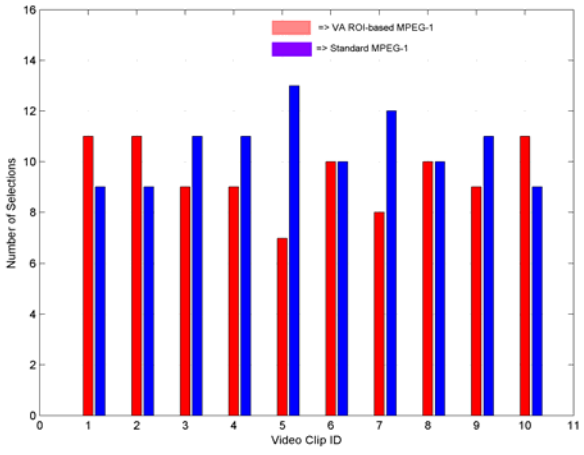

Figure 3: VA ROI-based encoding (left bar) and standard MPEG-1 encoding (right bar) preferences on the eye_witness (1), fashion (2), grandma (3), justice(4), lecturer (5), news_cast1(6), news_cast2 (7), night_interview (8), old_man (9) and soldier (10) video-clips.

In Figure 3, are shown the selections made per video clip. In two of them (*lecturer*, *news_cast2*) there is a clear preference to standard MPEG-1, while VA-ROI based encoding 'wins' three times (*eye_witness*, *fashion*, *soldier*). 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). Preference to VA ROI based encoding, in the above mentioned clips, may be assigned to statistical error or to denoising, performed on non-ROI areas by the smoothing filter. However, it is important to note that there is signifi-cant movement of the main objects in all three video sequences 'won' by VA-ROI based encoding. In contrary, in the two video clips at which there is a clear preference to standard MPEG-1 the main objects are mainly static. These two facts indicate that the deterioration caused by VA ROI- based encoding can be only perceived if the observer has enough time to focus to other areas than the visual salient ones. This is an important observation because it indicates that the areas marked by the VA algorithm as being visually salient are indeed salient.

One important reason for the reduction in preferences of the VA-ROI based encoding is due to the inherently uneven frame quality in the majority of ROI coded video clips. This results in video frames that do not appear natural as the ROI is in sharp focus, whilst the remaining area is blurred. A more gradual change in image quality between ROI and non-ROI would, however, increase both the size of the ROI, having a negative impact on ROI coding efficiency, as well as on the encoding complexity.

Table II presents the bit-rates achieved for both the VA ROI based encoding and standard MPEG-1 in the individual video clips. It is clear that the bit rate gain obtained is significant, ranging from 15% to 48%. Furthermore, it can be seen from the results obtained in the *night_interview* video sequence, that increased bit-rate gain does not necessarily mean worse quality of the VA ROI encoded video.

TABLE II

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

Video Clip	*Encoding Method*	*Bit Rate (Kbps)*	*Bit Rate Gain*
eye_witness,	VA-ROI	319	17 (%)
	Standard	386	
Fashion	VA-ROI	296	16 (%)
	Standard	354	
grandma	VA-ROI	217	15 (%)
	Standard	256	
justice	VA-ROI	228	28 (%)
	Standard	318	
lecturer	VA-ROI	201	27 (%)
	Standard	274	
news_cast1	VA-ROI	205	31 (%)
	Standard	297	
news_cast2	VA-ROI	170	37 (%)
	Standard	270	
night_interview	VA-ROI	174	48 (%)
	Standard	335	
old_man	VA-ROI	241	25 (%)
	Standard	321	
soldier	VA-ROI	193	29 (%)
	Standard	270	
Average	VA-ROI	224.4	27.2 (%)
	Standard	308.1	

Bit-rate gain achieved by JPEG encoding of the individual video frames (not shown in Table II) is on average about 21% (ranging from 14% to 28%). This indicates that the bit-rate gain is mainly due to the compression obtained for Intra-coded (I) frames than for the Inter coded (P,B) ones. This conclusion strengthens the argument that smoothing of non-ROI areas may decrease the efficiency of motion compensation.

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 [3] in which an additional map that accounts for the conscious search performed by humans when looking for faces in a scene, has been incorporated. 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 based encoding leads to better compression of both Intra-coded and Inter coded frames though the former is higher.

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.

Acknowledgment

The majority of the study presented in this paper has been undertaken in the framework of 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 [14].

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]. 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.

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

[5]. 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.

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

[7]. 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.

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

[9]. 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.

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

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

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

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

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

Appendix: Summary of the VA Coding Algorithm

```
-------------------------------------------------------------------------------
function Mask = TotalMap(fileIn,fileOut1 fileOut2, QualFactor)

 f = imread(fileIn);                % load individual frame file
SkinMap = SkinDetection(f);         % Compute skin map
% Threshold, and identify the major objects in skin map
% Model objects as ellipses to account for non- compactness
SkinMask = MainSkinMasks(SkinMap);
% Compute the combined bottom-up map which includes color, intensity
% and orientation conspicuity maps
AtMap = AttMap(f);
% Threshold bottom-up map using Otsu's method
AtMask = im2bw(AtMap,graythresh(AtMap));
% Combine top-down and bottom-up masks
Mask = SkinMask | AtMask;       % logical OR
% Filter (smooth) non-ROI areas
h = fspecial('disk',radi); g =imfilter(f,h);  g(Mask)=f(Mask);
% Encode smoothed frame as JPEG frame
imwrite(uint8(g), ,fileOut1,'jpeg','quality',QualFactor);
% Encode non-smoothed frame with the same quality factor
imwrite(uint8(f), ,fileOut2,'jpeg','quality',QualFactor);
-------------------------------------------------------------------------------
function AtMap = AttMap(f)

z = rgb2ycbcr(f);                  % Convert to YCbCr space
% Compute the depth of wavelet pyramid
M=size(f); N=min(M(1),M(2));    d=ceil(floor(log2(N))/2);
% Compute the pyramid for intensity orientation and color using
% Daubechie's wavelets
z1=z(:,:,1); z2=z(:,:,2); z3=z(:,:,3);
[c1,s]=wavedec2(z1,d,'db2'); [c2,s]=wavedec2(z2,d,'db2');
[c3,s]=wavedec2(z3,d,'db2');
% Compute orientation map as differences in gradient angle
hx = fspecial('sobel'); hy=hx'; Gx = double(imfilter(z1,hx,'replicate'));
Gy = double(imfilter(z1,hy,'replicate'));
for i=1:d
   % Lower approximation
   c01=[c1(1:s(i+1,1)*s(i+1,2)) zeros(1,length(c1)-s(i+1,1)*s(i+1,2))];
   c02=[c2(1:s(i+1,1)*s(i+1,2)) zeros(1,length(c2)-s(i+1,1)*s(i+1,2))];
   c03=[c3(1:s(i+1,1)*s(i+1,2)) zeros(1,length(c3)-s(i+1,1)*s(i+1,2))];
   a01=[a1(1:s(i+1,1)*s(i+1,2)) zeros(1,length(a1)-s(i+1,1)*s(i+1,2))];
   r01=waverec2(c01,s,'db2');     r02=waverec2(c02,s,'db2');
   r03=waverec2(c03,s,'db2');     ra01=waverec2(a01,s,'db2');
   % Higher approximation
   c11=[c1(1:s(i+2,1)*s(i+2,2)) zeros(1,length(c1)-s(i+2,1)*s(i+2,2))];
   c12=[c2(1:s(i+2,1)*s(i+2,2)) zeros(1,length(c2)-s(i+2,1)*s(i+2,2))];
   c13=[c3(1:s(i+2,1)*s(i+2,2)) zeros(1,length(c3)-s(i+2,1)*s(i+2,2))];
   a11=[a1(1:s(i+2,1)*s(i+2,2)) zeros(1,length(a1)-s(i+2,1)*s(i+2,2))];
    r11=waverec2(c11,s,'db2');    r12=waverec2(c12,s,'db2');
   r13=waverec2(c13,s,'db2');     ra11=waverec2(a11,s,'db2');
  % Compute the differences in fine and coarse scales
   Imap=Imap+abs(r11-r01) ;
   Cmap=Cmap+abs(r12-r02)+abs(r13-r03);
   Omap = Omap+abs(ra11-ra01);
end
```