BIOLOGICALLY INSPIRED REGION OF INTEREST SELECTION FOR LOW BIT-RATE VIDEO CODING

Nicolas Tsapatsoulis, Constantinos Pattichis

University of Cyprus
Department of Computer Science
CY-1678, Nicosia, Cyprus
email: {nicolast,pattichi}@ucy.ac.cy

Konstantinos Rapantzikos

National Technical University of Athens
School of Electrical & Computer Engineering
GR-15780, Zographou, Athens, Greece
email: rap@image.ntua.gr

ABSTRACT

A variety of approaches have been proposed in the literature for Region-Of-Interest (ROI) estimation. In most of them the ROI definition is highly subjective, i.e., lacks scientific evidence in supporting the claim that the areas defined as ROIs are indeed perceptually interesting. In this paper we attempt to model ROIs as the visually attended areas indicated by a saliency map in order to lower as much as possible the subjectivity of selection. For evaluation purposes we follow a ROI-based video compression setup and present comparisons with state-of-the-art algorithms in terms of perceived visual quality and video compression improvement,. 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.

Index Terms— Visual attention, saliency map, region of interest, video coding, MPEG-4

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 ROI coding [1]. In medical video transmission it is imperative to include in the ROI area the clinically important parts of the video frames. 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 [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 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 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 selected areas of interest are in conformity with human perception. In this paper we attempt to detect ROIs based on a saliency map [3] so as to enhance conformity with human perception. A computationally efficient way for identifying salient regions in images, based on bottom-up information [4], 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.

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

2. SALIENCY MAP ESTIMATION

The idea of attention deployment dates back to the pioneering work of James [5], the father of American psychology. Several theoretical models have been proposed in the past using a two component attention framework, consisting of a top-down and a bottom-up component, as proposed by James. A computational modelling of this theoretical framework was first proposed by Koch & Ullman [3]. The core idea is the existence of a saliency map that combines information from

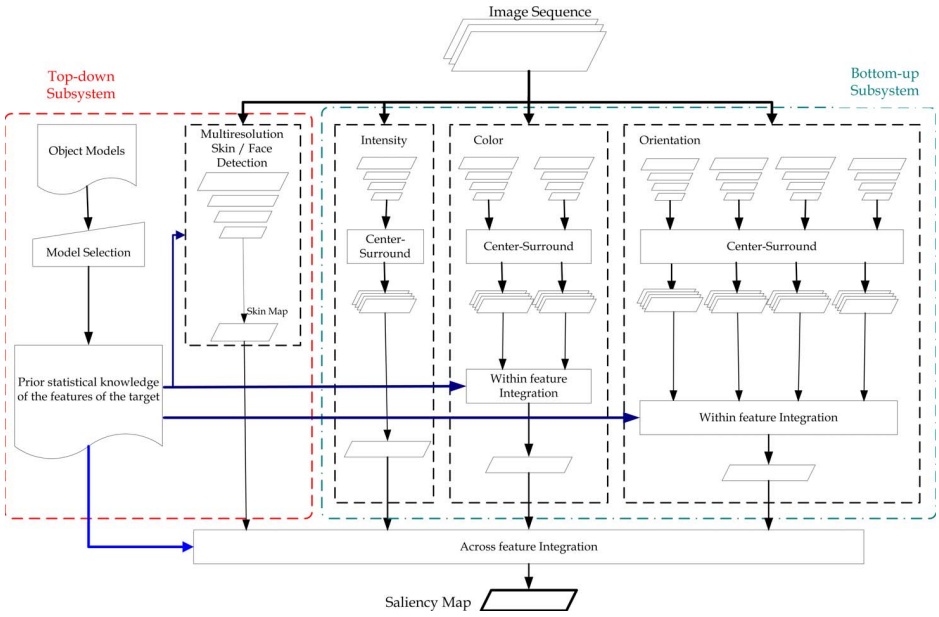

Fig. 1. The overall architecture of the Wavelet-based VA model for saliency map estimation

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. One of the most successful saliency-based computational models was proposed by Itti *et al.* [4]. It is based on the same principles of 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. Finally, all feature maps are linearly combined to generate the overall saliency map.

In addition to saliency-based visual attention models there is another category of models known as 'goal directed'. Attentional selection in these models is based on the principle that humans are directing their attention on known 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 of goal directed approaches simply outline a general framework [6] on how to use target knowledge for identifying visually salient areas.

2.1. The proposed Visual Attention Model

The proposed Visual Attention model for saliency map estimation 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. In the context of this work we only use a previously developed [7] statistical model for human skin representation 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. 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 . The feature level is related to the bias applied to specific features in order to enhance regions similar to the prior model.

Let us consider a video-telephony scenario, where one or more faces are present. The proposed scheme is, then, activated as follows: (1) The skin model is selected as the one to bias further analysis and the input sequence is analyzed in different feature dimensions; (2) 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; (3) 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); (4) all conspicuity maps are again weighted using top-down gains and finally fused to generate the saliency map.

Detailed presentation of the proposed model can be found at [8]. A Simulink$^{(R)}$ implementation of the proposed model, that includes the frame encoding process, is available online (*http://www.cs.ucy.ac.cy/∼nicolast/research/VAmodel.zip*).

3. ROI GENERATION AND VIDEO ENCODING

The visually interesting areas highlighted by a saliency map computed through the proposed model, can be used for the creation of ROIs. For this purpose 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 the spatial decorrelation obtained by applying either the DCT transform (MPEG-1) or wavelet decomposition (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 concentrates on visually salient areas and does not perceive deterioration in non-visually important areas. Smoothing non-ROI areas is not optimal in terms of expected encoding gain but has the advantage of producing compressed streams that are compatible with existing decoders. The quality of the VA-ROI based encoded videos is evaluated through a set of visual trial tests conducted on ten short video clips, namely: *fashion, eye_witness, grandma, justice, news_cast1, news_cast2, lecturer, night_interview, old_man, soldier* (see also: *http://www.cs.ucy.ac.cy/∼nicolast/research/frames.rar*). 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 with a frame resolution of 144x176 pixels and a frame rate of 15 fps. The *ImTOO MPEG Encoder* (*http://www.imtoo.com/*) 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 [9] (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.

4. VISUAL TRIAL TESTS AND CODING RESULTS

Visual trial tests were conducted to directly compare the subjective visual quality of VA-ROI based, IttiROI based, and streamline MPEG-1 and MPEG-4 video encoding. ROI areas were determined using the proposed saliency map estimator for the VA-ROI method and the Neuromorphic Vision Toolkit (*http://ilab.usc.edu/toolkit/*) for Itti's method. In both cases saliency maps were thresholded using Otsu's method [10] to create the binary masks that correspond to ROI areas. A three alternative forced choice (3AFC) methodology was selected, i.e., the observer views the three differently encoded video clips and then selects the one preferred. There were ten observers, all being non-experts in image compression (university students). The video clip triples were viewed one at a time in a random order. Each triple was viewed twice, giving (10x10x2) 200 comparisons. Video-clips were viewed on a Smartphone display in the case of the MPEG-4 videos and on a typical PC monitor in the case of the MPEG-1 videos. Both the MPEG-1 and the MPEG-4 encoded videos were tested through a visual trial. Table 1 shows the preferences, bit-rate and bit-rate gain w.r.t the standard MPEG-1, for the Itti-ROI and the proposed VA-ROI-based method. The results refer to five indicative video clips (due to space limitations). However, averages in this table, refer to all ten video clips (online at: *http://www.cs.ucy.ac.cy/∼nicolast/research/videos.rar*) to allow comparisons both in terms of visual quality and bit rate gain. A slight preference to standard MPEG-1, which is selected at 46.5% (9.3 selections on average) of the time as being of better quality, is observed. The difference in selections, between VA-ROI based (selected at 45.5% of the time) 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, 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. The slightly increased encoding gain (41% on average), compared to the VA-ROI method, does not trade off this lowering in perceived visual quality. In Table 2 the preferences and bit-rates achieved by VA-ROI based, IttiROI based and standard MPEG-4 encoding in five video clips are also presented. Once again averages refer to all ten video clips. 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. Encoding gain, in the IttiROI encoded videos, presents higher variance across the various video clips; 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. 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.

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

Video Clip	Encoding method	Prefs	Bit Rate (Kbps)	Bit Rate Gain
grandma	VA-ROI	11	1507	15.2%
	Itti-ROI	0	1300	33.6%
	Std MPEG-1	9	1737	
justice	VA-ROI	7	1468	30.5%
	Itti-ROI	5	1606	19.3%
	Std MPEG-1	8	1916	
lecturer	VA-ROI	7	950	57.3%
	Itti-ROI	0	848	76.3%
	Std MPEG-1	13	1495	
old_man	VA-ROI	10	1307	32.9%
	Itti-ROI	0	1085	60.0%
	Std MPEG-1	10	1737	
soldier	VA-ROI	8	500	63.9%
	Itti-ROI	4	614	33.3%
	Std MPEG-1	8	819	
Average (over all video clips)	VA-ROI	9.1	1125	35.7%
	Itti-ROI	1.6	1081	41.2%
	Std MPEG-1	9.3	1527	

Table 2. Comparison of VA-ROI based, IttiROI based and Standard MPEG-4 encoding in five video sequences

Video Clip	Encoding method	Prefs	Bit Rate (Kbps)	Bit Rate Gain
fashion	VA-ROI	8	288	10.4%
	Itti-ROI	4	381	11.7%
	Std MPEG-4	8	439	
grandma	VA-ROI	8	264	11.9%
	Itti-ROI	3	247	19.5%
	Std MPEG-4	9	296	
lecturer	VA-ROI	8	107	28.3%
	Itti-ROI	3	110	25.3%
	Std MPEG-4	9	138	
old_man	VA-ROI	10	217	13.3%
	Itti-ROI	0	189	30.3%
	Std MPEG-4	10	246	
soldier	VA-ROI	7	71	22.5%
	Itti-ROI	6	79	11.4%
	Std MPEG-4	7	87	
Average (over all video clips)	VA-ROI	7.9	197.5	13.7%
	Itti-ROI	4.2	194.0	15.7%
	Std MPEG-4	7.9	224.6	

5. CONCLUSIONS AND FUTURE WORK

In this paper a saliency-based visual attention model was applied to identify regions of interest for ROI-based video coding. Two separate information channels are processed; a high-level one which models conscious search for diagnostic rich image/farme regions and a low-level one which models subconscious attention attraction. ROI-based encoding is then applied by smoothing the non-ROI areas with a median 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-2 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 conformance with the ones identified by the humans, and (c) VA-ROI outperforms the VA method proposed by Itti [9] in terms of visual quality but achieves slightly lower encoding gains. Further work includes conducting experiments in an object basis framework where the disjoint ROI areas will be considered as objects.

6. REFERENCES

[1] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 970–982, 2000.

[2] B. A. Wandell, *Foundations of Vision*, Sinauer Associates, Sunderland, MA 01375, 1995.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[4] 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, no. 11, pp. 1254–1259, 1998.

[5] W. James, *The Principles of Psychology*, Cambridge, MA: Harvard University Press, 1890/1981.

[6] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, Springer Berlin/Heidelberg, 2006.

[7] N. Tsapatsoulis, Y. Avrithis, and S. Kollias, "Facial image indexing in multimedia databases," *Pattern Analysis and Applications*, vol. 4, no. 2/3, pp. 93–107, 2001.

[8] K. Rapantzikos and N. Tsapatsoulis, "A committee machine scheme for feature map fusion under uncertainty: the face detection case," *Int. J. of Intell. Syst. Tech. and Applications*, vol. 1, no. 3/4, pp. 346–358, 2006.

[9] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. on Image Proc.*, vol. 13, pp. 1304–1318, 2004.

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