Low-complexity Motion-based Saliency Map Estimation for Perceptual Video Coding

Ana Belén Mejía-Ocaña, Manuel de-Frutos-López, Sergio Sanz-Rodríguez, Óscar del-Ama-Esteban, Carmen Peláez-Moreno, Fernando Díaz-de-María
Department of Signal Theory and Communications
University Carlos III Madrid
Leganés, Spain
mfrutos@tsc.uc3m.es, abelen@tsc.uc3m.es

Abstract—**In this paper, a low-complexity motion-based saliency map estimation method for perceptual video coding is proposed. The method employs a camera motion compensated vector map computed by means of a hierarchical motion estimation (HME) procedure and a Restricted Affine Transformation (RAT)-based modeling of the camera motion. To allow for a computationally efficient solution, the number of layers of the HME has been restricted and the potential unreliable motion vectors due to homogeneous regions have been detected and specially managed by means of a smooth block detector. Special care has been taken of the smoothness of the resulting compensated camera motion vector map to avoid unpleasant artifacts in the perceptually-coded sequence, by including a final post-processing based on morphological filtering. The proposed saliency map has been both visually and subjectively assessed showing quality improvements when used as a part of the H.264/AVC standard codec at medium-to-low bitrates.**

Keywords-region of interest; perceptual video coding; visual saliency; visual attention; hierarchical motion estimation; camera motion estimation; mathematical morphology.

I. INTRODUCTION

Hybrid video coding has been the core technique in virtually every standard recommendation for the last two decades. This algorithmic structure aims to reduce temporal and spatial redundancy by optimizing the trade-off between the resulting bitrate and an objectively measured distortion.

Perceptual video coding, on the other hand, introduces knowledge about the Human Visual System (HVS) for allocating larger amounts of bits in the regions where visual attention is focused. This new paradigm is expected to provide a step forward in the achievement of better quality coding, especially for medium to low bit rates. The automatic estimation of the locations of the most prominent regions from a subjective point of view is, however, a challenging issue [1].

An optimal solution can be obtained from direct observation of the target viewer's gaze by using eye-tracking devices. Unfortunately, these devices are not usually available for video coding purposes and therefore suboptimal solutions must be sought.

Several authors concentrate their efforts in the detection of faces since, for some specific applications such as teleconferencing or broadcast news, they are likely to attract the viewers' attention [2]. Obviously, the main drawback of this approach is the lack of generalization ability to other kinds of video contents. In [3] an alternative approach proposes the use of learnt-by-example attention functions that could be adapted to a particular application.

Yet another alternative relies on the observation of an uneven perception of distortions by the HVS depending on the source of the distortion and the type of the region being encoded, allowing higher distortion in those regions of the image where it is likely to be less noticeable. This formulation does not take into account the saliency of these regions, that clearly affects the overall quality perception and, therefore, they are normally employed jointly (for example in [4]).

In this paper we describe a content-independent motion-based algorithm for the generation of a continuous-valued saliency map that is based on the displacement of the objects relative to the camera. Special attention has been given to the temporal consistency of the motion of the different objects, which has been inferred by combining three main modules:

1) A hierarchical motion estimation procedure that overcomes the well-known shortcomings of block-matching algorithms, while maintaining the computational complexity considerably lower than that of optical flow –based methods. A smooth block detector allows us to overcome the potential errors in large homogeneous regions of the images where the computed motion vectors become unreliable.

2) A camera motion estimation model. A preprocessing step is used for removing unreliable blocks that result from the previous hierarchical motion estimation.

3) A morphological filtering postprocessing that allows for smoothing temporal and spatial transitions.

This paper is organized as follows: Section II provides an overview of the proposed and describes in detail each one of its subsystems. Section III focuses on the evaluation of the method proposed, and Section IV draws some conclusions and outlines further work.

II. ALGORITHM DESCRIPTION

Motion is a highly salient feature which grabs one's attention and keeps it locked on important features and objects. Interest in motion perception has a long history and it can currently be considered a relatively well established discipline. However, the understanding of the contribution of top-down influences, like attention, on neural activity is still a research subject [5].

Our proposal focuses on detecting those regions in each frame of the video sequence that could potentially attract the attention of most viewers due to motion. In [4] and [6], a method to estimate motion attention based on the motion vectors of the video codec (derived using a common block-matching procedure) is employed.

On the other hand, the proposed algorithm (summarized in Figure 1) carries out a precise Hierarchical Motion Estimation (HME) for obtaining the real motion of the objects in the scene, followed by a Camera Motion Estimation process (CME). Then, a map of camera motion compensated vectors is obtained by subtracting the latter from the former.

However, due to computational constraints, the HME procedure exhibits a high sensitivity to smooth regions producing erroneous results in those frames where the amount of blocks of such type is significant. For this reason we have developed a 'smoothness detection' process that detects and removes those vectors from the CME. Then, the vector field is further filtered using a 'closing' operation to prevent isolated spurious blocks and, as will be described later, obtaining a more coherent and realistic motion map. Finally, the resulting saliency map is a normalized version of this compensated and filtered motion vector map. Detailed descriptions of each stage follow in the next subsections.

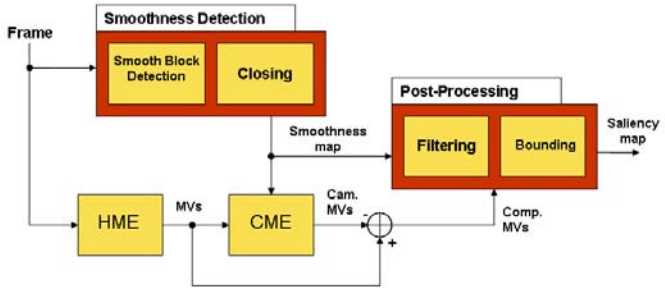

Figure 1. Block diagram of proposed saliency classification algorithm.

A. Hierarchical Motion Estimation

For our purpose of detecting the regions of interest (ROI) we must obtain an estimation of the real motion of the video sequence as it would have been perceived by a human observer. For this reason, we cannot employ common block-matching algorithms such as the ones usually employed in standard video codecs, where practical implementations usually employ suboptimal heuristic search algorithms. On the other hand, we are also concerned with the computational complexity of the overall process and therefore we rule out the use of optical flow motion estimation algorithms [7] that, in spite of offering a closer prediction to motion perception, they substantially increment the computational cost. Therefore, we propose an approach that provides a trade-off between efficiency and computational complexity: a Hierarchical Block-Matching Algorithm.

The hierarchical motion estimation computes local motion vectors following a sequence of progressive refinement stages that starts from an initial coarse estimation. Specifically, it involves the use of N levels where each one employs a target block size (BS) that decreases.

Bearing in mind that state-of-the-art video codecs employ 16x16 pixel macroblocks (MB) as the basic encoding units, we have decided to use the same size in the last layer (the finest) of the estimation process. However, after having tested the algorithm for several values of N, we detected that wrong estimations tend to ocurr in large homogeneous regions. The use of more levels with larger BS reduces the incidence of these errors, but in exchange for an additional computational cost. Therefore, a trade-off value of 3 has been adopted for N, resorting to a smoothness detector to avoid wrong estimations, as will be described in next subsection.

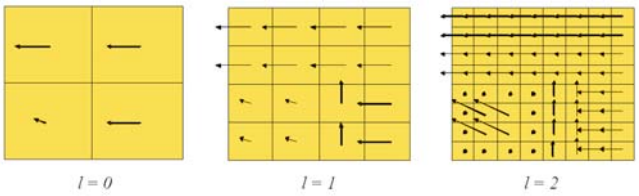

Figure 2. Hierarchical motion estimation layers

Consequently, blocks sizes for each layer are 64x64, 32x32 and 16x16 pixels respectively. An example of MV for these layers is depicted in Figure 2. In order to complete the algorithm configuration, the Search Area (SA), the region of the reference frame where the search of the best match takes place, is established, regardless of the level, to a square of 64x64 pixels centered in the left-top pixel of each analyzed block. Finally, it is worth noting that the abovementioned configuration parameters have proved to be suitable for both CIF (352x288) and SD (704x576) video resolutions.

Motion vectors ($\mathbf{mv}^l$) for each layer l are calculated by minimizing a cost function over all the possible search vectors ($\mathbf{sv}$) in the SA:

$$\mathbf{mv}^l = \arg\min_{\mathbf{sv}} \{Cost(l, \mathbf{sv})\}, \quad (1)$$

where the specific cost function, based on the Mean Absolute Differences (MAD), is described as follows:

$$Cost(l, \mathbf{sv}) = \begin{cases} MAD(\mathbf{sv}) + \alpha_1 |\mathbf{sv}|, & l = 0 \\ MAD(\mathbf{sv}) + \alpha_2 |\mathbf{sv} - \mathbf{mv}^{l-1}| + \alpha_3 |\mathbf{sv}|, & l > 0 \end{cases} \quad (2)$$

Constants α_1, α_2 and α_3 are regularization parameters: α_2 penalizes high deviations of the candidate, **sv,** from estimations in previous levels, represented by the difference with respect to $\mathbf{mv}^{l-1}$, whereas α_1 and α_3 prevent large motion vectors, namely

outliers, in the understanding that they do not represent real movement.

The final motion vector map (**MV**) consists of those motion vectors chosen for each block in the last layer, namely **mv**^{2}.

Figure 3. Motion vector map calculated by HME.

B. Smoothness Detection

Even though the use of HME avoids some of the common errors of standard block-matching in finding the actual motion, still some vectors fail to follow it. As can be observed in Figure 3, where the camera performs an horizontal panning to the left following the wandering couple with the dog (and, therefore, the background of the image should be showing an homogeneous field of horizontal vectors of the same modulus), that large smooth regions (like the facades of the buildings) tend to produce small MAD terms no matter the search position and, therefore, for these kind of blocks, the HME produces a 'zero' output.

In order to solve this drawback, without increasing L, a smooth block detector is proposed. Given that these homogeneous regions are typically encoded with very few bits regardless of the encoding mode, they will be excluded from the process in order to avoid any negative influence of the unreliability of the estimated MVs.

In particular, the first step in the detection of smooth blocks is the calculation of the portion of the total energy of the block associated with the DC coefficient. Therefore, for each block with coordinates (x, y) we compute:

$$E_{DC}(x,y)=\frac{(\sum_{i,j} p_{ij})^2}{\sum_{i,j} p_{ij}^2}, \tag{3}$$

as an indicator of its smoothness, where p_{ij} are the luminance values of every pixel within the block.

Given that we are only interested in large homogeneous regions, before deciding on the smoothness of a block, the E_{DC} map is further refined by means of a morphological opening.

The erosion process entails the substitution of each component of E_{DC} by the minimum value in a 2x2 square neighborhood, removing isolated high values and generating an eroded map E_{DC}^e of DC energy values:

$$E_{DC}^e(x,y)=\min\{E_{DC}(x+\Delta x, y+\Delta y)\}, \tag{4}$$

with Δx=0..1 and Δy=0..1. Next, a dilation process completes the morphological opening and generates the filtered DC energy map E_{DC}^o by means of the substitution of each value in E_{DC}^e by the maximum value in a 2x2 square neighborhood, recovering some high DC energy values previously eroded in large homogeneous regions:

$$E_{DC}^o(x,y)=\max\{E_{DC}^e(x-\Delta x, y-\Delta y)\} \tag{5}$$

After the morphological opening, map E_{DC}^o is compared to an empirical threshold, th_1. The result is a binary Smoothness Map (*SM*), with 'zero' value for non-homogeneous blocks and 'one' for those considered smooth:

$$SM(x,y)=\begin{cases}1, & if\ E_{DC}^o(x,y)>th_1\\ 0, & Otherwise\end{cases} \tag{6}$$

The example in Figure 4 illustrates how the opening operation over the original map improves the final *SM* (light gray meaning smooth regions).

C. Camera Motion Estimation

Similarly to [8], modeling of camera motion is based on a Restricted Affine Transformation (RAT) as follows:

$$\mathbf{RAT}=\begin{bmatrix} s\cdot\cos\theta & s\cdot\sin\theta & t_x\\ -s\cdot\sin\theta & s\cdot\cos\theta & t_y\\ 0 & 0 & 1\end{bmatrix}, \tag{7}$$

where s is the scale, θ is the rotation angle, and the vector (t_x, t_y) represents the motion translation. In order to obtain these parameters, we resort to the LMS algorithm (Least Median Squares) and a robust technique called Random Sample Consensus (RANSAC), as described in [9].

First, the camera motion is obtained from the MV field generated by the HME module. Considering that the camera usually focuses on a central object in the scene, and either keeps static or moves along with it, the central area of each frame does not provide significant information about camera motion, so we choose to exclude this region from the parameterization process. Besides, the SM obtained in the previous stage indicates which of the blocks are homogeneous and, consequently, may involve unreliable MV estimation. These blocks will also be removed from the camera motion calculation, completing the exclusion mask.

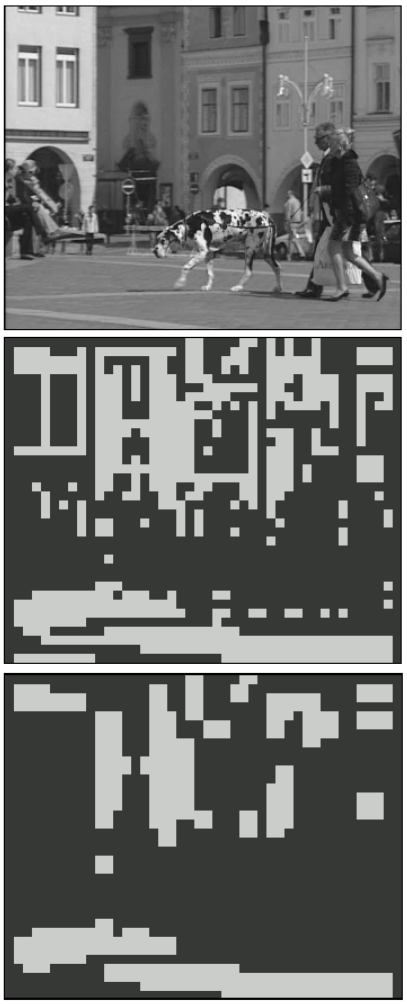

Figure 4. a) A sample frame of the sequence "bohemia". b) SM without morphological filtering. c) SM with morphological filtering.

Once the appropriate vectors to estimate the camera motion are known, the RAT model needs to be initialized for each frame. In the first frame, the rotation is initialized to zero, as well as the scale parameter, and the translation parameters are initialized to the average of MV map components. In subsequent frames the initial transformation parameters are the RAT values of the previous frame.

An LMS algorithm is trained with each one of M subsets of random valid vectors in order to achieve M RAT models, optimal from the MSE (Mean Square Error) point of view. According to RANSAC, the optimum number of subsets to represent the global camera motion is calculated as follows:

$$M = \frac{\log(1-P)}{\log(1-\rho^k)}, \tag{8}$$

where P, fixed to 0.99, is the probability to find at least one subset containing only background vectors; k is the number of elements per group, and it is set to 4 due to complexity reasons; and ρ is a lower threshold on the proportion between the background blocks and the total blocks in the frame. Due to the use of the exclusion mask, we consider that at least half of the blocks in each subset belong to background, so $\rho = 0.5$.

The training of RAT parameters for each subset stops when the difference between the MSE measured in the current iteration and the previous is less than a threshold. In order to reduce potential abrupt variations from one frame to the next, the RAT parameters of each subset are also exponentially averaged with those of previous frame. Finally, the best RAT model among all subsets is the one that minimizes the prediction error over the MVs outside the aforementioned exclusion mask. Though in [9] the median square error is recommended as a robust measure, the use of the mean is simpler and barely damages the performance.

Once the camera motion map or $\mathbf{MV}_{\text{cam}}$ is built, the compensated motion vector map is computed by subtracting it from the original motion vector field as follows:

$$\mathbf{MV}_{comp}(x,y) = \mathbf{MV}(x,y) - \mathbf{MV}_{cam}(x,y) \tag{9}$$

D. Post-processing

This stage aims to obtain the final saliency map, which could be used by a perceptual video encoder to allocate more resources to those blocks considered as belonging to the moving ROI than to those associated with regions belonging to the background or non-relevant (static) objects, as will be shown by the second experiment described in Section III.

First of all, given that local abrupt variations in the resource allocation within the frame could result in noticeable and undesirable artifacts, the motion vector map obtained in the previous stage is smoothed by filtering each of the Cartesian components of the vectors in order to integrate those blocks that are coherent in both motion direction and magnitude.

This filtering entails the use of a 3x3 square mask, depicted in Figure 5 (a), in which the center value contributes a 40% to the output value and the remaining 60% is equally contributed by the eight nearest neighbors. Additional constraints need to be imposed to the filtering process in order to ignore those values belonging to the excluded smooth blocks. Figure 5 (b) illustrates the modified filtering mask, in which the energy of the ignored homogeneous blocks (shaded in the figure) is redistributed through the rest of the blocks according to:

$$\mathbf{MV}'_{comp}(x,y) = \sum_{u=x-1}^{x+1}\sum_{v=y-1}^{y+1} C_{uv}\cdot\mathbf{MV}_{comp}(u,v)\cdot(1-SM(x,y)), \tag{10}$$

with C_{uv} calculated as:

$$C_{uv} = \begin{cases} 0.4, & \text{if } u = x \text{ and } v = y \\ \dfrac{c_1}{N}, & \text{otherwise} \end{cases}, \tag{11}$$

where N is the number of non-smooth blocks within the 3x3 square mask, and c_1=0.6 for masks centered in non-smooth blocks and c_1=1 for masks centered in smooth blocks.

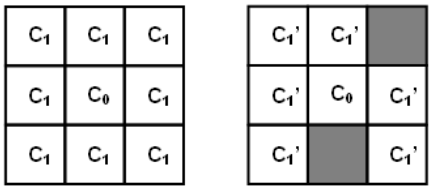

Figure 5. a) Standard 3x3 filtering mask, with C_0=0.4 and C_1=0.075. b) Mask with C_0=0.4 and C_1=0.1 due to the exclusion of smooth blocks (shaded).

Finally, the next step is the calculation of the saliency map, $S(x, y)$, which has been designed to be zero for irrelevant (static) regions and one for regions of interest (moving objects), and is calculated as follows:

$$S(x,y) = \frac{\min\left\{MV_{bound}, \left|\mathbf{MV}'_{comp}(x,y)\right|\right\}}{MV_{bound}}, \qquad (12)$$

where $|\cdot|$ denotes the magnitude of the motion vector and MV_{bound} is proportional to the video resolution, with values of 5 and 10 pixels respectively for CIF and SD resolutions. The use of this upper bound is motivated by the fact that, exceeding a certain amount of motion, faster objects do not attract more intensely the observer's attention due to the limitations of the HVS (an object with a 20 pixel displacement from one frame to the next is not more salient than another moving 10 pixels).

III. Experiments and Results

The evaluation of our proposal has been performed in two different ways: first, a visual inspection of the resulting saliency maps for different video typologies; and second, an attention-based rate control has been implemented for the H.264/AVC coding standard and evaluated by means of a subjective test.

First, the saliency map for various CIF and SD video sequences has been depicted in grayscale, with black representing minimum saliency and white representing maximum saliency. Sample frames can be observed in Figures 6 to 9. As can be seen, for sequences with no camera motion such as those in Figure 6 and Figure 8, moving objects are considered as the ROI and obtain higher values of S. On the other hand, for video sequences with camera motion, such as those in Figure 7 and Figure 9, the higher values of saliency correspond to those objects being followed by the camera and, therefore, static with respect to the observer. Interestingly, in the case of Figure 7, a small region around the logo bug is classified as a high interest region given that its static situation appears as if it was perfectly followed by the camera.

Additionally, the figures confirm the smoothness of the saliency map, which is a key aspect for the integration of our proposal into a perceptual rate control. It is also worth noting that even in sequences with large homogeneous regions, such as that in Figure 9, the ROI is properly detected.

Figure 6. Saliency map for a frame of CIF sequence "football".

Figure 7. Saliency map for a frame of CIF sequence "bus".

Figure 8. Saliency map for a frame of SD sequence "paris".

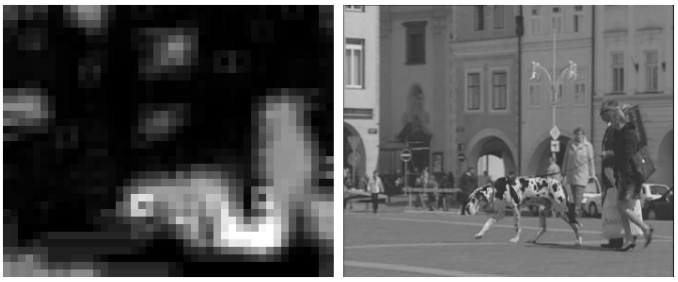

Figure 9. Saliency map for a frame of SD sequence "bohemia".

The second part of the experiments performed involves the integration of our proposal in a perceptual video encoder. From the three main techniques for perceptual resource allocation described in [1], we have chosen the selective application of blur filtering, in the form of a 3x3 Gaussian mask, which removes higher frequencies and, therefore, reduces the bit-rate in those regions less relevant to the HVS. The standard deviation for the Gaussian filter is thus determined according to the inverse of our S map. However, the procedure for selectively introducing distortion in low attention regions is a complex mechanism out of the scope of this paper.

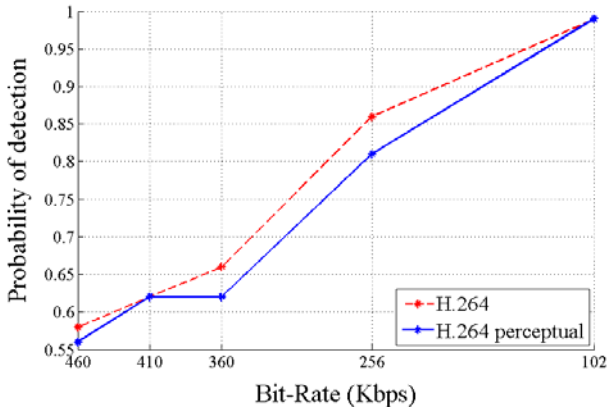

Figure 10. Probability of detecting the low bit-rate encoded version

A subjective test for assessing the performance of this perceptual encoder has been designed following the ITU recommendation for subjective quality assessment in video coding [10]. Specifically, a simplified version of the Pair Comparison (PC) method, recommended for its high discriminatory ability when the test items are almost identical in quality, has been carried out in order to determine whether a bit-rate reduction is less noticeable with the perceptual video encoder than with the standard video encoder,. To this end, a database of 45 video clips with different resolutions and motion contents has been selected and two sets of sequences have been generated: the first set consist of H.264 encodings at different bit-rates (512, 460, 410, 360, 256 and 102 Kbps) of the original video clips; the second set is composed of encodings at the same bit-rates of the corresponding blurred versions of the sequences, with stronger blur filtering applied to regions with lower *S* values. The H.264 encoder has been configured in Main profile at 25 frames per second, with rate distortion optimization and CABAC. IPPP… has been selected as encoding pattern and the full search motion estimation has been performed with 5 references.

The test subjects were instructed to focus on those regions in the video clips they consider more important from the video content point of view. Taking the highest rate as reference, video sequences were shown in pairs for every video clip and bit-rate reduction; the subjects were asked to decide what sequence was better. This task was repeated twice, the first time with the set of non-perceptually encoded sequences and the second time with the set of perceptually-encoded sequences. The results of the subjective test are summarized in terms of average probability of detecting the reduced-rate version of each pair of encoded sequences. Given that the original high-rate encoded versions of the sequences are perceptually indistinguishable, the comparison of detection probabilities, which measure the sensitivity to distortion, able us to establish also a comparison between the performance of both encoders and, therefore, can be employed in this context as a subjective quality score.

As can be seen in Figure 10, for higher rates, subjective score (probability of detecting the low bit-rate version) is very similar for both encoder versions, since the distortion is barely noticeable. On the other hand, for lower rates, the use of our proposed saliency map embedded in a perceptual encoder improves the subjective score of the encoded sequences, making more difficult the detection of bit-rate reductions with respect to the basic non-perceptual coding model. Nonetheless, if the bit-rate decreases significantly, the distortion detection is easier regardless of the encoding strategy.

IV. CONCLUSSIONS AND FURTHER WORK

In this paper, a low-complexity motion-based saliency classifier has been proposed for its use in a perceptual video encoder. The algorithm aims to detect those regions of the image that are susceptible of attracting the attention of a potential viewer due to their motion characteristics. The complexity of the algorithm is kept low by means of the use of HME combined with specific solutions for homogeneous region management and an efficient algorithm for camera motion estimation. The results show that the saliency characterization properly responds to the motion as perceived by the observer. Furthermore, a preliminary version of a perceptual encoder guided by the proposed saliency map has been successfully evaluated by mean of a subjective test.

Although the algorithm has been designed with low-complexity constraints, additional techniques could be explored in order to further reduce its computational cost.

It is also worth mentioning that detecting motion is not the only way of producing saliency maps and our proposal could be combined with several other existing methods (see for example [3]) for providing indications of the locations of the regions where bits can be spared with minimum visual impact.

The exploration of alternative methods for this attention-guided bit allocation is also a promising line of research that could provide further improvements in the subjective scores.

REFERENCES

[1] Z. Li, S. Qin and L. Itti, “Visual attention guided bit allocation in video compresion”, Image and Vision Computing. vol. 29, pp. 1-14, 2010.

[2] A. Bhat, I.E. Richardson, and L.J. Muir, “Perceptually Optimised Variable Bilateral Filter for Low Bit-rate Video Coding,” PGNet Conference, Liverpool, 27-28 June 2007.

[3] Ç. Dikici and H. Bozma, “Attention-based video streaming,” Signal Processing : Image Communication, vol. 25 (10), pp. 745-760. 2010.

[4] C.-W. Tang, C.-H. Chen, Y.-H. Yu; C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *Multimedia, IEEE Transactions on* , vol.8, no.1, pp. 11- 18, Feb. 2006.

[5] J. Culham, S. He, S. Dukelow, F.A.J. Verstraten, “Visual motion and the human brain: what has neuroimaging told us?,” Acta Psychologica, vol. 107, Issues 1-3, pp. 69-94, April 2001.

[6] Y.-F. Ma and H.-J. Zhang, “A model of motion attention for video skimming,” in *Proc. ICIP*, vol. 1, Sept. 2002, pp. I-129–I-132.

[7] D.J. Fleet and Y. Weiss. "Optical Flow Estimation". In Paragios et al.. Handbook of Mathematical Models in Computer Vision. Springer.2006.

[8] C.R. del Blanco, F. Jaureguizar, L. Salgado, and N. García, “Target detection through robust motion segmentation and tracking restrictions in aerial flir images,” *IEEE International Conference on Image Processing, ICIP 2007*. vol. 5, pp. V-445-V-448, 2007.

[9] C.V. Stewart, “Robust parameter estimation in computer vision”, *SIAM Reviews*, vol. 41, no.3, pp. 513-537, 1999.

[10] ITU-T Recommendation P-910, “Subjective video quality assessment methods for multimedia applications,” Tech. Rep., International Telecommunication Union, April 2008.