A Scheme for Attentional Video Compression

Rupesh Gupta and Santanu Chaudhury

Dept. of EE, Indian Institute of Technology Delhi, New Delhi, India
rupesh.iitdelhi@gmail.com, santanuc@ee.iitd.ac.in

Abstract. In this paper an improved, macroblock (MB) level, visual saliency algorithm, aimed at video compression, is presented. A Relevance Vector Machine (RVM) is trained over 3 dimensional feature vectors, pertaining to global, local and rarity measures of conspicuity, to yield probabalistic values which form the saliency map. These saliency values are used for non-uniform bit-allocation over video frames. A video compression architecture for propagation of saliency values, saving tremendous amount of computation, is also proposed.

1 Introduction

The acuity of the human eye is limited to only 1-2° of visual angle. This means that when viewed from a recommended distance of 1.2 m, the eye can crisply perceive only a 2 cm radial region (computed as $1.2\times\tan(2°/2)$) on a standard definition 32 inch LCD. Also, a recent eye-tracking study [1] on inter-observer saliency variations in task-free viewing of natural images has concluded that images known to have salient regions generate highly correlated saliency maps for different viewers. However, correctly estimating the points of human eye fixation still remains a challenge. Itti et. al. [2] model visual attention as a combination of low level features pertaining to the degree of dissimilarity between a region and its surroundings. Novel center-surround approaches like [3] model saliency as the fraction of dissimilar pixels in concentric annular regions around each pixel. Hou et. al. [4] take a completely different approach, suppressing the response to frequently occurring features while capturing deviances. Other transform domain approaches like [5,6] follow a similar line of thought. Although these approaches work on psychological patterns with high accuracy, they often fail to detect salient objects in real life images. Some failure cases of these approaches will be shown in our comparison results in Fig. 2.

The failure of these approaches can be attributed to Gestaltś grouping principle which concerns the effect produced when the collective presence of a set of elements becomes more meaningful than their presence as separate elements. Thus, we model saliency as a combination of low level, as well as high level features which become important at the higher-level visual cortex. Many authors like [7] resort to a linear combination of features such as contrast, skin color, etc., but do not provide any explanation for the weights chosen. Hence, we propose a learning based feature integration algorithm where we train an RVM with 3 dimensional feature vectors to output probabalistic saliency values.

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 458–465, 2011.

One of the earliest automated (as opposed to gaze contingent), visual saliency based, video compression model was proposed by Itti in [8]. In [8] a small number of virtual foveas attempt to track the salient objects, over the video frames; and the non-salient regions are Gaussian blurred to achieve compression. Guo et. al. [5] use their PQFT approach for proto-object detection, and apply a multiresolution wavelet domain foveation filter suppressing coefficients corresponding to background. Selective blurring can however lead to unpleasing artifacts and generally scores low on subjective evaluation. A novel bit allocation model, achieving compression while preserving visual quality is presented in [9] which we adopt here. In all these compression approaches, the saliency map is computed for each frame which is avoidable considering the inherent temporal redundancy in videos. We propose here a video coding architecture, incorporating visual saliency propagation, to save on a large amount of saliency computation, and hence time. This architecture is most effective for natural video sequences.

The rest of this paper is organized as follows. In Sect. 2, we describe the steps for computing the saliency map. Since all video coding operations are MB based, we learn saliency at MB level to save on unnecessary computation. Section 3 describes a video coding architecture in which various issues relating to saliency propagation/ re-calculation and bit allocation are addressed. We conclude with some conclusions and directions for future research in Sect. 4

2 Generation of Saliency Map

We use color spatial variance, center-surround multi scale ratio of dissimilarity and pulse DCT to construct 3 feature maps. Then, a soft, learning based approach is used to arrive at the final saliency map.

2.1 Global Conspicuity: Color Spatial Variance

The lesser a particular color is globally present in a frame, the more it is likely to catch the viewerś attention. However, a color sparsely distributed over the entire frame need not be conspicuous owing to Gestaltś principles. Hence, spatial variance of colors can be employed as a measure of global conspicuity. We follow the method given in [10], based on representation of color clusters by Gaussian mixture models to calculate their spatial variance, to get this feature map. The feature map is normalized to the range [0,1]

2.2 Local Conspicuity: Multi-scale Ratio of Dissimilarity

The ‘pop-out’ effect has, since long [2], been attributed to the degree of dissimilarity between a stimulus and its surroundings. A simple method to accurately capture local saliency has been recently proposed in [3]. In this method, a multiscale filter is designed to simulate the visual field. A summation of the fraction of dissimilar pixels in concentric ring-like regions around each pixel gives a measure of conspicuity. We use this method to construct our second feature map.

However, this approach is slow, since a large number of computations and comparisons are carried out for every pixel. Noting that background pixels generally have very low values of saliency, computation of saliency for these pixels is superfluous. Hence, we first run a SIFT algorithm and locate the keypoints in the image, which are salient not only spatially but also across different scales. We take one keypoint at a time and compute its saliency using [3]. If the saliency of this point is above a threshold (0.4 here, required since a keypoint may lie on a cluttered background), we start growing a region from that point. The saliency value of neighboring pixels is used as region membership criterion and all pixels visited are marked so that they are not re-visited when a different seed point is chosen. We stop when the distance between the new pixel and region mean exceeds a threshold (0.2 here). This feature map is also normalized to [0,1].

2.3 Rarity Conspicuity: Pulse Discrete Cosine Transform

A biologically plausible, real time model simulating lateral inhibition in the receptive field has been proposed in [6]. It has also been shown to outperform other transform domain approaches like [5] both in terms of speed as well as accuracy over psychological patterns. We apply the pulse DCT algorithm to smoothened images to produce our rarity feature map. A Gaussian blurred image simulates the scene viewed from a distance and thus finer edge details in a cluttered background are not noticed, leading to a sparser feature map. We normalize it to the range [0,1].

2.4 Learning to Integrate the Feature Maps

The steps followed for combining the 3 feature maps are as follows. First, we selected 30 images, of size 300×400, encompassing the failure cases of each of the 3 feature maps. 5 viewers were asked to mark each part of the image they considered salient. In accordance with [1], our images (mostly taken from [10]) had well-defined salient regions and hence the markings turned out to be exactly the same for almost all images. Then, an MB level, 3 dimensional training data (total 450×30 points) was prepared taking average values of each of the 3 feature maps over each MB of size 16×16. A target class label '1' was assigned to an MB if more than half of the pixels of that MB were marked salient; else class label '0' was assigned. Next, we trained an RVM over this training data as a binary classification problem. Here we must point out that we are not really interested in a binary label (salient/non-salient) but the relative saliency value of each MB which will later be used for bit allocation. A potential advantage of RVM over SVM, which is desired here, is that it provides posterior probabilities. Also, RVM has better generalization ability and its sparser kernel function leads to faster decisions. The probabilistic outputs of the RVM formed our final saliency map.

To test the machine, we generated a testing data from 120 images (450×120 points) and evaluated the saliency maps obtained against ground truth. Various authors like Bruce et. al. [11] have used area under the ROC curves to quantify the quality of their algorithms. The ROC curve obtained on our own ground truth data is shown in Fig. 1. Also shown in the same figure is a comparison of our result with another leading graph based visual saliency approach [12], which has been shown to outperform various other approaches like [2]. We obtained a 0.90048 (s.e. 0.00136) area under the curve compared to 0.87009 (s.e. 00161) for [12]. In the context of application of saliency to video compression, an FN (actually salient but classified non-salient) is costlier compared to an FP. A very low FN rate, less than 2%, at the cut-off point reflects the potential of our algorithm for such applications. Some results and comparisons with [12] and [11] are shown in Fig. 2. A comparison with [3] and [6] is inherent in these results as our local and rarity feature maps respectively. It is apparent that our approach is better or at least at par with these other high-ranking approaches.

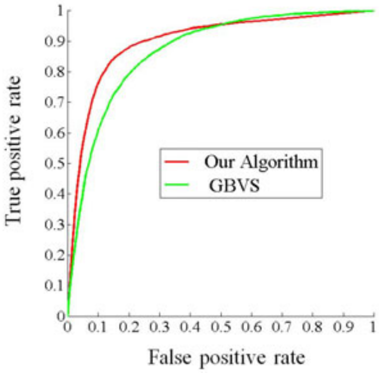

Fig. 1. ROC curves for our approach and [12] obtained by varying thresholds on saliency values

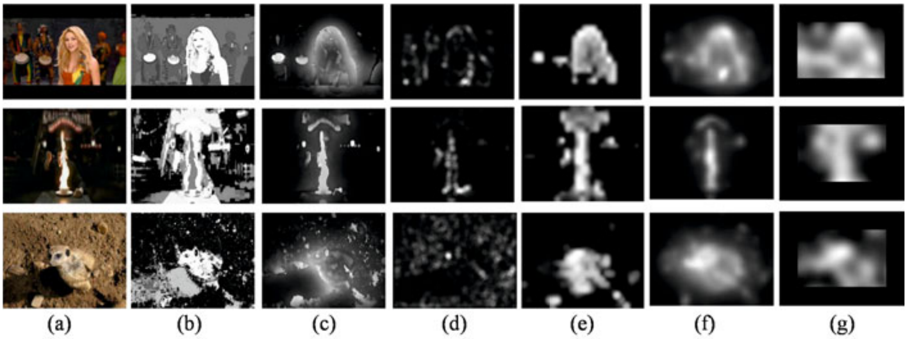

Fig. 2. (a) Input image, (b) global, (c) local [3], (d) rarity [6] feature maps, (e) our resized saliency map, (f) saliency map obtained from [12] and (g) [11]

3 Video Compression Architecture

We wish to employ saliency for the purpose of video compression. However, computation of feature maps for each video frame can prove to be computationally very expensive if we rely on techniques such as those proposed in [5,8,9] as they necessitate calculation of saliency map of each frame. We propose here the use of temporal redundancy inherent in videos to propagate saliency values. Ideally the saliency map should be re-calculated only when there is a large change in saliency. However, to measure this change, we require the saliency for the next frame which is unavailable. Hence, we also propose a workaround to detect the frames for which re-computation of saliency map is indispensable. A block diagram of the architecture is shown in Fig. 3 which is discussed in detail in the following subsections

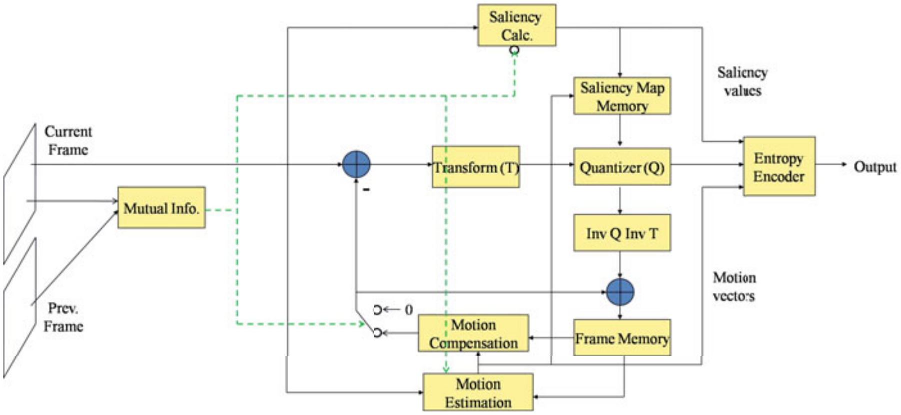

Fig. 3. Our video compression architecture incorporating saliency propagation

3.1 Propagation of Saliency Values

Firstly, we describe the need for the mutual information (MI) computation unit. The idea is that we perform a re-calculation of saliency map on the basis of MI between successive frames. A concise information theoretic shot detection algorithm has been proposed by Cernekova et. al. in [13] and an improved version of the same using motion prediction in [14]. The authors compute the MI between consecutive frames and argue that a small value of MI indicates existence of a cut. We experimented with this method over some video sequences, with saliency map of each frame pre-computed, and plotted the MI distributions for color as well as saliency. MI for an Airtel ad sequence with 9 scene changes is plotted in Fig. 4. It is apparent that not only does this method effectively capture changes in saliency as shown in Fig. 4(a), but also, that the RGB and saliency plots follow a very similar distribution (Fig. 4(b)). Figure 4(b) implies that we can detect the frames requiring re-computation of saliency maps by calculating MI over the color channels. The frame where a large change is detected should be coded as an I frame (or I MBs in H.264) and saliency re-computed for this frame and stored. The method has been found to work best on natural video sequences.

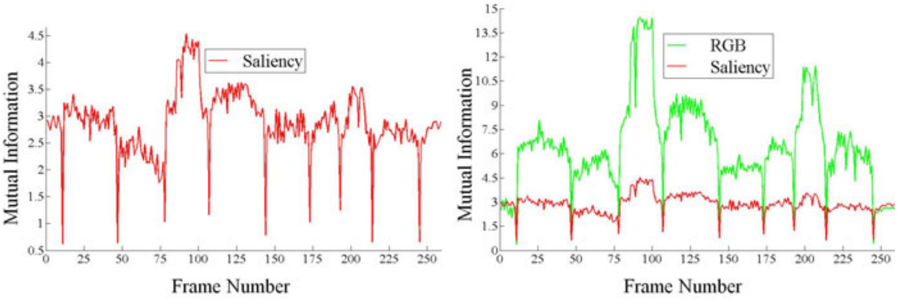

Fig. 4. (a) MI plot for saliency maps, (b) MI plots of RGB and saliency overlaid. An Airtel ad sequence with 9 cuts is used here.

For P frames, we make use of motion vectors to approximate saliency values. We select an MB in the current frame and look for the best match in the reference frame. This best match may or may not exactly overlap an MB in the reference frame, but we have the saliency values for only non overlapping 16×16 MBs. Therefore, we take a weighted average of the saliency values of each of the MBs under the best match region the in reference frame, as the saliency value for the MB in current frame. The weights correspond to the amount of area overlap as shown in Fig. 5

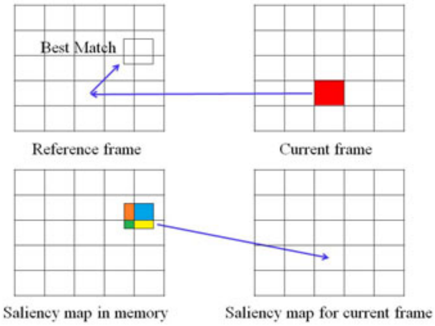

Fig. 5. Image illustrating a weighted averaging of saliency values, the orange, blue, yellow, green colors denote the amount of overlap and hence weights

3.2 Selection of Quantization Parameters

Once the saliency map is obtained, bits may be non-uniformly distributed across a frame. We require a function which can optimally tune the quantization parameters of salient and non-salient MBs to achieve compression, i.e, reduce rate (R), without any significant loss of perceptual quality, i.e, constant distortion (D). In [9], this is posed as a global optimization problem and solved using the method of Lagrange multipliers. The final result for quantization step Q_{istep} for the i^{th} MB having a saliency value w_i is given as:

$$Q_{istep} = \frac{Ws}{w_i S} Q_{step} \quad . \tag{1}$$

where W is the sum of saliency values over all MBs, s is the area of MB_i (16×16 here), S is the area of entire frame and Q_{step} is a fixed value depending on the amount of distortion tolerable. This formula implies that the quantization step size should be inversely proportional to the saliency value which is completely justified. We present here a short verification of how this formulation achieves compression without compromising on perceptual quality. Assuming a R-D function [15] for an MB_i is given by:

$$D_i = \sigma_i^2 e^{-\gamma R_i} \ or \ R_i = \frac{1}{\gamma} log \left(\frac{\sigma_i^2}{D_i} \right) \quad . \tag{2}$$

where σ_i^2 is variance of encoding signal and γ is a constant coefficient. Ignoring the constant term γ and taking $\sigma_i^2 = 1/\alpha$ we get:

$$R_i = log \left(\frac{1}{\alpha D_i} \right) \quad . \tag{3}$$

Now, the average rate R is calculated as $\sum_{i=1}^{N} sR_i/S$, where N is the number of MBs. Noting that $D_i \propto Q_{istep}$, we get after replacing Q_{istep} by (1):

$$R = \frac{Ns}{S} \left[log \left(\frac{1}{\alpha Q_{step}} \right) + log \left(\frac{(w_1.w_2...w_N)^{\frac{1}{N}}}{w_1 + w_2 + ... + w_N} \right) + log \left(\frac{S}{s} \right) \right] \quad . \tag{4}$$

From the above equation it is clear that the first term denotes the rate if every MB was quantized with the same parameter Q_{step}, the second term is always ≤ 0 by the AM-GM inequality and the third term is a constant. Thus R is reduced. It can also be readily observed from (1) that overall D ($\sum w_i D_i / W$) remains constant. We limit the Q_{istep} to minimum and maximum values of max($0.5 \times Q_{step}$, Q_{istep}) and min($1.5 \times Q_{step}$, Q_{istep}) respectively. Also, we smoothen our saliency map using a Gaussian filter before computing the quantization step. This serves two purposes, firstly, it ensures that the salient object/region is covered completely and secondly, it ensures a smooth transition from salient to non-salient regions.

4 Conclusion

A vast amount of research has gone into modelling of the human visual system with each model having its own merits and shortcomings. The potential which lies in an integration of these models has been demonstrated by the accuracy of our results. A simple and effective learning based approach for such a unification has been presented. Though we make use of only 3 features, this model is easily extendible to more features if desired. We computed saliency at MB level to save computation, however our model is equally applicable at pixel level. The compression framework proposed, to approximate saliency of P frames, can save a lot of computation, speeding-up compression. We plan to integrate our it into the H.264 coding system which remains a challenge owing to the complex mode decision metrics and hybrid coding structures in this standard [16].

References

1. Engelke, U., Maeder, A., Zepernick, H.J.: Analysing Inter-observer Saliency Variations in Task-Free Viewing of Natural Images. In: ICIP, pp. 1085–1088 (2010)
2. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. PAMI 20(11), 1254–1259 (1998)
3. Huang, R., Sang, N., Liu, L., Tang, Q.: Saliency Based on Multi-scale Ratio of Dissimilarity. In: ICPR, pp. 13–16 (2010)
4. Hou, X., Zhang, L.: Saliency Detection: A Spectral Residual Approach. In: CVPR, pp. 1–8 (2007)
5. Guo, C., Zhang, L.: A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. IEEE Trans. Image Proc. 19(1), 185–198 (2010)
6. Yu, Y., Wang, B., Zhang, L.: Pulse Discrete Cosine Transform for Saliency-Based Visual Attention. In: ICDL, pp. 1–6 (2009)
7. Chiang, J., Hsieh, C., Chang, G., Jou, F., Lie, W.: Region-of-Interest Based Rate Control Scheme with Flexible Quality on Demand. In: ICME, pp. 238–242 (2010)
8. Itti, L.: Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. IEEE Trans. Image Proc. 13(10), 1304–1318 (2004)
9. Li, Z., Qin, S., Itti, L.: Visual Attention Guided Bit Allocation in Video Compression. Image and Vision Computing 29(1), 1–14 (2011)
10. Liu, T., Sun, J., Zheng, N.-N., Tang, X., Shum, H.-Y.: Learning to Detect a Salient Object. In: CVPR, pp. 1–8 (2007)
11. Bruce, N.D.B., Tsotsos, J.K.: Saliency Based on Information Maximization. In: NIPS, pp. 155–162 (2006)
12. Harel, J., Koch, C., Perona, P.: Graph-Based Visual Saliency. In: NIPS, pp. 545–552 (2006)
13. Cernekova, Z., Pitas, I., Nikou, C.: Information Theory-Based Shot Cut/Fade Detection and Video Summarization. IEEE Trans. CSVT 16(1), 82–91 (2006)
14. Krulikovska, L., Pavlovic, J., Polec, J., Cernekova, Z.: Abrupt Cut Detection Based on Mutual Information and Motion Prediction. In: ELMAR, pp. 89–92 (2010)
15. Bhaskaran, V., Konstantinides, K.: Image and Video Compression Standards: Algorithms and Architectures. Springer, Heidelberg (1997)
16. Chen, Z., Lin, W., Ngan, K.N.: Perceptual Video Coding: Challenges and Approaches. In: ICME, pp. 784–789 (2010)