Visual attention based surveillance videos compression

Fahad Fazal Elahi Guraya, Victor Medina, Faouzi Alaya Cheikh; Faculty of Computer Science and Media Technology, Gjovik University College; Gjovik, Norway

Abstract

Visual attention models (VAM) try to mimic the human visual system in distinguishing salient regions from non-salient ones in the scene. Only a few attention models propose to detect salient motion in surveillance videos. These model utilizes static features such as color, intensity, orientation, face, and dynamic features such as motion to detect most salient regions in videos. This motivated us to propose a compression algorithm based on visual attention model that is developed specificly for surveillance videos. In this paper we are using a state of the art visual attention model developed by combining bottom-up, top-down, and motion cues. Based on its similarity with experimentally obtained gaze maps evaluated both visually and with quantitative measures, a compression model based on this attention model is proposed for H.264/AVC encoded videos. Our experimental results show that we can encode videos with same or better quality than those obtained with the standard baseline profile of the JM 18.0 reference encoder, while reducing the file size uptil 22%.

Introduction

Human visual system is attracted by salient objects or events. This is done unconsciously and effortlessly in the visual system when light passes through retina cells to the complex cells of the primary visual cortex. The retina cells distribute the light signal into two main outputs [1, 2], thanks to two different kinds of ganglion cells : magnocellular and parvocellular ganglion cells. Magnocellular ganglion outputs provide global information of the scene that contain lower bands of spatial frequencies. While parvocellular ganglion outputs provides detailed information of the scene, or high frequency components of the scene. The two signals then pass through cortical-like filter that decomposes these signals into elementary features by a bank of filters. The cortical-like filters give dynamic and static information of the scene in their output, that could be named as static or dynamic saliency maps. The ganglion cells output enhances contrast that attracts human gaze [3]. It is a rather challenging task to model such a complex phenomenon of human vision. Such computational models can however be used in many image and video processing applications such as compression, event detection, perceptual quality evaluation, etc. Most of the time due to high compression ratio, it is impossible to recognize people faces in surveillance videos [4]. VAM can be used as Region of interest (ROI) for compression algorithms which help increase the quality of salient regions of surveillance videos.

A region of an image or scene become salient when low level features such as color, texture, binocular disparity, intensity, orientation or motion etc. differ significantly from its variation in its neighborhood. Saliency is a unit that helps to determine the capability of attracting visual attention towards some region of an image [5, 6]. There are many factors involved in identifying these salient regions in a visual scene. These factors or visual cues are generally categorized into two groups, bottom up, and top down visual cues [7, 8]. The bottom-up stage of human visual system processes the input scene/image in parallel and pre-attentive manner and forward this information to a serial, attentive and computationally intensive top-down stage. In bottom-up approaches our visual system computes the salient regions from low-level features such as color, intensity, orientation, etc. It is evidently proved that human visual system combines low-level features in the early stage [6, 9]. Saliency computation models based on information theory have successfully model human attention from these local features [10, 11]. A famous computational model of bottom-up attention proposed by Itti and Koch [12], uses low level features such as color, intensity and orientation. Top-down approaches involve more complex visual activity such as object detection, face detection, etc. It is performed very fast and efficiently in human visual system. Combining bottom-up and top-down approaches guides the visual system towards the salient regions or regions of interest [13, 14, 15] in a visual scene. It is observed that the human visual system diverts the attention to faces 16.6 times more than other similar regions [16]. Therefore, face detection can significantly improve the short-comings of static saliency models such as Itti's saliency model [12], GBVS [17] and GAFE [18]. In [19], face detection as top-down visual cue is combined with Itti and Koch bottom up saliency computational model [12] which gave promising results. Bottom-up and top-down approaches can help us make a model which can detect salient regions in an image, but what about detecting saliency in videos? Videos have an extra dimension, which creates, a perceptual feeling of motion in human brain. Motion has great influence on identifying the salient regions in a complex dynamic visual scene. Many models have been introduced in the literature to detect salient motion [20, 21, 22, 23]. The dynamic visual selective attention approach proposed by [24] is not good in detecting moving regions as salient regions. An improved version of this model is proposed by [25], the improved model uses the information of each frame to obtain a dynamic saliency map.

Most saliency computational models [26, 27] are inspired by feature integration theory [28] also called late fusion of features in [12]. These models extract low-level features and integrate them to get salient regions. The question that can be raised on these models is how to combine low-level and high-level features to mimic human visual system? In [29], authors have proposed to use the neural networks for combining the visual cues or features such as color, intensity, orientation, faces and salient motion into a saliency model. We are using visual attanetion model proposed in [29] for compression of our surveillance videos.

Video compression standards like MPEG-2, MPEG-4 and H.264/AVC make use of a rate control algorithm to control the size of the encoded videos, so that they meet the bandwidth requirement imposed by each application. This mechanism ensures that it is possible to feed data to the applications in a fluent manner. Several methods have been proposed to control the bitrate of the encoded videos while maximizing their quality. In this model, we propose to use information obtained from visual attention model, i.e. saliency maps to increase the quality of those areas that are more salient in the scene and reduce the quality of less salient areas.

Normally, all the pixels are allocated the same amount of resources by the encoder, regardless of their importance in the scene. However, if we consider that some parts of the video might be less important to a human observer than others, there is no reason why the same amount of resources should be allocated to all the pixels in the scene. For example, an observer might unconsciuosly be more attracted towards areas such as human faces, or pay more attention to familiar objects that might be of interest in specific situations – like, for instance, a suspicious backpack in a surveillance video, or the main character's face in a feature film.In this model, we take those observations into account by assuming that the more salient pixels are those which viewers are more likely to look at, thus encoding them with a higher amount of bits – which, in turn, results in a higher quality. At the same time, we want to maintain the mean average difference (MAD) with the original video at frame level, which means that we must also lower the quality of less salient pixels.

In this paper we have reimplemented the model described in [29], improving the results by maintaining – or even reducing – the bitrate of the resulting videos, while maintaining or increasing their quality.

The rest of paper is organized as follows, in the next section we will describe the visual attention model that we have used for video compression. In later section the computation of the quantization parameter in the compression model is presented. Then we describe the compression experiment results. Finally we conclude the paper and point to some future directions.

Visual attention model

This section explains the visual attention model [29] used in this paper. This VAM includes top-down, bottom-up visual cues and salient motion information. Bottom-up visual features such as color, intensity and orientation are used in our proposed model. The method of computing bottom-up visual cues in [12] is adapted in this model. Apart from bottom-up visual cues, top-down visual cue i.e. face is also incorporated in this model. Faces have significant importance in surveillance videos and attract human visual attention [16] much more than other objects or regions. A similar model that incorporates bottom-up and top-down visual cues for images has been presented in [19]. The VAM proposed in [29] also include salient motion in addition to bottom-up and top-down visual cues, and the method of salient motion is inspired by [20].

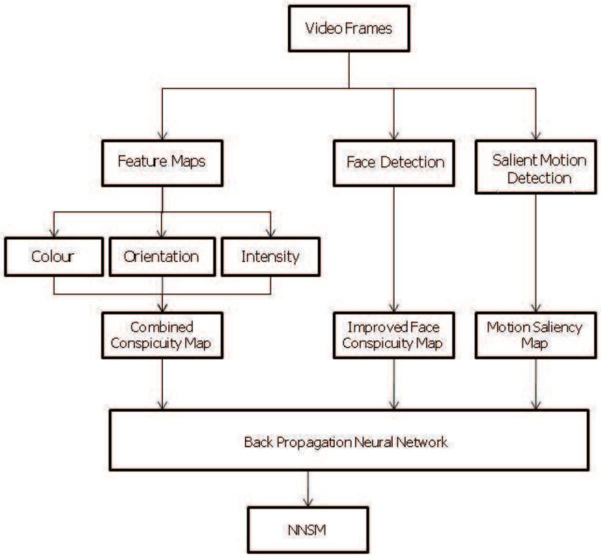

Figure 1: Visual attention model.

Several static or stationary saliency models have been proposed in the literature, the famous ones include Itti and koch [12], GAFE [18], GBVS [17]. But the most popular stationary saliency model is the one proposed by Itti and Koch [12]. This model generates the saliency maps based on the combination of color, orientation and intensity conspicuity maps(color C_c, intensity C_i, and orientation C_o). Itti's saliency model computes the saliency map by averaging the three conspicuity maps :

$$SM_{itti} = \frac{1}{3}(C_i + C_c + C_o) \qquad (1)$$

Apart from low-level saliency features, experiments show that high level features such as faces attract more attention than the other low-level features [16, 30]. Itti's saliency model is based on low-level features and does not consider high-level features, that is why it does not perform well for complex scenes that have faces, or other objects. To overcome this problem, a model that incorporates high level visual cues such as human faces is needed for surveillance videos. There is not so much research done on high level feature's use in saliency models. A saliency model proposed in [19] uses color, intensity, orientation and face features. This model gives 33% improvement in the saliency maps for the images with faces in them. The authors in [19] have used the face detection model by Walther et al [31]. They have proposed that face conspicuity map C_{Face} are important and their experiment shows that they should be given four times higher weightage than the low-level features in the combination step. Equation 2 describes the resultant saliency model.

$$SM_{Sharma} = \frac{1}{7}(C_i + C_c + C_o + 4C_{Face}) \qquad (2)$$

The model shown above provides an overall 33% performance improvement over other stationary models [19]. In this paper we have used the attention model that use face features, low-level features and motion feature together as shown in figure 1. As you can see in the figure, this model also use salient motion. Salient motion is the motion that grabs or attracts the attention of the viewer. Salient motion detection is a complex phenomenon that depends highly on the specific scene, environment, or scenario. It is also heavily dependent on the viewers interests and interpretation. So normal motion detection methods such as Lukas and Kanade method are not enough to detect the salient motion. Because if we use temporal difference of adjacent frames or compute the motion vectors from one frame to another, we might be able to find the moving regions of the image. But it cannot distinguish between regions with salient motion and regions with non-salient motion. For salient motion detection the non-salient motion has to be filtered out or ignored. Therefore in addition to motion detection models we also need a filter that can filter out non-salient motion. Literature has much research work done on salient motion detection [20, 32, 33, 21]. A Motion saliency map using spatio-temporal energy accumulation of coherent moving objects by Gabor filtering is proposed by [21]. Another salient motion detection algorithm proposed by [32] is using the motion vectors magnitude and phase histograms. These histograms are later combined by a proposed formula in such a way that the motion entropy of the salient regions increase.

The authors in [20] proposed a method based on temporal differencing, filtering and segmentation. [29] has proposed salient motion model, that uses Gaussian filter instead of segmentation as the last step of salient motion detection model, to make salient motion detection more robust for real time videos.

Combining different conspicuity maps or visual cues such as color, intensity, orientation, face and motion into one saliency map is a challenging task. It is vital to consider HVS perception characteristics during the combination phase. We need to mimic the neurons between retina and the visual cortex, and the best way to do it, is through learning from known gaze maps and neural networks. The method proposed in [29] uses neural network to combine the visual cues. The neural networks can be trained by using the provided input-output datasets. In training phase, inputs will be visual cues, low-level visual features and gaze maps obtained from psychophysical experiment for a given set of surveillance videos are used as output. In this way, neural networks can learn and mimic exactly the behavior performed by the neurons in the HVS. Every neural network is based on several neuron layers. One neuron can be considered as a summing device that has inputs and produces output. To be more precise, a neuron also has some sort of weighting mechanism. The inputs of the neuron are multiplied by the corresponding weights and the sum of these products is sent to the output. The neurons can be placed in 3 different layers of a neural network. These three layers are called input neuron layer, hidden layers, and output layer. Each neural network has to be trained before being utilized for actual task. During training neurons get the inputs and corresponding outputs so that it can learn and configure the weights of each neuron in the neural network. In the next paragraph, neuron functionality is explained in greater detail.

The basic neuron consists of an activation function F(weightedSum T) where *weightedSum* is weighted sum of the inputs and T is the threshold as shown in eq.3. The weights are initialized to random values and get updated in the training phase.

$$weightedSum = \sum_{i=1}^{n} weight_i * input_i \tag{3}$$

Various functions can be used as activation function F. Sigmoid function is used as activation function in this paper. The sigmoid activation function is shown in eq.4 and eq.5.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

$$\frac{d\sigma(x)}{dx} = \sigma(x) * (1 - \sigma(x)) \tag{5}$$

The most basic one is Back Propagation Neural Network (BPNN). The property of BPNN is that it back propagates the output through the neural network to update the internal weights of each neuron. BPNN learns in such a way that any mistakes or errors made during the training phase is sent backwards through the network to correct the weights. This process is called backward propagation of errors. The back propagation requires the activation function used by the artificial neurons or network nodes to be differentiable. This learning method is called supervised learning method, and is a generalization of the delta rule. The learning phase of BPNN is divided into two phases: propagation, and weight update phase. The propagation phase involves 2 steps:

1. Forward propagation of training datasets through the FFNN, to generate the neurons output activation.

2. Back propagation of the neurons output activation through the network with training pattern's target to generate the deltas of all outputs and hidden neurons.

The weight updation phase include 2 steps.

1. Get the error gradient of the weights by multiplying the output delta and activation. The error gradient can be computed by eq.6.

$$\delta_k = y_k(1 - y_k)(d_k - y_k) \tag{6}$$

   where y_k is the value at output neuron k and d_k is the desired value at output neuron k. Error gradients at output and at hidden layers are different. The hidden layers error gradient is based on the output layers error gradient due to back propagation. The error gradient for each hidden neuron is the gradient of the activation function multiplied by the weighted sum of the errors at the output layer as shown in eq.7.

$$\delta_j = y_j(1 - y_j) \sum_{k=1}^{n} w_{jk}\ \delta_k \tag{7}$$

2. Subtract the ratio of the gradient from the weight, that is called learning rate that influences the speed and quality of learning. The sign of the weights gradient represent the direction where the error is increasing. The weight updation equations are shown in eq.8,9,10,11.

$$w_{ij} = w_{ij} * \Delta w_{ij} \tag{8}$$

$$w_{jk} = w_{jk} * \Delta w_{jk} \tag{9}$$

   where

$$\Delta\, w_{ij}(t) = \alpha.inputNeuron_i.\delta_j \tag{10}$$

$$\Delta\, w_{jk}(t) = \alpha.hiddenNeuron_j.\delta_k \tag{11}$$

where α is learning rate and δ is error gradient.

Repeat the propagation and weight updation phase until the neural network converges. There are other internal factors like number of hidden neurons, number of iteration of the datasets, learning rate and momentum that can help the network to converge. For right combination of these values, alot of testing is required.

Computation of the quantization parameter

We propose an alternative model to compute the QP. Our method computes a new QP for every macroblock in each frame – as opposed to the method used in H.264, which computes a new QP for every frame – from the saliency values obtained from the corresponding saliency maps. The number of bits allocated for a given frame in the standard H.264/AVC rate control algorithm is computed based on a Mean Average Difference (MAD) criteria. The bitrate is related with the MAD and QP according to the following formula [34]:

$$T_i = c_1 \frac{MAD_i}{Qstep_i} + c_2 \frac{MAD_i}{Qstep_i{}^2} - h_i \tag{12}$$

where T is the assigned bitrate for the basic unit, Qstep is the quantizer step size (from which QP is computed), h_i corresponds to the number of bits due to overhead data, and c_1 and c_2 are model coefficients.

Depending on the configuration used, the bitrate can be fixed in advance – in which case the encoder gives preference to the bitrate over the quality, adjusting the QP dinamically to maintain a constant MAD – or variable – in which case the quantization is fixed in advance, having preference over the bitrate. We propose a hybrid approach, where the QP changes dynamically for every macroblock, but aiming for a target bitrate by keeping the average QP constant at frame-level. For each macroblock, the corresponding saliency values are obtained from the saliency map of the frame it belongs to. The saliency of all the pixels in the macroblock is then averaged. Once the average saliency has been computed for all the macroblocks, we can compute how much each macroblock deviates from the average saliency, and assign the corresponding QP accordingly so that, in the average, the QP is still constant for the entire frame. The QP is calculated as follows:

$$QP_i = \lceil QP_f - (s_i - \bar{s}) \cdot W \rceil \tag{13}$$

where QP_i is the QP for macroblock i, QP_f is the target QP at frame level, s_i is the average saliency of all the pixels in macroblock i, $\bar{s}$ is the average saliency of all the macroblocks in the frame, and W is a weighting factor. Let us note as D_i the absolute deviation of the saliency of macroblock i from the average, so that $D_i = (s_i - \bar{s})$. Then we compute the weighting factor W according to the following rule:

$$W = \begin{cases} -D_i \cdot \min\{QP_f - QP_{min}, QP_{max}/3\} \; where \; D_i > 0 \\ D_i \cdot \min\{QP_{max} - QP_f, QP_{max}/3\} \; where \; D_i \leq 0 \end{cases} \tag{14}$$

where QP_{min} and QP_{max} are defined by JM as 0 and 51 respectively. We must take into account the amount of overhead introduced by changing the QP very often. The Delta QP (DQP) is the difference between the QP assigned to the same macroblock in two consecutive frames. This means that, as the range of possible values for QP increases, so does the DQP and, in turn, the overhead data. Indeed, if QP can take values within a high interval, small changes in the saliency might translate into big changes in the QP, therefore increasing the DQP; this is especially true in the case of scenes with a lot of motion where the pixels change very rapidly from one frame to the next. In our model, we use the weighting factor W to constrain the QP range within a certain limit, so that the quantization of a given macroblock will always be, at most, in the interval $QP_f \pm QP_{max}/3$. After testing with several values, this range proved to be the best compromise between quality improvement and amoun of overhead data.

The amount of salient pixels in a frame tend to be much smaller than then amount of non-salient pixels; in fact, most pixels normally have a zero saliency, whereas only a small amount of them will be over zero, which means that the average saliency in a frame is typically very low –nearly zero. From equation 13 we can observe that the amount of units that the QP is increased or reduced for each macroblock is directly dependant on the deviation from the average D_i of the saliency of the macroblock. This way, frequent low-saliency macroblocks will get assigned a QP similar to the average, whereas less common ones with a higher saliency will get assigned a much lower QP –corresponding to a much higher bitrate; with this mechanism, we make sure that the average QP is maintained.

Experimental Results

In this section we will show the experimental results obtained with our model, and compare them quantitatively and qualitatively with the results obtained with the standard H.26/AVC algorithm. We have performed two experiments. The first one is about quanitative comparison and second one is psychophysical experiment. In quantitative experiment, we have compared the two output videos using quality matrices and in pscyhophysical experiment we have obtained the results from the subjective experiment. The first experiment is quantitative experiment, for this purpose, we use a reference software which implements the standard methods in the H.264/AVC standard. The reference software that we used in this project is JM 18.0 [35]. Five different surveillance videos were tested: videos 1, 2 and 6 belong to the iLIDS dataset of the IEEE International Conference AVSS 2007, and contain images from a surveillance camera at a subway station. Video 3 shows a man picking an object from a store and leaving. Video 4 shows people passing by at a waiting room in a train station. We tested our model for different QP values, and compared the results with the ones obtained by the standard model. Firtly, 25 was chosen as the middle point between the maximum and minimum possible quantization, and then different increments were tested between 25 and 45 –at this point features start to be unrecognizable.

The JM encoder works with configuration profiles, where most encoding parameters – including the QP – are set up. For this experiment we chose the baseline profile, which uses only I-frames and P-frames. The reason to choose this profile is that our tests showed that using B-frames resulted in additional encoding time that did not compensate the improved performance.

We used several metrics to compare the results. For this model, common metrics such as PSNR or SSIM [36] are not representative because, given that the average QP at frame level is the same, these metrics also yield similar results in the average at frame level. Therefore, we chose to focus on specific areas where we know that the saliency is high, comparing the results with the standard model. Figure 2 shows how the ssim index in the frame encoded with our model is much closer to one – an ssim index of one means that the image is identical to the original – than the one encoded with the standard model. We can see that the areas containing the face and the lights in the upper left corner will be given a higher quantization than the darker areas which do not contain relevant information for the viewer.

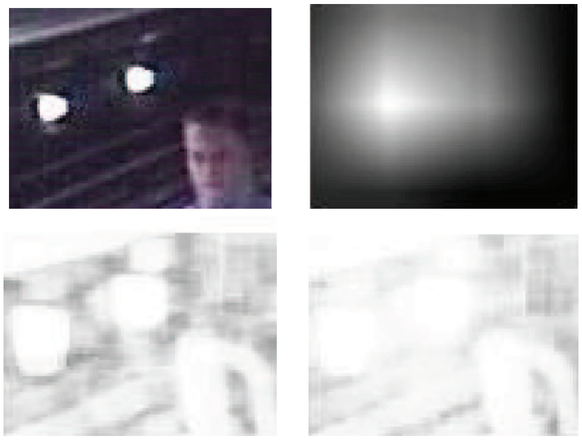

Figure 2: From left to right and up-down [a,b,c,d]: Closeup from frame 650 of video 1; saliency map for the zoomed area; SSIM disparity map of the frame obtained with the standard model; SSIM disparity map of the frame obtained with our model.

If we look at the entire frame from which the patch in figure 2.a was taken, we can see that the ssim index of the frame obtained with the standard model is 0.9497, whereas the frame obtained with our model has an index of 0.9395. Again, since some macroblocks have worse quality than others, this decreases the global ssim index. Hoewever, this is not a problem because we are only interested in the salient areas. Indeed, the ssim index corresponding to the images in figure 2.c and 2.d, respectively, is 0.8917 and 0.9413, which shows that our model clearly achieves much better quality in areas of high saliency.

Another point to compare is the file size against the visual quality. Our results have shown that using an intermediate value between the maximun and minimum possible QP – that is, 25 since, as we mentioned earlier, the minimum is set at 0 and the maximum at 51 – results in files of a simillar size to those encoded using the standard model. However, as we increase the quantization, we can also observe an increase in the amount of overhead bits which are added to the bitstream. As we explained earlier in quantiztion parameter section, this increase in the amount of overhead data is due to the fact that using higher QPs implies a higher range of QP values for salient macroblocks. This, in turn, translates into an increase in the DQP. Using a weighting factor – as shown in equation 13, – constrains the QP range, thus reducing the overhead data.

Furthermore, the QP for the chroma component is rescaled for high values [34] – above 39, – which means that the chroma components might not keep the average quantization because the amount of under-quantized bits will be higher than the amount of over-quantized bits, resulting in higher-size videos than those obtained with the standard model.

We have presented the final results for 3 videos of our dataset in figure 3, we have computed the file sizes of these videos for avearge frame QP_f of 25, 30 and 35 when compressed with or without saliency maps. The zero line in the graph show the video size when encoded JM baseline encoder withouth using the saliency maps. It can be noticed in the graph 3 that if we use QP = 35 then we always increase the filesize while improving the quality of video. But if we use lower quality improvement, i.e. QP = 30 or QP = 25, we get quality improvement while reducing the file size. These are our preliminary results, in the final paper we are planning to add the results of two more videos and the results from our psychophysical experiment.

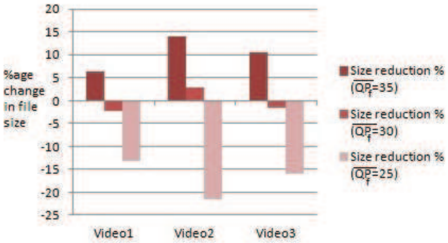

Figure 3: %age increase/decrease in saliency based compressed file size when compared with non-saliency based compressed file.

The subjective quality experiment is performed where 33 subjects participated, the participants ages varies between 20 and 50 years. Each subject has been shown 9 sets of videos, each video set contains two videos, one video with H.264 compression and the other with attention model based compression. The videos are displayed on a computer screen in normal viewing conditions and under normal luminance. The subjects were provided 3 options: Video-1's quality is better than Video-2's quality, Both are same(no quality difference), Video-2's quality is better than Video-1's quality. The score is calculated by assigning 1 when the subject choose H.264 compressed video over attentional model based compressed video, 2 for when the subject did not find any qulity difference in both videos, and 3 when the subject has choosen attentional model based compressed video as the better quality video. The mean opinion score of the experiment is shown in table 1. The graph of subjective experiment results is shown in figure 4. The mean opinion score of subjective results for most of the video is around 2.0, it shows that the attention based compressed video has same quality as the non-attention based compressed video. It also implies that we have reduced the file size of the compressed video by using the attention model.

Conclussion and future directions

In this paper, an attentional model is used for two basic reasons, one to reduce the file size without quality degradation, second to improve the video quality without increasing the filesize. The experimental results show that our model is capable of obtaining videos with the same or better quality while reducing the file size. The file size of videos obtained with the standard reference encoder JM is used as ground truth and we manage to reduce

Table 1: Subjective Experimental Results

Video set #	Mean Opinion Score
1	2.15
2	1.85
3	1.73
4	2.03
5	1.94
6	2.09
7	2.06
8	1.97
9	2.00

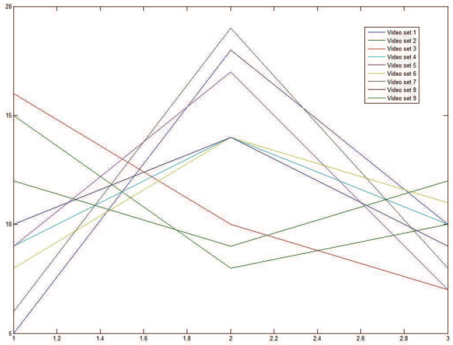

Figure 4: Subjective results graph for video set 1 to 9.

the file size upto 22%.

Our tests have been done with surveillance videos, but this model could easily be applied to any other type of videos where non-salient information is not to be paid as much attention by the observers. However, it is worth noting that, in general, the saliency maps are not available to the encoder beforehand, so a practical application of this model should integrate in the encoder the visual attention model to generate the saliency maps. Although using saliency maps at the encoding stage requires several additional operations, our tests have shown that our model only adds a small fraction of computation time – around 0.15% at most. Nonetheless, as mentioned above, the generation of the saliency maps will add an additional time, so one must pay attention to whether or not this approach could be adequate for the desired application. The psychophysical experimental results show that we have produced compressed videos using attention model with smaller size than the one compressed by H.264 encoder. However the quality of attention based compressed videos is similar or better than the one compressed using H.264 encoder.

References

[1] W. H. Beaudot, *The neural information in the vertebra retina: a melting pot of ideas for artificial vision*. PhD thesis, Tirf laboratory, Grenoble, France, 1994.

[2] S. Schwartz, *Visual Perception: A Clinical Orientation*. McGraw-Hill Medical Pub. Division, 2009.

[3] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," in *Network: Computation in Neural Systems*, pp. 341–350, 1999.

[4] P. Kovesi, "Video surveillance: Legally blind?," in *Digital Image Computing: Techniques and Applications, 2009. DICTA '09.*, pp. 204 –211, dec. 2009.

[5] E. Titchener, *Elementary Psychology of Feeling and Attention*. Ayer Co Pub. ISBN 0405051662, 1973 (original 1908).

[6] A. R. Koene and L. Zhaoping, "Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in v1," *Journal of Vision*, vol. 7, no. 7, pp. 6.1–14, 2007.

[7] L. Itti, *Models of bottom-up and top-down visual attention*. PhD thesis, California Institute of Technology, January 2000.

[8] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194–203, Mar 2001.

[9] J. Krummenacher, H. J. Muller, and D. Heller, "Visual search for dimensionally redundant pop-out targets: evidence for parallel-coactive processing of dimensions.," *Percept Psychophys*, vol. 63, no. 5, pp. 901–17, 2001.

[10] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," *Image Rochester NY*, vol. 3021, no. 6, pp. 228–241, 2004.

[11] M. Mancas, D. Unay, B. Gosselin, and B. Macq, "Computational attention for defect localisation," in *Proceedings of 5th International Conference on Computer Vision Systems*, 2007.

[12] 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, pp. 1254–1259, Nov 1998.

[13] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search.," *Journal of Experimental Psychology: Human Perception & Performance*, vol. 15(3), pp. 419–433, 1989.

[14] J. M. and Wolfe, "Visual search in continuous, naturalistic stimuli," *Vision Research*, vol. 34, no. 9, pp. 1187 – 1195, 1994.

[15] J. M and Wolfe, "Visual memory: What do you know about what you saw?," *Current Biology*, vol. 8, no. 9, pp. R303 – R304, 1998.

[16] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model.," *Journal of vision*, vol. 9, no. 12, pp. 1–15, 2009.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, pp. 545–552, MIT Press, 2007.

[18] U. Rajashekar, I. van der Linde, A. Bovik, and L. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *Image Processing, IEEE Transactions on*, vol. 17, pp. 564 –573, april 2008.

[19] P. Sharma, F. Cheikh, and J. Hardeberg, "Face saliency in various human visual saliency models," in *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*, pp. 327 –332, sept. 2009.

[20] Y.-L. Tian and A. Hampapur, "Robust salient motion detection with complex background for real-time video surveillance," in *Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION'05) - Volume 2 - Volume 02*, WACV-MOTION '05, (Washington, DC, USA), pp. 30–35, IEEE Computer Society, 2005.

[21] A. Belardinelli, F. Pirri, and A. Carbone, "Motion saliency maps from spatiotemporal filtering.," in *WAPCV'08*, pp. 112–123, 2008.

[22] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, pp. 1093–1123, 2005.

[23] F. F. E. Guraya, F. A. Cheikh, A. Tremeau, Y. Tong, and H. Konik, "Predictive saliency maps for surveillance videos," in *Proceedings of the 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*,

DCABES '10, (Washington, DC, USA), pp. 508–513, IEEE Computer Society, 2010.

[24] S.-W. Ban, I. Lee, and M. Lee, "Dynamic visual selective attention model," *Neurocomput.*, vol. 71, pp. 853–856, Jan. 2008.

[25] N. S. L. Wei and Y. Wang, "A spatiotemporal saliency model of visual attention on maximum entropy," in *In proceedings of Canadian Geomatics Conference 2010*, 2010.

[26] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 12, pp. 507 – 545, 1995. ¡ce:title¿Special Volume on Computer Vision¡/ce:title¿.

[27] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, pp. 802 –817, may 2006.

[28] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, Jan. 1980.

[29] F. F. E. Guraya, F. A. Cheikh, and V. Medina, "A novel visual saliency model for surveillance video compression," 2011.

[30] R. Desimone, T. Albright, C. Gross, and C. Bruce, "Stimulus selective properties of inferior temporal neurons in the macaque," *The Journal of Neuroscience*, vol. 4(8), pp. 2051–2062, 1984.

[31] D. Walther and C. Koch, "Modeling attention to salient protoobjects.," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[32] Y.-F. Ma and H.-J. Zhang, "A model of motion attention for video skimming," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, pp. I–129 – I–132 vol.1, 2002.

[33] S.-C. Y. Dwarikanath Mahapatra, Stefan Winkler, "Motion saliency outweighs other low-level features while watching videos," in *In proceeding of SPIE Human Vision and Electronic Imaging*, vol. 6806, pp. 68060P–68060P–10, 2008.

[34] I. Richardson, *The H.264 Advanced Video Compression Standard*. John Wiley & Sons, 2010.

[35] "Jm 18.0," http://iphome.hhi.de/suehring/tml/download/.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 13, no. 4, pp. 600–612, 2004.