Himawan, Ivan, Song, Wei, & Tjondronegoro, Dian W.

(2013)
Automatic region-of-interest detection and prioritisation for visually optimised coding of low bit rate videos. In
Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, Clearwater Beach, Florida, pp. 76-82.

Automatic Region-of-Interest Detection and Prioritisation for Visually Optimised Coding of Low Bit Rate Videos

Ivan Himawan, Wei Song, Dian Tjondronegoro
Science and Engineering Faculty, Queensland University of Technology
Brisbane, QLD, Australia
i.himawan@qut.edu.au, w1.song@qut.edu.au, dian@qut.edu.au

Abstract

The increasing popularity of video consumption from mobile devices requires an effective video coding strategy. To overcome diverse communication networks, video services often need to maintain sustainable quality when the available bandwidth is limited. One of the strategy for a visually-optimised video adaptation is by implementing a region-of-interest (ROI) based scalability, whereby important regions can be encoded at a higher quality while maintaining sufficient quality for the rest of the frame. The result is an improved perceived quality at the same bit rate as normal encoding, which is particularly obvious at the range of lower bit rate. However, because of the difficulties of predicting region-of-interest (ROI) accurately, there is a limited research and development of ROI-based video coding for general videos. In this paper, the phase spectrum quaternion of Fourier Transform (PQFT) method is adopted to determine the ROI. To improve the results of ROI detection, the saliency map from the PQFT is augmented with maps created from high level knowledge of factors that are known to attract human attention. Hence, maps that locate faces and emphasise the centre of the screen are used in combination with the saliency map to determine the ROI. The contribution of this paper lies on the automatic ROI detection technique for coding a low bit rate videos which include the ROI prioritisation technique to give different level of encoding qualities for multiple ROIs, and the evaluation of the proposed automatic ROI detection that is shown to have a close performance to human ROI, based on the eye fixation data.

1. Introduction

The emerging mobile technologies have created a huge demand for video streaming applications. Users increasingly choose portable devices, including smart phones and tablets, to consume visual media for the freedom to watch anytime and anywhere in the easy to carry pocketsize device. In order to deliver high quality videos smoothly over the wireless networks, it is important for content providers to optimise the media for customers's devices while at the same time maintaining a high quality of service, especially when the bandwidth becomes limited.

Detection of region of interest can be useful to provide users with high quality videos by preserving both the content and the perceived quality. In video retargeting task for example, the important regions can be preserved using methods such as cropping and scaling [16, 17], warping [23], and seam carving [2, 7] when adapting large resolution video to a smaller screen size. It is also possible to allocate more bits to the region of interest at the cost of reduced bits in the background. In the videoconferencing application for example, face is the object of interest. Rather than transmitting the whole scene in a high quality video, only the face region is encoded in high quality [5]. Similar strategy can be applied to various types of videos, assuming that viewers are mainly interested at ROI and would not mind lower quality on the rest of the frame when the bandwidth is limited. In addition, a bitrate adaptation technique can be employed in such situation to adjust differently the regions within each frame of streamed videos depending on the network conditions [6]. Recent study in [21] shows that ROI enhancement improves user perceived quality in low bit rate sports video using a mobile phone (i.e. iPhone). More recent results using the same device shows that in the lower bit rate range (e.g. 300-500kbps), people prefers watching ROI-based video, albeit with manually annotated ROI [9].

One challenge for ROI-based video coding is to accurately predict where people look when watching a particular scene. An ideal solution would be to track the viewer's eye movements in real time and selecting areas surrounding the point of gaze. This ultimate method of determining ROI is unfortunately not practical due to the necessity for a user wearing eye tracking device which is obtrusive. It also suffers from technical and cost implications to be deployed in real time applications. The more practical approach is to automatically predict the ROI by analysing the video contents.

This paper proposes a novel technique for automatic ROI detection and prioritisation of multiple ROIs for a visually optimised video coding, which is then evaluated using the eye tracking data from videos in DIEM database [1]. The videos contain documentaries, game video trailers, music clips that are commonly watched in a daily basis. The ROI-based video coding is implemented based on the open source x264 to generate videos with H.264/AVC format, a widely adopted standard for video encoding. Section 2 presents the related works. Section 3 describes the implementation of automatic region of interest detection, showing the steps of how ROIs are obtained and prioritised. Section 4 describes the implementation of ROI-based video encoder. The evaluation of automatic ROI detection and video encoding are presented in Section 5. This is then followed by conclusion remarks in Section 6.

2. Related Works

Due to the difficulties in designing a generic ROI that would work effectively for arbitrary videos, research on ROI-based video coding is usually limited to specific applications, such as videoconferencing based on face and skin detection. Moreover, determining the region of interest is a very subjective task since it tries to automate the process of human cognition. Human's vision is capable of naturally focusing attention towards the interesting pixels that stand out from the rest of a video frame, which in turn can be interpreted semantically as a person and other objects. Computationally modelling of this ability has been studied over a long period of time based on physiology, psychology, and neural systems. In the field of computer vision, early work by [13] and subsequent work building on this research [10] suggest that visual attention is the result of a fast, pre-attentive, bottom-up, data driven saliency detection; in conjunction with slower, task-dependent, top-down, goal driven saliency detection. This biologically inspired model is quite dominant in the field because it has a strong theoretical foundation in the study of human attention mechanisms.

Recently, many research efforts have directed their interest toward computational model without strictly reproducing biological structures of human vision systems. For example, Ma and Zhang [18] generate saliency maps using contrast analysis and then extract the attended areas using fuzzy growing techniques. Zhai and Shah [24] calculate image color statistics in order to generate pixel-level saliency. Further, Achanta *et al.* [1] analyse a couple of saliency detection methods and treat them as a filtering operation in the frequency domain. A bandpass filter is then designed to extract salient objects in the image. All the aforementioned research in general uses a core principle, whereby regions that stand out from the surroundings capture attention. Thus, the saliency map may be generated using one or more features of intensity, color, and orientation by evaluating the relative contrast of image regions compared to the entire image.

3. Automatic ROI Detector

The ROI is determined from the video's saliency map. The saliency map is generated from the reconstruction of phase spectrum of the image's Fourier transform based on the technique presented in [8]. To enable the handling of salient features of the image such as color, intensity, and motion, in the frequency domain in a holistic manner, the quaternion Fourier transform is used [20].

A video with a total number of T frames is processed frame by frame as an image $I(x, y, t)$, where x and y are the location of each pixel and $t = 1, 2, ..., T$. For each input image, the RGB color frame is decomposed into luminance Y and two chrominance components, C_r and C_b [17]. The motion feature $M(t)$ is calculated by frame differencing, $M(t) = \|Y(t) - Y(t-\tau)\|$ to capture the temporal saliency between frames with the latency of τ.

The four features are represented by a quaternion image $q(t)$ which has four channels,

$$q(t) = M(t) + C_r(t)\mu_1 + C_b(t)\mu_2 + Y(t)\mu_3 \qquad (1)$$

where $\mu_i, i = 1, 2, 3$, satisfies $\mu_i^2 = -1$, $\mu_1 \perp \mu_2, \mu_2 \perp \mu_3, \mu_1 \perp \mu_3, \mu_3 = \mu_1\mu_2$. The $q(t)$ can be further represented in *symplectic* form, $q(t) = q_1(t) + q_2(t)\mu_2$, where $q_1(t) = M(t) + C_r(t)\mu_1$, $q_2(t) = C_b(t) + Y(t)\mu_1$. The Quaternion Fourier Transform (QFT) of a quaternion image $q(t)$ is then computed to obtain $Q(t) = Q_1(t) + Q_2(t)$, where $Q(t)$ is the frequency domain representation of $q(t)$.

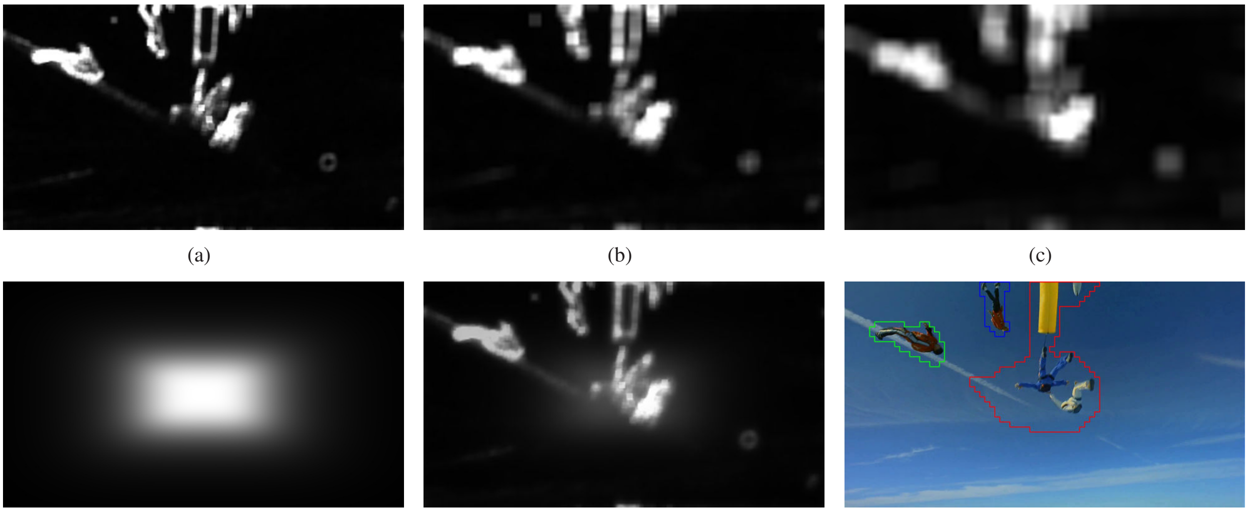
In polar form, Q (t is dropped for clarity sake) can be represented as $Q = ||Q|| \exp^{\mu\Phi}$, where Φ is the phase spectrum of Q and μ is a unit pure quaternion. To obtain the image's phase spectrum Q_p, the $||Q||$ is set to unity (i.e. $||Q|| = 1$) or by computing $Q_p = \frac{Q}{||Q||}$. The spatio-temporal saliency map is obtained by convolving the q_p with 2D smoothing filter g,

$$sM(x, y, t) = g * ||q_p||^2 \qquad (2)$$

where q_p is the inverse Fourier transform of Q_p.

The saliency map can be constructed in different sizes. The variation in sizes simulates the various view distances between the observer and the scene. The coarser resolution is obtained from a smaller size, mimicking the observer looking at the scene from a long distance. In this case, the

[1] http://thediemproject.wordpress.com



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1. The saliency map of (a), (b), and (c) are combined together with a generalized bell function map (d) to obtain the final saliency map (e). Multiple ROIs in (f) are enclosed in red lines for the 1^{st} priority ROI, in green lines for the 2^{nd} priority ROI, and in blue lines for the 3^{rd} priority ROI respectively. The total area of ROIs are constrained to be 15% of a frame size.

global features are emphasised whereas the fine details of features are omitted. Conversely, the fine details of local features are displayed when the saliency map is constructed at larger sizes. In order to reliably extracting salient objects, a scale-invariant saliency map is constructed through a multi-scale analysis. The combination of saliency maps from different scales will hopefully highlight both local and global features. Three maps which are from $1/4$, $1/8$, and $1/16$ of the original frame size are combined, with weights are empirically set to be 0.6, 0.3, and 0.1 respectively [12]. Note that a larger weight is given to a larger map to emphasise local features and restraint non-salient areas marked as salient. Figure 1.a, 1.b, and 1.c show saliency maps generated with 320x180 pixels, 160x90 pixels, and 80x45 pixels respectively.

The final region of interest is obtained by augmenting the saliency map with maps created from high level knowledge of factors that are known to attract human attention:

1. **Human Faces**. Faces are very significant cue in human perceptions. Study in [4] shows that when there is no instruction for human observers to look for anything, they will fixate on faces within the first two fixations with a probability of over 80%. In this paper, the Viola Jones face detector based on the implementation of Intel Open Source Computer Vision Library ("OpenCV") [3] is used to detect faces. The estimated face map is determined by convolving a delta function (x,y) at every detected faces centre location with 2D Gaussians. To reduce the computation time in this experiment, the face is detected for every half a second and assumed to stay for 0.5s duration.

2. **Screen Centre**. In typical sports/news coverage, the cameramen filming the action will locate region or object of interest and place it in the centre of the view. Thus salient region have tendencies to be located in the centre of screen [19, 11]. In order to emphasis the centre of screen, generalized bell function approximately centred in the middle of the frame is used to emphasize a combined saliency map after multiscale analysis. The generalized bell function $b(x, y)$ is defined as,

$$b(x,y) = \frac{1}{1 + \| \frac{x-\mu_r}{\sigma_r} \|^{2\beta} + \| \frac{y-\mu_c}{\sigma_c} \|^{2\beta}} \qquad (3)$$

where μ_r and μ_c are set to be a coordinate (i.e. row and column respectively) approximately in the middle of screen, and σ_r and σ_c control the width of the main beam of the bell function. Figure 1.d and Figure 1.e show the generalized bell function map and the final saliency map respectively.

To generate the ROI area, each frame of an input video is divided into macroblocks (16x16 pixels in H.264/AVC format [22]). Given the final saliency map, a macroblock is defined as a ROI area when the intensity of map within the macroblock bounding box is above certain threshold (the threshold is determined using a binary search to constraint the ROI within the specified area). The transition between ROI region between consecutive frames are also smoothed by setting an upper and lower values so that the ROI region in a next frame will be changed only when the intensity of map is not between these two values.

3.1. Region-of-Interest Prioritisation

The region of interest can be located in multiple positions within a frame and with different sizes. Assuming a viewer focuses only on a particular region at one instance of time, it would be more efficient if this particular ROI is encoded at a higher quality than the other ROIs. Rather than equally encoding ROI regions at higher quality, the less prominent ROIs should be encoded in less quality, so that more bits can be allocated to the non-ROI region. This strategy is assumed to reduce coding artefacts in the non-ROI region. It can also be implemented in adaptive video streaming whereby the quality of different ROIs is adjusted accordingly depending on the bandwidth conditions.

The different priority levels of ROIs are determined by first evaluating the sum of intensity of saliency map within each ROI area. The N number of ROIs with their values are given as $rM_1, rM_2, ..., rM_N$. These values are then quantised into three levels with uniform sampling, so that the ROI with the largest value will be assigned the 1^{st} priority and the ROI with smallest value will be assigned the 3^{rd} priority. Figure 1.f shows the example of three ROI locations with three quality levels, each is enclosed by bounding boxes with specific color.

4. Region-of-Interest Encoder

The ROI encoder is a custom-developed x264 codec[1] that is able to allocate different amount of bits to the ROI/non-ROI area by changing the quantisation parameter (QP) values of macroblocks. The region of interest are pointed out to the encoder as macroblock positions within a frame. To generate the ROI encoded video, the two pass encoding process was used. In the first pass, a video source is encoded with a higher bitrate than the target bitrate. In the second pass, the quality of the 1^{st}, 2^{nd}, 3^{rd} priority ROI area and the non-ROI area (the 4^{th} priority) is reduced successively by increasing the quantisation parameter value using the equation below:

- $QP_1 = QP_o + 1$ for the 1^{st} priority ROI
- $QP_2 = QP_o + 2$ for the 2^{nd} priority ROI
- $QP_3 = QP_o + 3$ for the 3^{rd} priority ROI
- $QP_{BG} = QP_o + 6$ for the non-ROI region.

where QP_o is the QP values which are assigned by the encoder in the first pass.

Compared to the normal encoded video, the ROI encoded video has a higher quality in the ROI area and a lower quality in the non-ROI area at a given bitrate. The target bitrate was set to hit approximately 500kbps for experiments in this paper. No B-frames were used and a maximum GOP size was set to 50. The subpixel motion estimation and mode decision was set to sum of absolute transformed differences (SATD) mode decision and number of reference frames used is one.

5. Region-of-Interest Video Coding Evaluation

Twelve videos from the DIEM project were selected for experiments in this paper. The project used various video contents such as advertisement, documentaries, film trailers, music clips to collect the eye tracking data. The twelve videos have a resolution of 1280x720 with no audio. The videos' name, length, number of individuals watching the video in which the eye tracking data were collected and description, are shown in Table 1 below.

5.1. Accuracy of Automatic ROI

The accuracy of the proposed automatic ROI detector is evaluated by calculating the percentage of subject's fixated points that fell within the automatic ROI region. The final result is obtained by averaging this percentage for all frames for each video. The higher the percentage is an indication that the automatic ROI region corresponds to the region where human looks. As a baseline comparison, the percentage of subject's fixated points that fell within the rectangular ROI that is centered in the middle of the frame is calculated.

Human performance to predict eye fixations is also evaluated using a method outlined in [11]. First, the human ROI is determined by choosing one subject's fixation point and draw a rectangular ROI with video's aspect ratio centred around this fixated point. This one human's ROI is then used to evaluate other humans' fixation points excluding the one chosen, whether it fell inside this one human's ROI for every frame. By varying the area of human's ROI, the ROC curve in Figure 2 is drawn, averaged for all subjects and all twelves videos' frames. Higher human performance is observed on videos which contain a ROI that pops out where fixations are likely clustered together on that particular ROI location, compared to the videos without ROI that pops out [14].

For experiments in this paper, the ROI area is set to be 15% for both automatic ROI and rectangular ROI. This covers approximately 70% of the ground truth human fixations (see Figure 2). There is also a trade-off between ROI area, ROI encoding quality, and the resulting ROI-based video perceived quality. For example by giving a stronger quality in the ROI area will require more bits to encode the ROI at the expenses of non-ROI bits. This may degrade the overall perceived quality if coding artefacts in the non-ROI area is too annoying [15]. If the ROI area is larger, lesser bits will also be allocated to the non-ROI area. On the other hand, if the ROI area is too small, there is a high chance that it will not predict human fixations well and the video will not

[1] Original source code available online at http://www.videolan.org

Table 1. *Twelve videos from DIEM project used in the experiments.*

Video Name	**Length**	**Number of Subjects**	**Descriptions**
coral_reef_adventure (CR)	1:58	42	Underwater shots of people diving in coral reefs.
dolphins (DP)	2:07	124	Underwater shots of dolphins swimming in the sea.
london_no_voices (LV)	2:33	50	Faces of various people talking to the camera on the street.
adrenaline_rush (AR)	2:11	123	Shots of people parachuting with a blue sky in the background.
discoverers (DC)	3:02	42	Mix scenes of observatories on hills, inside the obsevatory, and people inside a building.
mystery_nile (MN)	1:44	42	Shots of water rafting in the rough river.
nine_inch_nails (NI)	0:51	42	A band playing in a concert. Most of scenes are quite dark.
bullet_witch (BW)	2:28	123	A video game trailer showing gameplay animation.
ghostbusters (GB)	2:04	219	A video game trailer showing gameplay animation.
lego_indiana_jones (LI)	2:12	218	A video game trailer showing gameplay animation.
barcelona_extreme (BE)	1:08	46	Mix scenes of waterskiing and skateboarding, and the spectators watching the sports.
F1_slick_tyres (FS)	1:30	46	Shots of F1 racing car with a short interview of the driver.

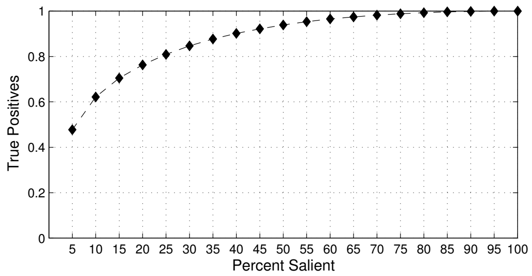


Figure 2. The ROC curve of human performance. Fifty percent of humans will fixate within the 5% of a novel viewer's ROI area, and 90% are within the 40% ROI area.

benefit from the ROI coding. The overall quality will also depend on the actual objects or the low-level visual cues the ROI area covers. More research is needed to understand this relationship but not in the scope of this paper.

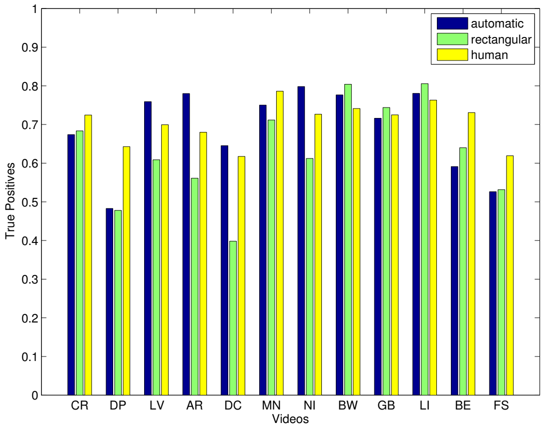


Figure 3. The bar plot performance of automatic ROI detection along with rectangular ROI and human for the twelve test videos.

The average performance of automatic ROI along with human and rectangular ROI for the twelve videos are given in Table 2, while the details for each video is shown in Figure 3. For most of videos, automatic ROI has either better or having comparable performance with rectangular ROI, except the BW, GB, LI, and BE video where rectangular ROI performs better. The average of the twelve videos shows that the use of automatic ROI detection is promising and comparable to human performance. Note that there is a strong cluster for human fixations to be located around the centre of the frame, with an average of 63% fixations are within 15% of rectangular ROI area. Another interesting point from the bar plot is that it is possible for the automatic ROI to exceed the human performance. This is possible if there are two or more distinguish ROIs in a frame so that one group of subjects fixate on one of ROIs, and the other fixate on the other ROI, while a novel's viewer ROI can cover only one of ROIs. Therefore, it is an advantage to have multiple ROIs in a frame. Comparable performance between the automatic ROI detector and the human shows that the strategy employed by the proposed automatic ROI detection is able to predict human gaze quite well.

Table 2. *The average performance of automatic ROI detection (in %) in comparison with human and rectangular ROI when the ROI area is 15% of the frame's area.*

automatic	*human*	*rectangular ROI*
69.0	70.4	63.1

5.2. Performance of ROI-based Video Encoding

Figure 5.2 shows the snapshots from Normal and ROI encoded videos at 500kbps. The graph presented in Figure 5 shows the average bits within ROI and non-ROI area for both ROI and Normal encoded videos. The improvement in the ROI area while reduction of bits in the non-ROI area for ROI-based video coding is shown clearly in the figure. The bits allocation for each ROI priority area are shown in Table 3.

Having multiple qualities for ROIs (i.e. with different

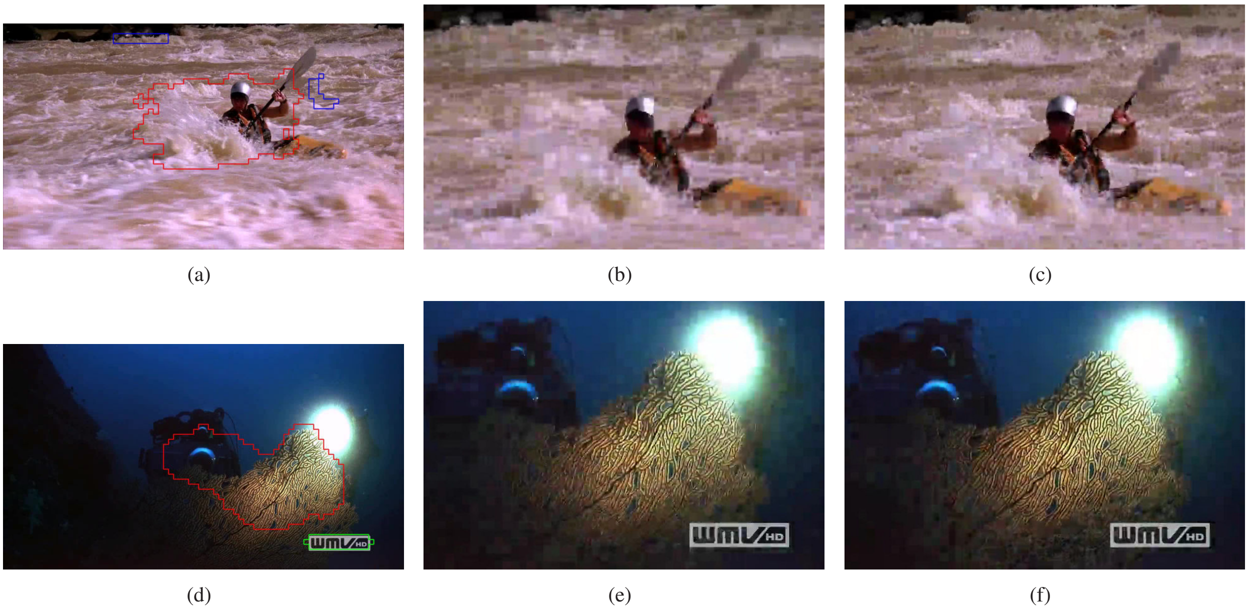


Figure 4. Screenshots from Normal and ROI encoded videos at 500kbps for MN (top) and BE (bottom). The frames with ROI boundaries are displayed in the orginal frame size (a and d), while the Normal (b and e) and ROI encoded (c and f) videos are cropped and zoomed to highlight the difference of encoding quality in the ROI region.

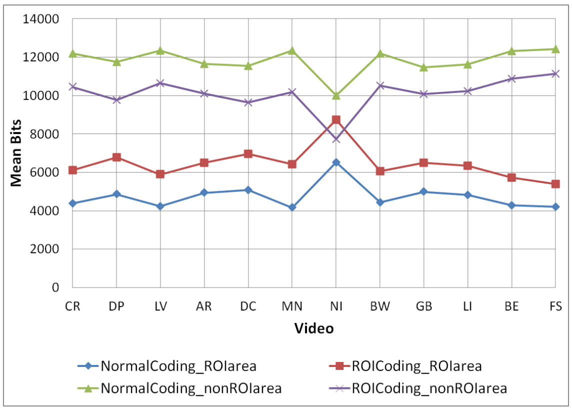


Figure 5. The average bits within ROI and non-ROI area for the 12 test videos.

Table 3. *The average bits of different priority area between Normal and ROI-based encoded videos.*

Type of Encoding	1^{st} *ROI*	2^{nd} *ROI*	3^{rd} *ROI*
Normal	3622	1010	1221
ROI-based	5112	1287	1407

priorities) can be advantageous since more bits can be allocated to the non-ROI region. Figure 6 shows the ROI-based encoded videos that are generated with 4 priorities in comparison to videos that are generated with 2 priorities (i.e. 1^{st} priority is ROI region and 2^{nd} priority is non-ROI region). The average bits allocation from the twelve videos in Figure 6 shows that having 4 priorities increase the bits allocation to the non-ROI area while reducing the bits allocation to the ROI area. However, this improvement is small for experiments in this paper which is caused by smaller areas of the 2^{nd} and 3^{rd} priority ROIs compared to the first priority ROI area. Analysing the relationship between the area of ROIs and bits allocation will be left for future works.

The time to encode ROI-based video in the current x264 two pass implementation takes approximately three and half times longer than the time to encode normal video using one pass. Using the CR video with the length of 1 minute and 38 seconds, the ROI encoding takes an average of 8 minutes and 36 seconds while the normal encoding takes 2 minutes and 36 seconds in 10 trials. The test were run on a Windows platform, equipped with Core2 Duo 2.33GHz and 2GB of memory.

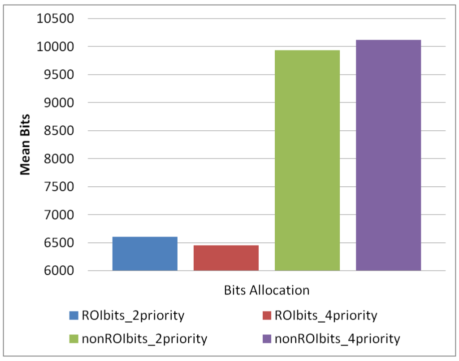


Figure 6. The average bits within ROI and non-ROI area for the 12 test videos when the ROI-based encoded video has either two or four priorities encoding strategy.

6. Conclusion

This paper presents a novel technique for automatic region-of-interest detection for the purpose of visually optimised video coding for general videos. The ROI detection is developed from the saliency analysis of video frames in combination with face detection, while emphasising the importance of the centre of the frame. The result is the ROI location that likely to sit in the centre location which then expand to cover any stand-out objects incuding faces that appears on the screen. The ROI detection accuracy is shown to have a comparable performance with human. Screenshots of videos which are encoded at the low bit rate with ROI video coding show a better perceptual quality compared to the normal encoded videos (with further evidence is available in the submitted supplementary materials). Ongoing research investigates the impact of ROI-based encoded videos, at various target bit rates, on the perceived quality for various mobile devices through subjective experiments.

7. Acknowledgments

This research was carried out as part of the activities of, and funded by, the Smart Services Cooperative Research Centre (CRC) through the Australian Governments CRC Programme (Department of Innovation, Industry, Science and Research).

References

[1] R. Achanta et al. Frequency-tuned salient region detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597 –1604, June 2009. 2

[2] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *ACM SIGGRAPH*, SIGGRAPH '07, New York, USA, 2007. ACM. 1

[3] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with Intel's open source computer vision library. *Intel Technology Journal*, 9(2), 2005. 3

[4] M. Cerf et al. Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, 2007. 3

[5] M.-J. Chen et al. ROI video coding based on H.263+ with robust skin-color detection technique. *IEEE Transactions on Consumer Electronics*, 49(3):724–730, 2003. 1

[6] B. Ciubotaru et al. Objective assessment of region of interest-aware adaptive multimedia streaming quality. *IEEE Transactions on Broadcasting*, 55(2):202 –212, June 2009. 1

[7] M. Grundmann et al. Discontinuous seam-carving for video retargeting. pages 569 –576, june 2010. 1

[8] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1):185–198, 2010. 2

[9] I. Himawan et al. Impact of region-of-interest video coding on perceived quality in mobile video. In *Proceedings of IEEE International Conference on Multimedia & Expo*, 2012. 1

[10] L. Itti et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 2

[11] T. Judd et al. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, 2009. 3, 4

[12] W. Kim et al. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):446 –456, 2011. 3

[13] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuity. *Human Neurobiology*, 4:219–227, 1985. 2

[14] O. Le Meur et al. Do video coding impairments disturb the visual attention deployment? *Image Communication*, 25(8):597–609, Sept. 2010. 4

[15] J.-S. Lee et al. Subjective quality evaluation of foveated video coding using audio-visual focus of attention. *IEEE Journal of Selected Topics in Signal Processing*, 5(7):1322–1331, 2011. 4

[16] F. Liu and M. Gleicher. Video retargeting: automating pan and scan. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 241–250, 2006. 1

[17] T. Lu et al. Video retargeting with nonlinear spatial-temporal saliency fusion. In *Proceedings of IEEE International Conference on Image Processing*, pages 1801 –1804, 2010. 1, 2

[18] Y. F. Ma and H. J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the 11th ACM international conference on multimedia*, 2003. 2

[19] J. D. McCarthy et al. Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 535–542, 2004. 3

[20] S. Sangwine. The discrete quaternion fourier transform. In *International Conference on Image Processing and Its Applications*, volume 2, pages 790 –793 vol.2, July 1997. 2

[21] W. Song et al. Impact of zooming and enhancing region of interest for optimizing user experience on mobile sports video. In *Proceedings of ACM International Conference on Multimedia*, 2010. 1

[22] T. Wiegand, G. Sullivan, and A. Luthra. Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC). *Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT - G050r1*, 2003. 3

[23] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video-retargeting. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007. 1

[24] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th annual ACM international conference on multimedia*, 2006. 2