FAST REGION OF INTEREST SELECTION IN THE TRANSFORM DOMAIN FOR VIDEO TRANSCODING

Safak Dogan, Abdul H. Sadka and Ahmet M. Kondoz

CCSR, University of Surrey, Guildford GU2 7XH, Surrey, UK
e-mail: {S.Dogan, A.Sadka, A.Kondoz@surrey.ac.uk}

ABSTRACT

Finding an area of visual attention in coded video is not as straightforward as selecting it in the raw video domain. This is due to the fact that coded scenes do not reveal the content-specific characteristics easily. Nevertheless, selection of a region of interest (ROI) in compressed video has a great advantage, as it can be used for object-based video transcoding to perform content adaptation over heterogeneous networks. This paper presents a fast ROI selection algorithm for video transcoding purposes. This algorithm is used to identify a visually important area in a coded video scene during the downscaling of a high resolution input to lower resolutions in the transform domain. The test results demonstrate the accuracy of the selected ROI.

1. INTRODUCTION

Compressed video communications have gained great momentum in recent years owing to the significant progress in signal processing and telecommunication technologies. As a result, coded video is today widely exchanged within a number of application scenarios, such as digital video broadcasting, video-on-demand and video-conferencing/telephony. Each of these applications is associated with a specific video compression standard and a dedicated access network as well as a pre-determined set of communication protocols for ensuring an acceptable level of service quality. However, due to the rapid proliferation of new video services targeting a very wide range of networking platforms, terminals and users, the boundaries between different application scenarios have also become increasingly vague. Consequently, this has caused services to cross over several application domains resulting in highly heterogeneous video communication systems.

As more diverse terminals and users with varying requirements are connected to a particular video service, the coded video traverses a higher number of access networks with different characteristics. When compressed video is exchanged between heterogeneous communication networks, its coding and transmission parameters need to be either re-negotiated or re-configured to retain the pre-set level of service quality. If extra care is not taken during the set-up of such video connections, application, network and device-centric mismatches impede the distribution of video to users, which in some cases may even lead to the loss of the flow and integrity of entire communications. In literature, it is addressed as the problem of seamlessly accessing the video (or multimedia in general) content from any network, with any device and at any time, which constitutes the central theme of the universal multimedia access (UMA) concept [1]. To circumvent this problem, several strategies have been developed within UMA [2,3]. The key technology of these strategies is based on the video content adaptation techniques. These techniques aim at improving the success rate and service quality of the end-to-end video delivery, both of which are greatly affected by the growing heterogeneity in multimedia:

- Client-device capabilities.
- Access network characteristics.
- Content representation formats.
- User preferences.
- Natural environment of users and end-terminals.

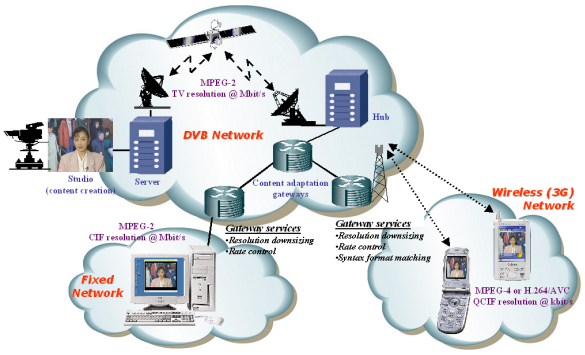

Fig. 1: Heterogeneous video communications scenario

The work presented in this paper focuses specifically on one of the content adaptation techniques, namely video transcoding, which performs online or offline re-formatting of an input compressed video stream [4]. In this paper, a gateway operation is utilised to deliver high quality and high bit rate video services to a number of users connected by different access networks, as illustrated in Fig. 1. For this purpose, the video transcoding is performed at the gateway to downscale an input high resolution video to lower resolutions, as dictated by the variety of display capabilities of heterogeneous client-terminals. However, high resolution video mostly contains large areas of scenes that are typically not of high importance to human visual attention. Therefore, a fast region of interest (ROI) selection algorithm is proposed in the paper to identify a region or an object of interest during the downscaling process. The fast operation of this algorithm is achieved in the transform domain with minimal decoding of the compressed video. The rest of the paper is

The work presented was developed within VISNET, a European Network of Excellence (http://visnet-noe.org), funded under the European Commission IST FP6 programme.

organised as follows: The next section gives an overview on the role of visual attention in selecting an ROI in video. The third section discusses the proposed ROI selection algorithm in the transform domain. The fourth section presents the test results along with discussions. Finally, the last section outlines the concluding remarks and respective future work items.

2. ROLE OF VISUAL ATTENTION IN ROI SELECTION

Whilst viewing an image or a sequence of images, the human visual system (HVS) usually focuses on parts of a particular scene. This is due to the fact that some part or parts of an image or video draw more attention than the others, as they may be of particular interest to the viewer. Moreover, different viewers may have different visual attention points within the same scene. For instance, Fig. 2 depicts a scene of an accident that takes place at a Formula-1 car race. In such a scene, many of the viewers would direct their viewpoints to the particular racing car that is directly involved in the accident, which can be identified as the primary area of visual attention. However, some viewers may also focus their attention on the crashing car as well as the car next to it, so as to observe whether or how the second car will be affected from this tragic event. Therefore, the area of visual attention now contains two objects of interest, one of which is the primary area/object to the point of view. Of course, this is a very extreme example, but it is always possible to extend this logic to other video scenes, such as movies, news, sports and music videos, etc, in which more than one area of visual attention could exist. Throughout a video sequence, this point of attention may or may not vary within one scene when new objects are introduced. However, all the areas of visual attention are not always equally important to the HVS, and thus viewers usually tend to ignore the regions of low importance, such as the spectators, billboards, racetrack or other racing cars in Fig. 2. In light of these discussions, the relationship between an area of visual attention and an ROI can be generalised as follows [5-7]:

- Several factors (e.g. motion, shape, colour or contrast in relation to the background) and distinctive events are more likely to draw the attention of a viewer in a scene.
- The area of visual attention is thus focused on the visually salient or most interesting object or region within an image, which represents the ROI of a scene.
- This area may comprise one or more ROIs.

Various methods have been presented to determine an area of visual attention in literature to date. Moreover, these methods have been exploited to develop a number of algorithms for ROI selection [7-9]. Nevertheless, most of these algorithms were employed to select an ROI in the pixel domain during the encoding of a video sequence [6,7,10]. Therefore, they are not quite adequate for gateway operations for heterogeneous video communications with quick system responses. Thus, recent research has focused on finding ROI in the compressed domain to allow for a number of fast applications, such as: transcoding systems, object detection, tracking and identification techniques, image and video retrieval/summarisation schemes based on MPEG-7 descriptors, event detection, and audiovisual content analysis and understanding tools, etc [11-17].

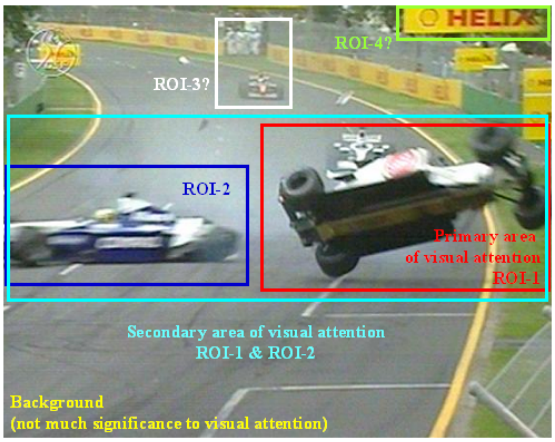

Fig. 2: Areas of high and low visual attention and ROI in video

3. ROI SELECTION IN TRANSFORM DOMAIN

The ROI selection provides a key advantage during transcoding, as it identifies a visually important area or object in the compressed video. The advantage is particularly significant when video services are distributed across a wide range of heterogeneous client-terminals with diverse display capabilities [5]. Selecting an ROI in compressed video allows a transcoding gateway to accurately re-format the resolution of input video whilst focusing on the main region or object of visual attention. In this way, the gateway is enabled to re-organise the pre-defined scene priorities allowing for unequal video parameter allocation to different parts of a scene based on their perceptual qualities [11,18].

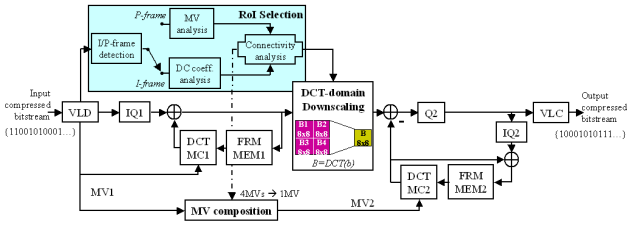

Fig. 3: ROI selection in downscaling transcoder architecture

In this paper, the ROI selection algorithm is employed during the transcoding operation in the transform domain, as shown by the shaded function block in Fig. 3. For this purpose, a transcoder has been utilised, which performs MPEG-2 video downscaling. This transcoder is designed to reduce the spatial resolution of a high quality input coded video to much lower resolutions in the discrete cosine transform (DCT) domain [19]. Whilst downsizing the high resolution video, the DCT-domain downscaling system makes both the transform coefficients and the valuable motion information available to the ROI selection algorithm. However, finding the ROI in the transform domain is not as straightforward as finding it in the pixel domain. This is due to the fact that only the DCT coefficients are accessible rather than the pixel values themselves.

The proposed ROI selection method is designed to treat the intra (I) and inter (P) frames separately in an MPEG coded video sequence, as shown in Fig. 3. These frames can be easily recognised by analysing the video frame headers. The transform

domain data in I-frames gives a good indication of where the main object or ROI might be in a video scene by providing the different colour and illumination characteristics without having to obtain the original pixel values. With this knowledge in mind, the ROI selection algorithm analyses the DC colour information for each macroblock (MB) prior to the downscaling of the input video. The DC coefficients are available to the transcoder in the transform domain as well as the rest of the DCT coefficients. In [14], a method is presented to identify an ROI based on extracting the Dominant Colour Descriptor (DCD) using the detected DC values. The DCD is a standardised descriptor defined by the MPEG-7 standard [15]. However in our proposed method, the DC colour differences are directly evaluated and the different areas with significant contrast are identified. This is a very straightforward and fast approach without needing to consult with a standardised colour descriptor. The weakness of this method however is that it relies on the colour contrast between the ROI (e.g. a visually significant foreground object) and the rest of the video scene (e.g. the background).

Whilst finding the ROI in P-frames, it is considered that the ROI consists of a motion-active area or object. Based on this assumption, the available motion vector information is exploited before the downscaling of these vectors. For this purpose, the regions without any motion characteristics are categorised as the background, and hence they are accepted to be insignificant to the human visual perception. On the contrary, the areas in motion are identified as the regions for visual attention, and thus classified as the ROI. Moreover, the detected motion in P-frames allows for the temporal tracking of the identified ROI from the previous I-frames in a sequence of I- and P-frames. This is achieved by determining the shape of the ROI roughly in an I-frame, and retaining this shape in line with the detected motion information in the subsequent P-frames. In this way, the connectivity between neighbouring regions, which present identical motion/colour characteristics belonging to one ROI, is maintained whilst avoiding the erroneous detection of uncorrelated (isolated) regions in the scene. Currently, the transform-domain ROI selection algorithm has been devised to process a primary motion-active foreground object. When there is more than one moving object in a scene, more elaborate motion analysis is required for identifying different active regions. [12,13] present methods for detecting and tracking various moving objects in a sequence of coded video.

4. RESULTS AND DISCUSSIONS

This section demonstrates the test results obtained from the developed ROI selection algorithm. The experiments were conducted using a standard CIF resolution (352×288 pixels) video test sequence, called *Bream*. This is a single-object video scene, which contains a bright yellow fish swimming in water with relatively moderate motion characteristics. The water is represented in the scene as the stationary dark blue background. Thus, it is anticipated that the transform-domain ROI selection algorithm performs sufficiently well to identify the active foreground object with significant colour contrast compared to the background.

The ROI selection is carried out during video downscaling, as discussed previously. Therefore, the input to the transcoder was the compressed video sequence, which was originally encoded with MPEG-2 in progressive mode using frame predictions at 4Mbit/s and 25fr/s. The video clip was coded as an I-P-P-…-I-P-P-… sequence with a distance of 12 frames between the regular I-frames (i.e. the group of pictures with a length of N=12 frames). Bi-directional (B) frames are not considered in the tests, as they do not contribute to the finding of ROI in this method. Nevertheless, the insertion of B-frames would not disturb the operation of the developed ROI selection algorithm, and hence N=12 and M=3 (i.e. a sub-group of 3 frames) coding would still be admissible, which is a typical MPEG-2 coding layout.

Fig. 4 shows the results of ROI identification in a set of different I-frames in CIF resolution prior to downscaling. In this figure, the # and – signs represent the detected MBs that belong to the ROI (based on the colour distinction) and the background, respectively. As can be observed from this figure, there is a close match between the foreground fish object and the selected ROI represented with the # signs. The regular I-frames help to maintain the approximate shape of the detected ROI for the duration of the video clip. The results show that the proposed DC coefficient analysis in the transform domain provides adequate information about the ROI when there is perceptible colour contrast between the foreground and background regions.

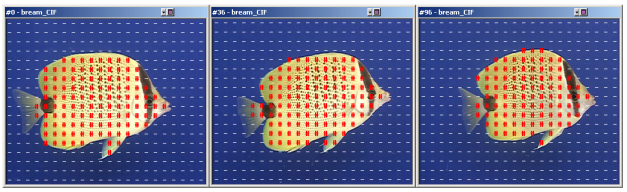

Fig. 4: ROI selection in CIF resolution I-frames

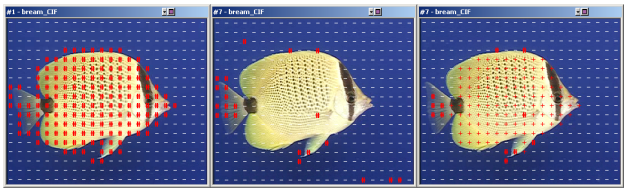

Fig. 5: ROI selection in CIF resolution P-frames

Fig. 5 presents the test results obtained from a number of P-frames. Similarly, the # and – signs represent the detected MBs that belong to the ROI (based on the motion activity) and the background, respectively. The results demonstrate that when there is any detectable motion, the ROI is accurately identified, as seen on the left-hand side of the figure. However, when the motion-active region/object becomes stationary, as in the centre frame, the ROI selection method, which is based only on the detection of motion activity, fails to identify the correct regions. Moreover, some isolated areas are incorrectly identified as the ROI, and hence the # signs are scattered in the scene, as shown in the figure. For this reason, the proposed method applies connectivity analysis to merge the neighbouring regions, which have similar motion/colour characteristics whilst also utilising the previously detected DCT-domain data (e.g. shape, motion, texture, etc) from the former I/P-frames. In this way, the correct tracking of the selected ROI is maintained, as indicated by the + signs in addition to the existing # signs on the right-hand side of Fig. 5. As a result, the wrongly detected regions are eliminated with this method.

Finally, Fig. 6 demonstrates the results of the ROI selection in CIF resolution before downscaling and the accurate tracking of this identified ROI in the corresponding downscaled QCIF (176×144 pixels) video in a set of I- and P-frames. These results show that the information of a selected ROI in high resolution video can be passed to the lower resolution counterpart during video downscaling for object-based transcoding. The DCT-domain video downscaling transcoder was estimated to achieve around 40~50% speed-up in computational complexity compared to the pixel-domain method. In addition to this fast operation of the transcoder, the proposed ROI method does not incur any further significant complexity. This is due to the fact that the proposed method exploits the video information already available for the downscaling process, such as the I/P-frame headers, DC/DCT coefficients and motion vectors. The connectivity analysis is performed using a frame-based ROI map with regular updates of the selected MBs, which itself is also a straightforward operation. Therefore, the overall added complexity is negligible. The proposed algorithm has been tested with other test sequences, which have similar colour contrast and motion characteristics to *Bream*, and similar results have been obtained. However, it has also been noted that more complex video scenes, which contain multiple ROIs without significant colour contrasts between each other and the background, require more elaborate ROI selection techniques at the expense of higher computational complexity.

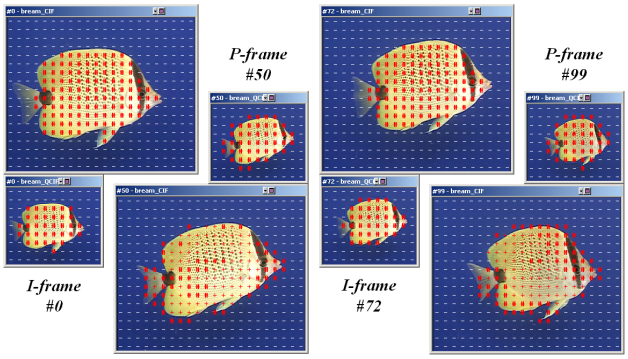

Fig. 6: ROI selection/tracking in CIF→QCIF downscaled video

5. CONCLUSION

This paper has presented a fast ROI selection algorithm whilst downscaling a high resolution coded video sequence. The fast operation of the algorithm is provided in the transform domain without the need for fully decoding the input compressed video stream. The selection of the ROI has been performed in I- and P-frames prior to the downscaling process at the video transcoder, and the test results have demonstrated the accurate detection and tracking of the selected regions in both high and low resolution video scenes. The future work will focus on developing an object-based video transcoding mechanism using the developed fast ROI algorithm to support better end-to-end video service qualities over heterogeneous networks in real time.

6. REFERENCES

[1] Special Issue on Universal Multimedia Access, *IEEE Sig. Proc. Mag.*, Vol. 20(2), Mar. 2003.

[2] Special Issue on Multimedia Adaptation, *Sig. Proc.: Image Commun.*, Vol. 18(8), Sep. 2003.

[3] S. Dogan, S. Eminsoy, A.H. Sadka and A.M. Kondoz, "Video content adaptation using transcoding for enabling UMA over UMTS", in *Proc. WIAMIS'2004*, Lisbon, Portugal, 21-23 Apr. 2004.

[4] J. Xin, C.-W. Lin and M.-T. Sun, "Digital video transcoding", *Proc. IEEE*, Vol. 93(1), pp. 84-97, Jan. 2005.

[5] K.B. Shimoga, "Region-of-interest based video image transcoding for heterogeneous client displays", in *Proc. PV'2002*, Pittsburgh, PA, USA, 24-26 Apr. 2002.

[6] A.P. Bradley and F.W.M. Stentiford, "Visual attention for region of interest coding in JPEG 2000", *J. of Visual Commun. and Image Rep.*, Vol. 14(3), pp. 232-250, Sep. 2003.

[7] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang and H.-Q. Zhou, "A visual attention model for adapting images on small displays", *ACM Multimedia Syst. J.*, Vol. 9(4), pp. 353-364, Oct. 2003.

[8] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans. Pattern Anal. and Machine Int.*, Vol. 20(11), pp. 1254-59, Nov. 1998.

[9] W. Osberger and A.J. Maeder, "Automatic identification of perceptually important regions in an image", in *Proc. ICPR'98*, Brisbane, Australia, 16-20 Aug. 1998.

[10]D.S. Cruz, T. Ebrahimi, M. Larsson, J. Askelof and C. Christopoulos, "Region of interest coding in JPEG 2000 for interactive client/server applications", in *Proc. MMSP'99*, Copenhagen, Denmark, 13-15 Sep. 1999.

[11]A. Sinha, G. Agarwal and A. Anbu, "Region-of-interest based compressed domain video transcoding scheme", in *Proc. ICASSP'2004*, Montreal, Canada, 17-21 May 2004.

[12]H. Zen, T. Hasegawa and S. Ozawa, "Moving object detection from MPEG coded picture", in *Proc. ICIP'99*, Kobe, Japan, 24-28 Oct. 1999.

[13]L. Favalli, A. Mecocci and F. Moschetti, "Object tracking for retrieval applications in MPEG-2", *IEEE Trans. CSVT*, Vol. 10(3), pp.427-432, Apr. 2000.

[14]V. Mezaris, I. Kompatsiaris and M.G. Strintzis, "Compressed-domain object detection for video understanding", in *Proc. WIAMIS'2004*, Lisbon, Portugal, 21-23 Apr. 2004.

[15]B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan and A. Yamada, "Color and texture descriptors", *IEEE Trans. CSVT*, Vol. 11(6), pp.703-715, Jun. 2001.

[16]K. Yoon, D. DeMenthon and D. Doermann, "Event detection from MPEG video in the compressed domain", in *Proc. ICPR'2000*, Barcelona, Spain, 3-7 Sep. 2000.

[17]G. Agarwal, A. Anbu and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed MPEG domain", in *Proc. ICME'2003*, Baltimore, MD, USA, 6-9 Jul. 2003.

[18]A. Vetro, H. Sun and Y. Wang, "Object-based transcoding for adaptable video content delivery", *IEEE Trans. CSVT*, Vol. 11(3), pp.387-401, Mar. 2001.

[19]S. Dogan, S.T. Worrall, A.H. Sadka and A.M. Kondoz, "DCT-domain downscaling for transcoding MPEG-2 video", in *Proc. ICCVG'2004*, Warsaw, Poland, 22-24 Sep. 2004.