IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 4, Issue 5, May 2016

FPGA Implementation of Best Suitable Video Compression With Respect to Region of Interest

Mohammad Sarparajul Ambiya[1], B Ramesh[2], Praveen J[3], Raghavendra Rao A[4]

M.Tech Student, Dept. of ECE, Alva's Institute of Engg & Technology, Mijar, Moodbidri, Karnataka, India[1]

Associate Professor, Dept. of ECE, Alva's Institute of Engg &Tech, Mijar, Moodbidri, Karnataka, India[2]

Sr. Associate Professor, Dept. of ECE, Alva's Institute of Engg &Tech, Mijar, Moodbidri, Karnataka, India[3, 4]

Abstract: ROI (Region of interest), defines the individual choice and this ROI principle applied on compression of video and transmission methods like foveation targets on exploiting the certain demerits of human visualisation power. The observing quality of the human decreases exponentially as the distance gets increased from the statically situated of video frame detection and correction of errors, speed control and calculation of performance in compression of a video. This paper provides a individual choice of location system based on prediction for HD football broadcast video. the proposed method makes uses of information about the context which is produced from analysis of individual's choosed location study that are experimented, in order to construct a flexible prior map. In addition to this, classified the complexity into sub categories through classification of various shots thus providing the model to pre understand the task relating directly to every object category and therefore constructing automatically the prior map. Final results conclude that the proposed technique has good performance for the gaze prediction when compared to various other top-down models that have made used in this paper. Exact view power gives the best possible outcome as it has the tendency to advance bit allocation accurately.

Keywords: ROI (Region of interest), foveation, human visualisation power, HD (High Definition).

I. INTRODUCTION

While analysing the content of images and videos one can notice that importance of the content within the frame is not equal. We are living in the age of information revolution Images may represent different aspects of our live everyday routine, events, holiday trips and arts. Like in any means of information exchange only some parts of an image contain the desired information. Indeed, due to the way images are created there is no way of full control over their content.

This peculiarities leads to the competition of information streams. Thus the way how our brain is functioning orders incoming visual information by its importance. The saccade search and selectivity process are guided by bottom-up and top-down stimulus. Top-down stimulus usually represents selection based on knowledge, for example, a subject is looking for a picture of an animal. Bottom-up stimulus is driven by properties of perceived visual information, such as high contrast, difference in orientation etc.

Ability of automatic detection of important regions can be a priceless tool for a broad variety of multimedia applications. A wide spectrum of application may benefit from separate processing of important and less important regions of an image. Thus development of a robust method for automatic detection of important regions in image may lead to significant progress in multimedia processing. As it will be shown later there already exist a number of approaches to automatically determine important regions, or as it is often called saliency. The main concern in video coding is to compress the video frames as much as possible without significant degradation of visual quality.

Real-time video streaming over wireless network is subject to impairments, either due to high error rate or bandwidth channel limitations and are unable to handle the amount of data and cannot guarantee that all the frames could meet their deadlines. Bandwidth channel limitations are considered as the major challenge to the video stream over wireless networks.

The higher the compression ratio is, the smaller is the bandwidth consumption. However, there is a price to pay for this compression: increasing compression causes an increasing degradation of the image. These are called artifacts. Compression basically means reducing image data [1]. As mentioned previously, a digitized analog video sequence can comprise of up to 165 Mbps of data. To reduce the media overheads for distributing these sequences, the following techniques are commonly employed to achieve desirable reductions in image data: reduce colour nuances within the image, reduce the colour resolution with respect to the prevailing light intensity, remove small, invisible parts, of the picture, compare adjacent images and remove details that are unchanged between two images.

There are two basic categories of compression; lossless and lossy. Lossless compression is a class of algorithms that will allow for the exact original data to be reconstructed from the compressed data. That means that a limited amount of techniques are made available for the data reduction, and the result is limited reduction of data. GIF is an example of lossless images compression, but is because of its limited abilities not relevant in video

IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 4, Issue 5, May 2016

surveillance. Lossy compression on the contrary means that through the compression data is reduced to an extent where the original information cannot be obtained when the video is decompressed.

Video source coding in general aims to preserve quality, while reducing the bit rate. In most cases the quality is defined by the extent of the error introduced by the compression independent to its position in the video sequence. This is a simplification, which disregards the Complexity of the human visual system (HVS). The perceptual quality is highly dependent on the information being transmitted at the location of the error.

A. GAZE LOCATION PREDICTION PRINCIPLE

Herein we propose and describe a context specific gaze location prediction model for broadcast football videos. Similar to [9] and [10], our model incorporates both bottom-up features and top-down visual attention cues. Our proposed prediction system has two novelties. Firstly, multiple object categories are considered, which in this context are the ball and players. In most existing gaze location prediction systems, a single object category is usually considered e.g. human faces. Secondly and most importantly, we classify the complex context into different categories according to shot type, thus allowing our model to pre-learn the task pertinence of each object category and build the prior map automatically.

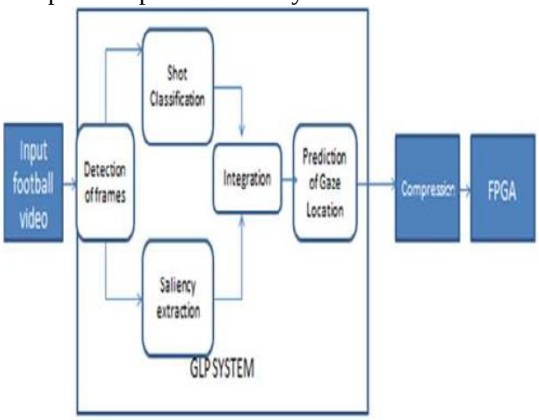

Fig 1: Basic block diagram of GLP

The Fig 1 shows the basic block diagram of the GLP. The flow of system can described as that initially the football video is taken and frames are detected from the video. Again these frames under go into two different process called shot classification and saliency. Finally the output from these two parts is combined to get gaze part of the frame. Then compression is done according to the gaze principle and finally gets into FPGA in order to speed up the process.

The Content-dependent, frame-selective compression technique which is developed wholly as a pre conditioner that can be used with existing digital video compression techniques. The technique is heavily dependent on a priori knowledge of the general content of the video which uses content knowledge to make smart decisions concerning the frames selected for storage or transmission [2]. The feature of each frame is calculated to determine the frames with the most active changes.

The selective compression depends on Region of Interest (ROI) which can be obtained from an image or video by using concepts of human visual systems (HVS) the visual system in the human body forms the eyes. The main goal associated with video source coding is to reduce the amount of data used to describe the video sequence with as limited an effect on the quality as is possible. Region of interest (ROI) coding advances this concept by allowing higher quality within interesting regions of the video sequence without increasing the total amount of data [3]. In low bit rate video transmission the necessary encoding causes reduced quality in regions of the video sequence with high detail and motion content. The perceived quality is experienced as particularly poor if the region contains information, which is important to the viewer.

ROI video coding methods increases the quality in regions of interest of the viewer at the expense of the quality in the background, compared to that when using ordinary encoding. In applications involving, for example, video conferencing, surveillance and transmission of sports, the interesting regions within the sequence can be identified. The ROI video coding consists of two main steps. The ROI must firstly be detected, which requires previous knowledge of what a human would find interesting in the sequence. The perceived quality may even be reduced if the correct ROI is not detected. Secondly the video sequence is compressed using different amounts of encoding based on the detected ROI. This is achieved by bit allocation, which controls how many bits will be allocated to the different parts of the video sequence. The idea of increasing quality within the ROI by decreasing quality in the background is called ROI video coding.

This can be divided into two separate steps. Firstly the ROI is detected by predicating the type of content in a region that attracts the viewer's gaze and communicates the greatest amount of information. Based on these characteristics the position of the ROI is extracted. In the second step the bit allocation is controlled in order to ensure that more bits are allocated to the ROI to increase the visual experience A related research area to ROI video coding is to use foveas instead of ROI's.

The key to a successful ROI video coding is to correctly predict and detect the ROI, since a falsely detected ROI gives a lower perceptual quality than for ordinary video coding at low bit rates. This is achieved either by applying a generalized or an application-based approach. In the application-based approaches the type of content present in an interesting region is predicted a prior for a particular application. These include video conferencing and videophone applications where faces are of interest, surveillance of people and vehicles or in sports applications.

The human visual system can be subdivided into two major components: the eyes, which capture light and convert it into signals that can be understood by the nervous system, and the visual pathways in the brain, along which these signals are transmitted and processed [4].From an optical point of view, the eye is the equivalent of a photographic camera. It comprises a system of lenses and a variable aperture to focus images on the light-sensitive retina. The optics of the eye relies on the physical

principles of refraction. The photoreceptors are specialized neurons that make use of light-sensitive photo chemicals to convert the incident light energy into signals that can be interpreted by the brain.

There are two different types of photoreceptors, namely rods and cones. In well-lit conditions the content of a digital image or video frame is mainly captured by cones and then delivered to the viewers' visual cortex. Over 90% of cones are located at the fovea of the human eye. The viewer's gaze location (fixation) in the picture corresponds to the part of the image/video frame which will be processed by the fovea. This part is considered to have maximum perceivability. The other parts of the picture, as the distance to the gaze location increases, will have less perceive-power in the viewer's brain.

The compression of video frames based on the foveation behaviour of the human visual system. Eye fixations on a video frame, as depicted by eye-gaze trace data, define an imaginary region of interest. The perceived resolution of the frame by the human eye depends totally on this eye-gaze (fixation) point. The resolution, then, decreases dramatically with the distance from the fovea. This behaviour of the HVS has gained interest in the image and video processing area recently especially in compression of images or video frames. We present an approach where eye-gaze trace data are integral to the compression process which has demonstrated its usefulness in yielding high compression performance. We partition a video frame into three regions: the inner-most includes a point of eye-gaze for which we apply lossless compression; an outer region which encompasses the first and for which we apply visually lossless (near-lossless) compression, and finally an outmost region where lossy compression is applied.

II. LITERATURE SURVEY

Various computational models of visual attention have been proposed. The output of these models is a visual importance map (or gaze density map), the values of which represent each pixel's probability of being the fixation point. Bottom-up approaches rely on the assumption that the higher the saliency of a certain region, the more likely it is to attract the viewer's location. Saliency is calculated from bottom-up features such as colour, intensity, orientation, motion, flicker, [9], [11], [13]. However, as mentioned above, a top-down predictor in a given context usually offers better performance than a pure bottom-up model [14], [15], [17]. Top-down visual attention depends on the content of the video and the given task. Most existing top-down gaze prediction models try to analyse the image/frame and bias the visual importance map towards certain high level object/concept locations [9], [15], [17]. In a simple context, for example a person's picture, the face can act as the top-down cue and this alone can lead to very good prediction results. But there is no common strategy in complex contexts such as broadcast football video.

Peters and Itti classify video frames into different categories -using a classifier trained on eye tracking data- and select corresponding top-down prior maps [14]. A support vector machine (SVM) was built based on previously obtained eye tracking data and a "gist" descriptor of the video frames, the latter including bottom-up pyramid features and Fourier features. The selected top-down prior map was then multiplied by saliency to form the combined prediction map. Their results indicate that the combined map offers the best prediction performance, and that the top-down prior map performs better than both the mean eye position prediction and saliency (worst performance). The context tested in [14] is rather constrained in that the locations of the top-down cues (the kart and the top part of the road) are almost fixed: the kart appears in the middle of the frame and the location of the top part of the road is also fixed. This leads to regular gaze patterns in the prior map. Considering that, in the context of BFV, small objects may appear anywhere in the same background, this method is unlikely to work. A second potential problem is that frames of different gist have to be distinguished by the SVM features.

The problem with the SVM model is that the context tested in SVM model is rather constrained in that the locations of the top-down cues are almost fixed. This leads to regular gaze patterns in the prior map in the context of BFV, small objects may appear anywhere in the same background, this method is unlikely to work. A second potential problem is that frames of different gist have to be distinguished by the SVM features. The problem with Context Prior model is that the task of looking for an object was given to the viewer's initially seems to be a drawback. Not only these limitations but their also another limitations such as they do not take into account multiple object categories in the same prediction map and the semantic aspect of the context is not investigated, which may offer important information for building the prior map.

III. APPLICATIONS

Torralba et al. propose a top down model based on object search and contextual guidance [15]. In particular they use a trained "context prior" based on eye-tracking data collected from viewers who were given the task of searching for a particular object in the picture. The "context prior" is a probability sum of all possible conditions. Their final prediction map results from Bayesian integration of saliency and this trained context prior. Torralba et al. split the bottom-up features into local and global features. Different predictions can be made when the global features of the training set are similar, by defining a clear task. For example, different prediction maps were generated for 'viewer looking at the mug' and 'viewer looking at the painting' tasks with the same picture. The fact that the task of looking for an object was given to the viewers initially seems to be a drawback. However in a good context where the viewer's attention is guided by certain objects this method may prove useful.

Inspired by Torralba's approach, Boccignone et al. proposed a new Bayesian integration model for gaze location prediction [17]. In their approach the input frame is analysed hierarchically. At the higher level, visual

attention is attracted to the location of objects. At the lower level, it is affected by saliency around the objects. Different from [14] and [15], Boccignone et al. did not use any pre-learnt context information. Instead of generating a descriptor for the whole frame, they presume that the top-down control is due to the presence of human face(s) in the image/video frame. The top-down prior distribution is modeled by a Gaussian around the center of detected face(s). The performance of such models depends on how the prior map is defined. Boccignone et al. did not examine the case where objects of multiple categories are present. In such cases the simple face-based prior map may not be enough, which may explain why their model did not perform well with the "News" sequence, where three objects are present: a man, a woman and a big screen behind them [17]. Their results however are better than those of the bottom-up model.

IV. CONCLUSION

Through the proposed system predicts the gaze part in the football video. From the concept of the region of interest the required part is obtained from the input image by using the saliency and shot classification. The obtained gaze part from the system matches almost same as that can be predicted with human eye.

This concept of predicting the gaze can be applied to television news where the chance of reducing the redundancy is very high. Here the gaze part forms the news reader face, highlights which are scrolling. These parts can be carefully extracted and easily predicted for the gaze location concept.

REFERENCES

[1] Winkler S. Digital Video Quality: Vision Models and Metrics. New York, NY, USA: Wiley, 2005.
[2] Komogortsev OV, Khan JI. Eye movement prediction by Kalman filter with integrated linear horizontal oculomotor plant mechanical model. Proc. Symp. Eye Tracking Res. Appl. 2008; 229–236.
[3] Feng Y, Cheung G, Tan WT, Ji Y. Hidden Markov model for eye gaze prediction in networked video streaming. Proc. IEEE ICME, Jul. 2011; 1–6.
[4] Frintrop S, Rome E, Christensen H. Computational visual attention systems and their cognitive foundations: A survey. ACM Trans. Appl. Perception 2010; 7(1): 1–6.
[5] Itti L. Automatic foveation for video compression using a neurobiological model of visual attention. Image Processing, IEEE Transactions on 2004; 13(10): 1304–1318.
[6] Wen-Fu L, Tai-Hsiang H, Su-Ling Y, Chen HH. Learning-based prediction of visual attention for video signals. Image Processing, IEEE Transactions on 2011; 20(11): 3028 –3038.
[7] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach.Intell. 1998; 20(11): 1254–1259.
[8] Peters R, Itti, L. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. ProcIEEE CVPR, Jun. 2007; 1-8.
[9] Torralba A, Oliva A, Castelhano M, Henderson J. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychol. Rev 2006; 113(4): 766–786.
[10] Boccignone G, Marcelli A, Napoletano P, Di Fiore G, Iacovoni G, Morsa S. Bayesian integration of face and lowlevel cues for foveated video coding. IEEE Trans. Circuits Syst. Video Technol. 2008; 18(12): 1727–1740.
[11] Ekin A, Tekalp A, Mehrotra R. Automatic soccer video analysis and summarization. IEEE Trans. Image Process 2003; 12(7): 796–807.
[12] Li L, Zhang X, Hu W, Li W, Zhu P. Soccer video shot classification based on color characterization using dominant sets clustering. Proc. 10th Pacific Rim Conf. Multimedia, Adv. MultimediaInf. Process., Dec. 2009; 923–929.
[13] Darrell T, Gordon G, Harville M, Woodfill J. Integrated person tracking using stereo, color, and pattern detection. Int. J. Comput. Vis 2000; 37(2): 175–185.
[14] R. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in Proc. IEEE CVPR, Jun. 2007, pp. 1–8.
[15] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search." Psychol. Rev., vol. 113, no. 4, pp. 766–786, Oct. 2006.
[16] S. Lee, G. J. Kim, and S. Choi, "Real-time tracking of visually attended objects in virtual environments and its application to LOD," IEEE Trans. Visualizat. Comput. Graph., vol. 15, no. 1, pp. 6–19, Jan. 2009.
[17] G. Boccignone, A. Marcelli, P. Napoletano, G. Di Fiore, G. Iacovoni, and S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 12, pp. 1727–1740, Dec. 2008.