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 model 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).

1. 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 can be in 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. 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 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.

2. CONVENTIONAL METHOD

Visual attention has become popular over the past decade, primarily for two reasons. First, models make testable by experimentalists as well as theoreticians. Second, models have practical and technological applications of interest to the applied science and engineering communities .
Gaze data can be extracted and predicted using interactive tools in most cases such tools are either not available or their usage is not applicable. Various computational models of visual attention have been developed aiming to predict the viewer's gaze location, there are basically two attention models such as

1) Bottom-up approach
2) Top down approach

The bottom-up aims to compute several low level features in parallel and fuse the individual saliencies into an overall gaze density map. Top-down visual attention models also exist. However, the cues in these models are based on task type, prior knowledge and context information. In other words, top-down visual attention models are context related or even context specific. It is widely agreed that using top-down information may outperform a pure bottom-up model in predicting the viewers' visual attention.

Top-down attention is driven by cognitive factors such acknowledge, expectations and current goals .Other names for top-down attention are endogenous, voluntary, or centrally cued attention. There are many examples of this such as car drivers are more likely to see the petrol stations in a street and cyclists notice cycle tracks. If you are looking for a yellow highlighter on your desk, yellow regions will attract the gaze more readily than other regions. The eye movements depend on the current task for the same scene an unexpected visitor which shows a room with a family and a person entering the room, subjects got different instructions such as estimate the material circumstances of the family, what are the ages of the people, or simply to freely examine the scene. Eye movements differed considerably for each of these cases. Visual context, such as the gist or the spatial layout of objects, also influence visual attention in a top-down manner. In psychophysics, top-down influences are often investigated by so called cueing experiments. In these experiments, a cue directs the attention to the target. Cues may have different characteristics they may indicate where the target will be, for example by a central arrow that points into the direction of the target, or what the target will be, for example the cue is a similar or exact picture of the target or a word or sentence that describes the target search for the black, vertical line.

2.1 COMPUTATIONAL MODELS

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, and flicker. However, as mentioned above, a top-down predictor in a given context usually offers better performance than a pure bottom-up model 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.

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. Several gaze location prediction models has been proposed from the long-time some these models are discussed below.

2.1.1 SVM Model

In SVM model we classify video frames into different categories using a classifier trained on eye tracking data- and select corresponding top-down prior maps. 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 top-down component of this gaze-prediction model is designed to learn to associate the gist of an image with likely task-relevant locations under the current task. The model proceeds in two stages.
First, in the training stage we build a training set using a leave-one-out approach when one clip is used as the test clip, the training set is formed from the remaining 23 clips, and this procedure is repeated for each of the 24 clips. From each frame in the training clips we collect the recorded eye position of the observer who played the game as the clip was recorded, and we also compute from the entire image a low-dimensional feature vector that is intended to be diagnostic of the image's gist two approaches for generating such feature vectors are described below. Second, in the testing stage, we pass the set of observed eye positions and corresponding feature vectors to a learning algorithm, which, after training, can take feature vectors extracted from new test frames and generate eye position prediction maps.

2.1.2 Context Prior

Context prior is a top down model based on object search

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

and contextual guidance. 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.

2.1.3 BI MODEL

This model can be considered as an advanced version of the context prior model. In this 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. In this model they 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 faces in the image/video frame. The top-down prior distribution is modelled by a Gaussian around the center of detected faces. The performance of such models depends on how the prior map is defined.

2.2 DISADVANTAGE OF THE PREVIOUS MODELS

There are limitations for the previously proposed models when comes to the context of football videos. 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. 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.

The problem with the BI model is that they 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. Not only these limitations but their also another limitations such as

1) They do not take into account multiple object categories in the same prediction map.
2) The semantic aspect of the context is not investigated, which may offer important information for building the prior map.

The another main drawback in the existing system comes in the shot classification model i.e.; the entire video of the football match is classified into different shots based on the green colour present in the different frames of the video. By using predefined RGB threshold the shot classification is done. The green pixels in each frame can extracted for this purpose. The process can be accurately done where no problems such as rainfall, shadows, change in the sunlight, and artificial pitch, when these problems occur the green colour cannot be exactly predicted as that of the natural case.

3. PROPOSED METHOD

The architecture of the proposed system is shown in below figure. 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.

Visual saliency is a broad term that refers to the idea that certain parts of a scene are pre-attentively distinctive and create some form of immediate significant visual arousal within the early stages of the Human Visual System. The salience (also called saliency) of an item be it an object, a person, a pixel, etc. is the state or quality by which it stands out relative to its neighbours

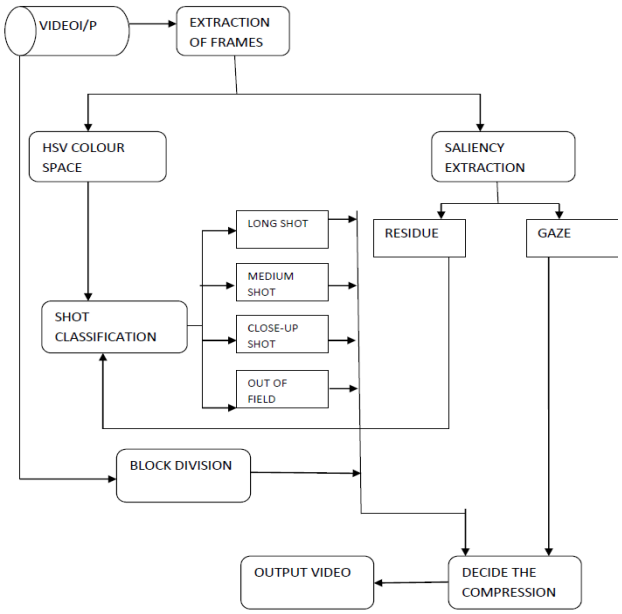

Fig 3.1 Architecture of the proposed system

Shot class information, when combined with other features, conveys interesting semantic cues. Motivated by this observation, we classify shots into four classes. They are

1) Long shots
2) Medium shots
3) Close-upshots
4) Out of field shots

Finally from the saliency part and shot classification part we will extract the features required for the gaze location. The saliency features and the face can also be detected from the close up shot of the shot classification technique. These features extracted from the frames are finally useful for the compression which will be done by the region of interest. Finally these GLP features are useful in many ways for real time broadcasting the football video, accurate prediction of a viewer's gaze location has the potential to improve Bit allocation, rate control, error resilience, quality.

4. RESULTS

The results of the proposed system can showed in the following figures.Which consists of original image and its saliency map followed by saliency output and resude part for different transformations.

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

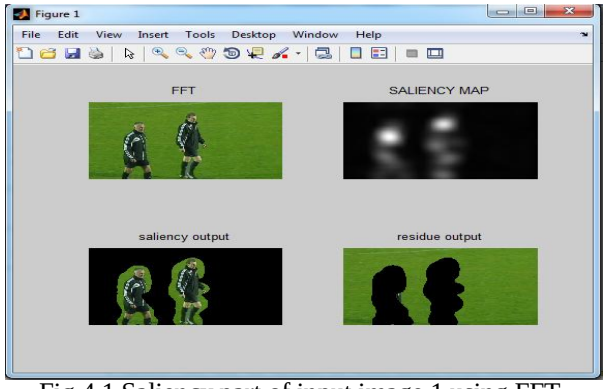

Fig 4.1 Saliency part of input image 1 using FFT

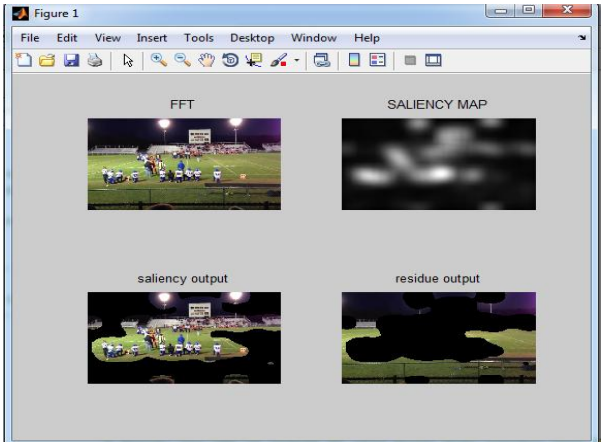

Fig 4.2 Saliency part of input image 2 using FFT

By applying our proposed technique we can compress the required file size without much effect in the visual quality as shown from the images below

Fig 4.3 Original image (883 kb)

Fig 4.4 Compressed image (655kb)

5. CONCLUSION

Through our proposed system we predicted 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 our system matches almost same as that we predict 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.