AUTHOR(S): CHILUKAMARI, J., KANNANGARA, S. and MAXWELL, G.

TITLE: A DCT based in-focus visual saliency detection algorithm.

YEAR: 2013

Publisher citation: CHILUKAMARI, J., KANNANGARA, S. and MAXWELL, G. 2013. A DCT based in-focus visual saliency detection algorithm. In Proceedings of the 3rd IEEE international conference on consumer electronics - Berlin (ICCE-Berlin 2013), 9-11 September 2013, Berlin, Germany. New York: IEEE [online], pages 1-5. Available from: https://doi.org/10.1109/ICCE-Berlin.2013.6698003

OpenAIR citation: CHILUKAMARI, J., KANNANGARA, S. and MAXWELL, G. 2013. A DCT based in-focus visual saliency detection algorithm. In Proceedings of the 3rd IEEE international conference on consumer electronics - Berlin (ICCE-Berlin 2013), 9-11 September 2013, Berlin, Germany. New York: IEEE, pages 1-5. Held on OpenAIR [online]. Available from: https://openair.rgu.ac.uk

A DCT Based In-Focus Visual Saliency Detection Algorithm

Jayachandra Chilukamari, Sampath Kannangara and Grant Maxwell
School of Engineering, Robert Gordon University, Aberdeen, UK

***Abstract*—A novel low-complexity visual saliency detection algorithm for detecting visually salient regions in images based on camera focus is presented. This fast in-focus region detection algorithm detects salient-frequencies present in in-focus areas using the characteristics of Discrete Cosine Transform coefficients. The performance of this algorithm is validated against five state-of-the-art saliency detection algorithms. The results show that this algorithm consistently outperforms other saliency detection algorithms in terms of complexity and prediction accuracy.**

***Index Terms*— in-focus, saliency, DCT, visual, attention**

I. Introduction

The nature of the Human Visual System (HVS) is to attend visually salient regions in an image whilst ignoring the regions of less interest. Determining these regions computationally has many applications in the areas of image and video compression, human robot interaction, face segmentation and tracking, salient object detection and visual tracking [1-5]. In the past decade, many different saliency models have been proposed to detect these salient regions in images and videos. Majority of them use bottom-up features in an image to detect the salient regions as these features are easier to identify, control and model. The bottom-up features are specific features that "pop out" from the scene to grab the viewer's attention. Some of the bottom-up features are colour, intensity, orientation, corners, flicker and curvature. The Phase Quaternion Fourier Transform (PQFT) [1] is a bottom-up saliency model with low complexity, however, it suffers from incorrectly detecting highly textured backgrounds. Itti's model [6] and Graph Based Visual Saliency (GBVS) [7] can powerfully predict human fixations. However, high computational cost is a limitation of these models. The saliency models such as PQFT, Itti's model and GBVS that incorporates bottom-up features do not consider cognitive "top-down" features. Some of the top-down features are faces of people, animals, prior knowledge about the target, expectation and goals of observers [8]. Among these top-down features, expectation, knowledge and goals of observers differ between individual subjects and are difficult to model mathematically. A few saliency models have incorporated top-down features along with bottom-up features with a goal of improving the prediction accuracy. Some saliency models have utilized face detection to improve the prediction accuracy [4]. Although the human visual system attends to faces more than other features, algorithms relying mostly on face detection perform poorly when other high level features such as text, hands and objects are encountered. The Saliency Using Natural statistics (SUN) [9], [10] model incorporates both bottom-up and top-down features. However, this model requires too many parameters to be estimated, resulting in high computational complexity. Although, Context Aware Saliency (CAS) [11] considers some low and high level features, significant higher computational complexity compared to other models is still a weakness of this model. Most of the authors use a number of channels (features) to determine the visually salient regions and thereby increasing complexity. Judd *et.al* [4] use several features (face detection, horizontal line detection, gist, objects, and people) to improve the prediction accuracy of their visual attention model. Their approach is considered to be a justifiable option from the point of accuracy; however, a higher number of features increase the complexity of their model. A computationally developed visual attention model can imitate the HVS only when it considers both top-down and bottom-up features. However, models designed with too many bottom-up and top-down features results in highly complex algorithms [12].

Typically, during video or image capture, the viewer attention is lead to a specific region of a video frame (or an image) by bringing the region into focus (in-focus). These in-focus regions may constitute bottom-up or top-down features of an image depending on the context of the scene in the video/image. An "in-focus" region in an image contains more high-frequency content which is of interest to the HVS when compared to an "out-of-focus" region [13]. Therefore, the objective of this work is to develop a fast visual saliency detection algorithm based on the following hypotheses:

(a) Salient regions of an image or video frame tend to be in-focus areas.

(b) In-focus areas of an image contain some significant spatial frequency components compared to out-of-focus areas.

II. In-Focus Detection

The authors of [14], [15] developed algorithms for computing the bottom-up features using the Discrete Cosine Transform (DCT) coefficients of an image. Although these algorithms are limited to detecting only bottom-up features, they highlight the use of the DCT to detect salient frequency components of an image. Moreover, DCT is widely used in video and image compression and is therefore implemented in most video and image processing products. The above facts motivated us to use the DCT (to identify salient frequencies in in-focus regions) in the development of In-focus visual saliency detection algorithm.

Fig.1. Observed in-focus region

A. Spatial Frequency Composition

As per initial hypotheses, the in-focus regions are sharp and are highly appealing to the human visual system. Therefore, in-focus regions would contain some frequency coefficients (spatial frequencies) that hold significantly large values when compared to out-of-focus regions.

An investigation was carried out to identify the salient frequency components of in-focus regions of an image compared to out-of-focus regions. Random images were selected and converted to YCbCr colour space. In-focus regions of the images were manually observed and identified using the luminance (Y) component of the image as shown in Fig.1. The face and body area enclosed by the outline is observed to be the in-focus region (region of interest).

The spatial frequency composition of in-focus (foreground) and out-of-focus (background) regions were analyzed as follows. The entire image is divided in to 8x8 blocks and the DCT of each 8x8 block is calculated. The foreground 8x8 superblock $FS_{8\times8,}$ background superblock $BS_{8\times8}$ and full image superblock $IS_{8\times8}$ are calculated as follows:

$$[FS]_{8x8} = \frac{1}{M}\sum_{c=1}^{M}\left|[f_c]_{ij}\right| \quad (1)$$

$$[BS]_{8x8} = \frac{1}{N}\sum_{c=1}^{N}\left|[b_c]_{ij}\right| \quad (2)$$

$$[IS]_{8x8} = \frac{1}{(M+N)}\sum_{c=1}^{M+N}\left|[I_c]_{ij}\right| \quad (3)$$

Where, $i \in (0,1,..,7)$, $j \in (0,1,..,7)$. f_c, b_c and I_c denote foreground, background and image (all) DCT blocks respectively. M is the number of foreground 8x8 DCT blocks and N is the number of background blocks. (M+N) denotes the total number of 8x8 blocks of an image.

Essentially, the in-focus region is manually selected and the foreground superblock is calculated by taking the mean absolute value of each coefficient of the 8x8 DCT blocks that correspond to the foreground region. Similarly the background superblock is calculated. Finally the image superblock is calculated by taking the sum of the foreground and background superblocks and dividing it with the total number of 8x8 blocks in the entire image. The DCT coefficients in all the three super

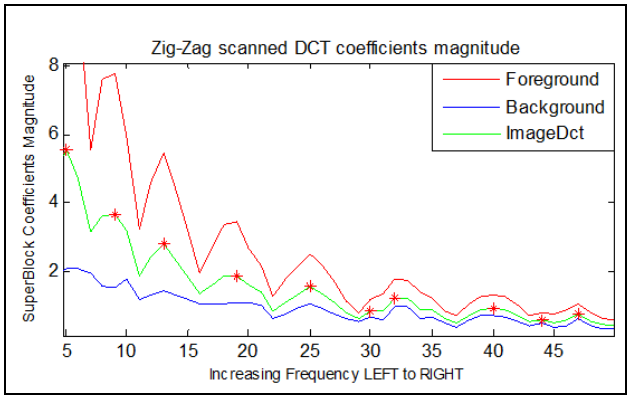


Fig.2. Superblock coefficients magnitude vs Zig-zag scanned frequencies

blocks are selected using zig-zag scanning method so that the low frequency components precede the high frequency components. These zig-zag scanned coefficients are subsequently converted into a one dimensional vector. The average frequency coefficient magnitudes of foreground, background and the overall image superblocks are plotted on a graph to determine the relationship between in-focus and out-of-focus coefficients. The DCT frequency coefficient amplitude pattern of a sample image is shown in Fig.2. The graphs reveal that the peaks of spatial frequency amplitude waveform of the foreground (in-focus) and the entire image almost coincide. The maximum magnitude difference between background and the foreground (also full image DCT) occurs at these peaks. This amplitude differences tends to be significant within a band of frequencies which excludes both very low and high frequencies. It also reveals that the peak magnitude frequencies in the in-focus regions are absent in out-of-focus regions. Further, these peak frequencies have a significant presence in the overall image DCT. Therefore, the peak frequency coefficients in the whole image can be used to identify in-focus regions. Experiments with a number of images revealed that very low frequencies (1-4) corresponding to gradual changes, do not show a significant difference between in-focus areas and out-of-focus areas. Very high frequencies (50-64) typically correspond to noise in the image. Noise is very high frequency information which occurs during the video capture and transmission phase. Therefore, excluding these frequencies limits the influence of noise on saliency computation. Hence, zigzag scanned coefficients are band-pass filtered. This filter eliminates these frequencies by inhibiting some of the very high and low frequency irrelevant DCT coefficients. All frequency coefficient positions that correspond to the peaks within the band of frequencies of the image superblock coefficients are identified and stored. These are the salient spatial frequencies present in the in-focus area of the image. Figure 3 shows the saliency map generated by plotting the sum of salient frequency coefficients (sum of peak frequencies present in image DCT superblock) in each 8x8 block of the image. It is evident that the saliency map clearly detects the salient in-focus area of the image as observed originally.

Fig.3. Image saliency map

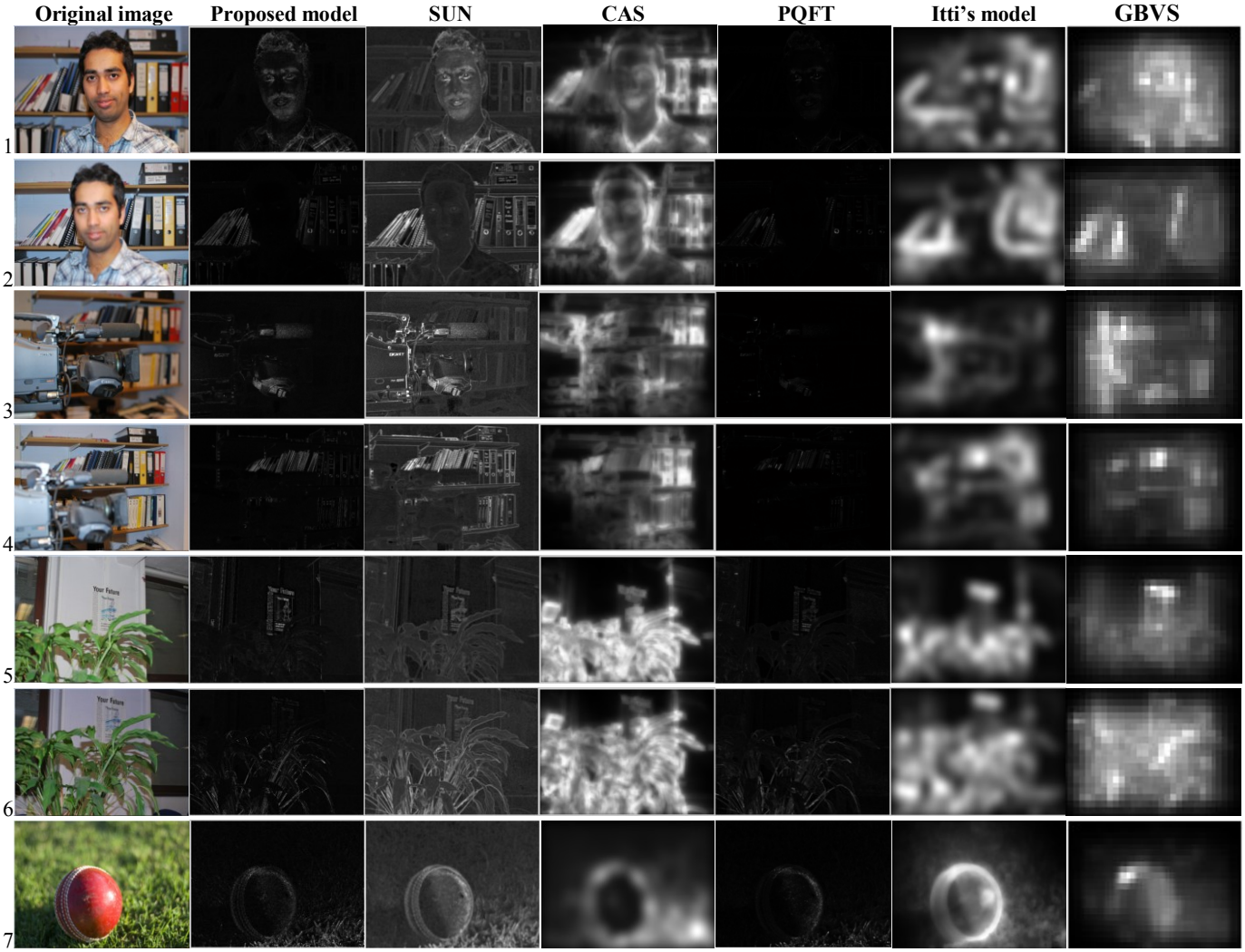
B. In-focus Visual Saliency Detection Algorithm

The complete novel algorithm for in-focus visual saliency detection model can be summarized as:

1. Extract the luminance component of an image.
2. Perform DCT across all 8x8 blocks of the image.
3. Calculate the image superblock and select the frequencies using zig-zag scan method.
4. These zig-zag scanned DCT coefficients are band pass filtered.
5. The peaks of the image DCT are obtained and the frequency coefficients corresponding to these peaks are summed up and displayed as a saliency map.

III. Results and Discussion

A qualitative validation of the proposed algorithm was carried out by comparing the results for 50 images against popular saliency models namely PQFT [1], GBVS [7], SUN [9], [10], Itti [6] and CAS [11]. Results for 13 sample images from the database are shown in Fig.4.



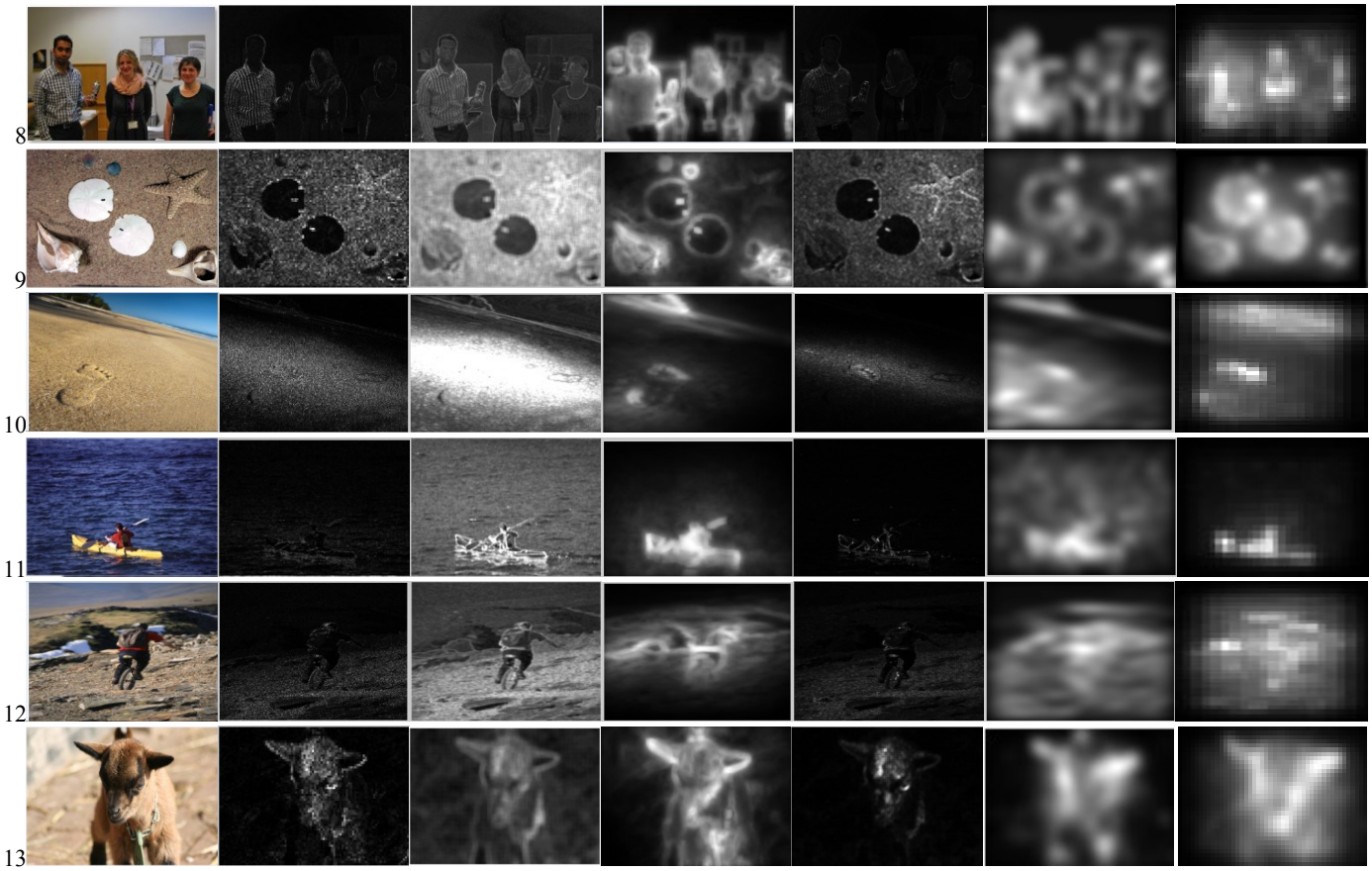


Fig.4. Qualitative comparison of saliency maps for sample images

Table 1. Complexity comparison of algorithms

Visual Attention Model	Complexity(secs)
PQFT	0.6025
Proposed model	0.9044
GBVS	1.4544
SUN	2.3081
Itti's model	4.3423
CAS	24.6172

In the first pair of images from the top in the Fig.4, the first original image focuses on the face. The second image focuses on the background (face is blurred). The proposed model is clearly able to detect the salient in-focus regions in the two images. The other models, especially SUN and CAS detect both face and the background as salient regions. Itti's model, PQFT and GBVS could not detect the face as they ignore top-down features. The first three pairs of pictures are captured by alternating the focus region between foreground object and the background region in order to verify the detection ability of the proposed model. A good saliency detector should have the ability to detect salient regions with varying backgrounds characteristics. In image 7 the proposed model detects the in-focus ball and grass region in the immediate vicinity of the ball (although the rest of background contains grass). The PQFT model marginally detects this; however Itti's model, GBVS, SUN and CAS detect both the ball and the most of the background as salient. In the eighth image with cluttered background the proposed model accurately detects high level features (3 faces & hands) and low level feature (bottle). The 10th image contains highly textured background (sand) which has high spatial frequencies throughout. However, the camera is clearly focused around the foot imprint. The proposed model clearly distinguishes the dominant frequencies within the in-focus region from the noisy frequencies contained in the background as evidenced by the saliency map. Other models clearly fail to detect the salient region. The last three images in Fig.4 are every day scenes with reasonable complexity. The proposed model effectively detects the salient regions compared to the other state-of-the art models.

A visual attention algorithm is always a trade off between time complexity and the prediction accuracy it achieves. The complexity of all tested saliency models were analyzed on a computer running with Microsoft windows 7 ultimate with 16 GB RAM and Quad core 3.40 GHz Intel core i7-2600K CPU. The average time required to compute a saliency map is calculated over 20 images with resolution 720x480 from our database. All the algorithms are developed in MATLAB environment. The complexity comparison is given in the Table

1. It is evident that PQFT is fastest among the models. However it lacks prediction accuracy when compared to the proposed model. Although, our model though ranks second in terms of complexity, with respect to the prediction accuracy it outperforms all the state-of-the art visual saliency models.

IV. Conclusion and Future Work

In this paper, a novel visual saliency model which detects in-focus regions is proposed. The model is qualitatively validated against five state-of-the-art visual saliency models. The model is tested with different types of images such as, same scene with different regions in focus, images with cluttered backgrounds, images with highly textured backgrounds and everyday scenes. The results show that the proposed model outperforms the state-of-the-art models in terms of complexity and prediction accuracy. One of the limitations of the current proposed model is that the saliency maps are $1/64^{th}$ of the original image resolution. Future developments include generating variable resolution saliency maps, verifying the model's performance over synthetic or psychological patterns and implementation for real time videos. The improved algorithm will be validated through complete qualitative and quantitative analysis.

References

[1] G. Chenlei and Z. Liming, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *Image Processing, IEEE Transactions on,* vol. 19, pp. 185-198, 2010.

[2] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on,* vol. 13, pp. 1304-1318, 2004.

[3] H. Li and K. N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *Journal of Visual Communication and Image Representation,* vol. 19, pp. 320-333, 2008.

[4] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 2106-2113.

[5] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1007-1013.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 20, pp. 1254-1259, 1998

[7] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.

[8] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience,* vol. 3, pp. 215-229, 2002.

[9] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top-down saliency using natural statistics," *Visual Cognition,* vol. 17, pp. 979-1003, 2009.

[10] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision,* vol. 8, 2008.

[11] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, pp. 1915-1926, 2012.

[12] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.,* vol. 7, pp. 1-39, 2010.

[13] S. Y. Lee, S. S. Park, C. S. Kim, Y. Kumar, and S. W. Kim, "Low-power auto focus algorithm using modified DCT for the mobile phones," in *Consumer Electronics, 2006. ICCE '06. 2006 Digest of Technical Papers. International Conference on*, 2006, pp. 67-68.

[14] J. Yiwei and X. De, "A Visual Attention Model Based on DCT Domain," in *TENCON 2005 2005 IEEE Region 10*, 2005, pp. 1-5.

[15] Y. Fang, W. Lin, Z. Chen, C.M. Tsai, and C.W. Lin, "Video saliency detection in the compressed domain," presented at the Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 2012.