International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2551-2557


# An Algorithm and Architecture for HEVC Video Compression

**K. Bhanu Rekha[1] and Dr. Ravi Kumar AV[2]**

*[1]Department of Electronics and Electrical Engineering, Sir M.V.I.T, Bangalore Karnataka, 562 157, India.*
*[1]E-mail: bhanublossom2014@gmail.com*
*[2]Department of Electronics and Electrical Engineering, SJBIT, Bangalore, Karnataka, 560060, India.*
*[2]E-mail: avr187@gmail.com*

**Abstract**

The video coding technique as modified HEVC coding based on saliency features is proposed tries to reduce attention-grabbing artifacts and provides high quality in areas where viewer's attention is drawn. The process allows visual saliency to increase quality in ROI parts and reduce in low-quality parts of the frame. This approach allows achieving the same subjective quality as contending state of the art methods at a low bit rate. The eye tracking datasets such as DIEM and SFU can be utilized to test the saliency algorithm for video and evaluate video quality assessment. The model provides better efficiency in terms of AUC, NSS and average PSNR.

**Keywords:** Saliency detection, HEVC, PSNR,AUC,NSS.

## INTRODUCTION

Visual attention in humans is a set of strategies in early stages of vision processing that filters the stream of data collected by eyes. Visual attention enables the visual system to parse complex and highly cluttered scenes rapidly. The complexity of visual scene analysis is reduced in Human Visual System (HVS) by automatically shifting the Focus of Attention (FOA) across the scene [1]. This ability allows the brain to restrict high-level processing of a scene to a relatively small part at any given time. Regions that draw attention are called salient and are subject to further processing for a high-level perception of the scene. An application such as image segmentation, recognition, video compression, video and image assessment uses saliency detection.

Itti et al. [1] extracted low-level features from still images such as intensity, color, and orientation are efficient in detecting saliency. The key to perceiving the around us is visual attention. The center-surround difference was used in their model to build the conspicuity maps. Finally, all the three conspicuity maps are linearly integrated to obtain saliency map. Two other features were added( i.e motion and flicker) into image saliency model for moving images[2]. Further, several video saliency models were proposed using advanced heuristics methods [3-9]. The most saliency approach is based on a bottom-up approach where the features are extracted to build a saliency map, but it often does not match the actual eye movement. Judd el al.[10] proposed a top down approach to build saliency model to predict fixation locations. To match the actual eye movements Rudoy el al [11] validated their method using gaze tracked video datasets and show it outperformed state of the art. A probabilistic multi-task learning for video saliency detection was developed, in which the stimulus function is learned for top-down attention models [12].

The formally approved the state-of-the-art new video coding standard in April 2013 by ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group is High-efficiency video coding (HEVC) [13]. It achieves double coding efficiency improvement over the other standards. Interestingly HEVC encoder can be explored as a feature extractor to efficiently predict video saliency in the compressed domain. The Compressed-domain approaches in contrast to pixel-domain methods make use of the data from the compressed video bitstream, such as motion vectors (MVs), block coding modes (BCMs), motion-compensated prediction residuals or their transform coefficients, etc. The main motivation of using HEVC features in our saliency detection is:

1. It takes advantage of HEVC encoder, to effectively extract for video saliency detection.
2. It takes the advantage of HEVC compressed domain data, so small part of data needs to be processed and decoded compared to pixel domain methods.This will reduce accuracy, but also reduces computational time and storage.

There are hardly any methods [14]–[19] proposed for detecting video saliency in the compressed domain for video coding standards such as H.264/AVC and MPEG standards. Among these methods, Ma and Zhang [14] used the magnitude of an object's motion and its duration as cues for detecting salient regions. Two new rules are included for the PMES model developed new model called Motion Attention Model (MAM) [15]. According to this new model false detection is detect if the MVs of a moving object is large and their angles are disjointed than the motion vector information is not reliable.The Perceptual Importance Map based on Zen method [16] (PIM-ZEN) computes the saliency map based on MVs and DCT values.Liu et al. [17] proposed a method for detecting saliency using MVs based on the Motion Center-Surround Difference Model (MCSDM).In the Gaussian Center-Surround difference (GAUS-CS) model [18] and the Parameterized Normalization, Sum and Product - Center-Surround difference (PNSP-CS) model, saliency is identified through still and dynamic saliency maps. The Motion Saliency Map - Similarity Map (MSM-SM) model [19] consists of two approaches for generating the final motion saliency map. In the first appr, the edge strength of each block is computed based on low-frequency AC coefficients of the luma and chroma channels. In the second approach, the dissimilarity of DC images of the luma and chroma channels among co-located blocks over the frames is calculated by entropy.

The rest of this paper is organized as follows. In Section 2, we review the related work on video saliency detection. In Section 3, we present our experimental results comparing eye tracking database as well as the analysis and observations on our database. In light of such analysis and observations. Finally, Section 4 concludes this paper.

## PROPOSED SALIENCY ESTIMATION ALGORITHMS

For Quality Saliency Assessment of a video, our model is divided into four components to assess saliency map in a spatial and temporal module.SSM (Spatial Saliency Map) is built using a convex approximation of IKN model, whereas TSM (Temporal Saliency Map) is used to predict saliency using Motion Vector Entropy and Smooth Residual feature while eliminating camera motion by global motion compensation to obtain high compression. The prediction of the quality saliency assessment map for a low-resolution video can be achieved in four components which are as follows:

### IKN convex approximation

Let N be a block within a current frame. Spatial Saliency map S*spatial*(N) will be computed by approximation of the IKN model by the convex approach. The normalized frequency (in radians/pixel) of the original image at level 0 is [0, π] in both horizontal and vertical directions, since the normalized frequency spectrum at level L of the pyramid is in the range [0, $\pi/2^L$]. Hence, the normalized frequency spectrum at levels 4 and 8 will be, respectively, in the range [0, π/16] and [0, π/256]. The centre-surround feature map at centre level is at level C ∈ {2, 3, 4} and surround level S = C+β, with β ∈ {3, 4}. The centre surround difference is computed by interpolating the surround level S to the centre level C followed by point-by-point subtraction. Hence, the normalized frequency spectrum of the centre-surround feature map at centre level C and surround level S will be, in the range [$\pi/2^s$, $\pi/2^c$]. All the computed centre-surround feature maps are resized to the size of level 4 to compute the conspicuity map of each feature channel. Hence, the upper limit of the normalized frequency spectrum of the obtained conspicuity map is capped by π/16. Since the smallest surround map is at level 8, we conclude that the IKN model uses the image content in the normalized frequency range [π/256, π/16] to construct the saliency map, as already observed in [1, 20]. We think of the image signal in the normalized frequency range [π/256, π/16] as the "signal," and the signal in the remaining part of the spectrum, [0, π/256) ∪ (π/16, π], as "noise," or "undesired signal." The Wiener filter is the optimum linear filter for extracting the signal from noise.[21]

To extract a block from the video frame of desired size windowing method used and then its 2-D DCT evaluated. The resultant DCT can be expressed as$X_d(p,q)$. Assume that another resultant DCT can be denoted as $X_w(p,q)$ which covers $[0-\pi/256) \cup (\pi-16/\pi]$ frequency band. Weiner transfer function in DCT domain can be expressed in equation1 as

$$T(p,q) = \frac{Xd^2(p,q)}{Xd^2(p,q) + Xw^2(p,q)}, \quad (1)$$

Where, $T(p,q)$ is the $(p,q)th$ Weiner Filter DCT Coefficient. For a macroblock (MB)$Z$, the energy of Wiener-filtered signal $X_Z^n$ can be described as for spatial saliency in equation2 as ,

$$Sspatial(N) = \sum_{(p,q)}\left(X_Z^n(p,q)\right)^2 = \sum_{(p,q)} T^2(p,q){X_Z}^2(p,q) \quad (2)$$

### Motion Compensated Global Saliency:

Motion vector(MV) during video coding, is accumulated by two factors: the camera motion and object motion. It is observed [22] that in a video, moving objects may receive extensive visual attention, while static background normally draws little attention. It is thus necessary to distinguish dynamic objects and still background. Unfortunately, MVs of static background may be as large as moving objects, due to the camera motion. On the other hand, although the temporal difference of MVs is able to make camera motion negligible for static background, it may also miss the moving objects. Therefore, the camera motion has to be eliminated from calculated MVs, to estimate object motion for saliency detection.

In [23], it was observed that the accuracy of the model degrades on scenes with camera motion. Global motion is associated with the motion of background caused by the camera motion. Global motion estimation (GME) to deal with the issues of separating foreground (outlier) from the background (inlier). In this approach GME is performed on the complete dataset, the estimation process removes the data that cause large errors and are declared as outliers. The 8-parameter perspective model is described by a vector of its parameters, V=[v0,v1….v7]. The perspective transformation for given (x,y) and (x',y') as the coordinates in the current and the reference frame, respectively, is defined in the equation 3 and 4 as:

$$x' = \frac{v0x + v1y + v2}{v6x + v7y + 1} \quad (3)$$

$$y' = \frac{v3x + v4y + v5}{v6x+v7y+1} \quad (4)$$

The X and Y components of the MV field at in the current frame are given by: $MV^x(x,y;v)=x'-x$, $MV^y(x,y;v)==y'-y$

A given vector of motion parameters usually exhibit strong spatial correlation. This property used to reject the outlier by cascading rejecters approach.

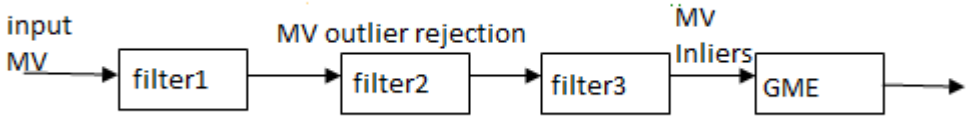


**Figure 1:** MV outlier removal cascade

The outliers are removed from an input motion vector filed using a chain of three filters to speed up GME, as shown in Fig. 1.

The operation of filtering is summarized as follows:

1) Symmetrically extend the motion vector field across the frame and indicate all MVs as inliers.
2) For each inlier, find the weighted count. Note that previously declared outlier MVs are included in the neighborhoods of inlier MVs.
3) Based on their weighted counts sort MVs in descending order.
4) MVs at the bottom of the sorted list declared as outliers.
5) If, then stop. Otherwise, set, and move on to the next filter, repeating steps 2-5.

The GME algorithm is followed by GMC(global motion compensation) to obtain GMC-MV per 4x4 block.

**Motion Vector Entropy**

For each block N(4x4), the average magnitude entropy of all GMC-MVs in it is taken as its motion saliency S*motion*(N).

The motion vector entropy can described by the equation 5 s follows,

$$mve(x) = -\log M^{-1} \sum_{j=K(\theta(x))} m_j . M^{-1} . \log m_j . M^{-1} \quad (5)$$

Here, $x$ represent a block and $\theta(x)$ represents a motion cube linked with $4 \times 4$ block, $m_j$ is the number of Motion Vectors present in the bin index$j$.The parameter $\log M^{-1}$ in (1) lies between 0 (min) and 1 (max).

The saliency map in [24] was constructed by using MVs only:

$$S = Zm \odot Zg$$

Where, $\odot$ element wise multiplication of Zm motion magnitude to Zg global angle.

Matrix Em was obtained from the average motion magnitude over a constant duration of a shot after removing outliers.

**Smoothed Residual Feature**

If the best matching block in the reference frame is not found in the current frame, it results in large motion compensation prediction residual. The block translation has failed to predict the motion of the current block. This is either due to higher order motion or due to dis-occlusion.The dis-occulusion occurs if a new object enters the scene or it is found behind another new object.Large DCT residual can detect dis-occulusion.

We define Smooth Residual feature (SRF) for any macroblock, as the normalized to the range[0:1] of the quantized transformed prediction residual of the MB.The size of residual depends on the normalization function which is a non-zero variable. Residual Normalization feature can be expressed in equation (6) as,

$$RF(N) = 256^{-1} \parallel Z \parallel_0 \quad (6)$$

Here, Z represents the residual transformation of 16 ×16macro blocks. Then, the spatial smoothness of residual normalization feature map calculated using the 3×3 filter and temporal smoothness obtained by utilizing average moving filter over total existing frames. Therefore, an SRF (Smooth Residual Feature) map generated using N( 4 ×4) block.

The combination of both the features MVE and SRF can be very crucial which can give a better saliency map. If both the features are having huge values then there is very high possibility of moving objects in that region as well as it can contain sudden objects. Therefore, our model consists of a combination of both MVE and SRN features. The saliency map estimation is represented in equation 7 below.

$$Smotion = \eta\,(mve + RF(Z) + mve \odot RF(Z)) \quad (7)$$

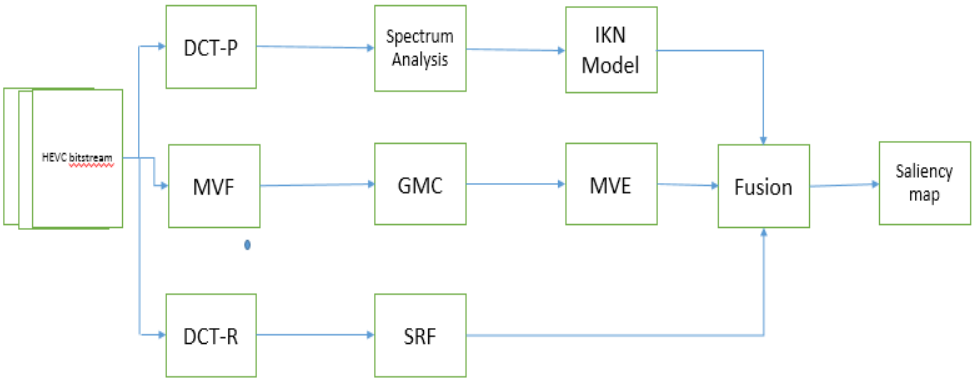


**Figure 2:** Compressed Domain Visual Saliency Model(MVF-motion vector feature, DCT-P-Discrete Cosine Transform pixel based, DCT-R-Discrete Cosine Transform Residual, GMC-Global Motion Cmpensation,SRF-Smooth Residual Feature,MVE-Motion Vector Entropy)

The Figure 2 shows Compressed Domain Visual Saliency Model.To get final spatial-temporal saliency map of $N$, we combine both spatial saliency obtained using IKN model in section2.1 and motion saliency of $N$ from MVE and SRF eliminating camera motion together with the help of coherent normalization technique based on fusion [25,26],

$$\text{Sspat\_temp(N)} = (1-\alpha)\text{Sspatial(N)} + \alpha\text{Smotion(N)} + \beta\text{Sspatial(N)Smotion(N)} \quad (8)$$

Where, positive constants α=0.9 and β=1 in our experiment. In equation (7), D(Z) represents MCGS (Motion Compensated Global Saliency) of Z. In equation (8), the first two terms show the spatial and temporal saliency respectively to compute an independent macroblock. In equation (8), the third term represents weights of temporal saliency in terms of spatial saliency and vice versa. Therefore, this can be defined as mutual reinforcement term which encourages spatially and temporally macro-blocks.

## EXPERIMENTAL RESULTS

In this paper, we have compared our experimental results with many state-of-the-art techniques such as $Itti$ [21], Surprise [30], $Judd$ [31], PQFT [32], $Rudoy$[33], Fang [34] and OBDL [35] existing techniques.

The proposed algorithm is evaluated on the SFU[26] and DIEM http://thediemproject.wordpress.com) datasets. The first viewing in the SFU dataset and right eye fixations in the DIEM[27] data set were used as the ground truth. FFM- PEG (www.ffmpeg.org) was used to encode with QP {20,28,36} and 1/4- pixel MV accuracy with no range restriction. Since the videos in the DIEM dataset and SFU dataset are at various resolutions, they were first resized to 288 pixels height, while preserving the original aspect ratio, resulting resolutions: 352X288, In this paper, all the raw videos sampled on YUV 4:2:0 sampling. All the videos are compressed to high quality (more than 30 $dB$).This videos includes contents such as sport events, video conferencing, surveillance, video games etc. All 12 raw videos in the are of $352 \times 288$ resolution with different frames.

To compare our algorithm in predict fixation we use two parameters Area Under receiver operating characteristic Curve (AUC) [28] and Normalized Scanpath Saliency (NSS)[29]. Fig. 3 shows the average AUC score (blue bar) as well as NSS score (red bar) of various algorithms across the test sequences. Not surprisingly, on average, pixel- domain methods perform better than compressed-domain ones. Here, we have shown average AUC and NSS comparison with other existing techniques for all 12 videos in **Table 4**.The average AUC and NSS for all 12 videos using our proposed method are 0.846 and 1.478 for SFU dataset respectively.

In **Table1** with Quantization parameter (QP=20) the average PSNR of encoded sequences was 45.2 dB. On the other hand, **Table 2** QP=28 represents with average PSNR of 36 dB. The average PSNR for Y, U and V channel using our proposed method is 30.739, 36.488 and 37.35 respectively presented in **Table 3** with QP=36. However, our proposed compressed-domain method tops all other methods on both metrics summarised in **Table 4** and **Figure3.**

**Table 1**:PSNR, CSNR and MSE RESULTS for Y, U, V CHANNEL using our proposed method using HM 20.0 considering both DIEM and SFU dataset

| VIDEO NAME | PSNR($dB$) | | |
|---|---|---|---|
| | Y | U | V |
| BEES(DIEM) | 45.208 | 47.512 | 48.132 |
| CLOSEUP(DIEM) | 45.966 | 49.092 | 48.696 |
| BUS(SFU) | 42.530 | 44.963 | 46.498 |
| CITY(SFU) | 42.063 | 45.765 | 47.024 |

**Table 2**: PSNR, CSNR and MSE RESULTS for Y, U, V CHANNEL using our proposed method using HM 28.0 DIEM and SFU dataset

| VIDEO NAME | PSNR($dB$) | | |
|---|---|---|---|
| | Y | U | V |
| BEES(DIEM) | 39.023 | 42.415 | 43.265 |
| CLOSEUP(DIEM) | 39.993 | 44.308 | 43.696 |
| BUS(SFU) | 29.768 | 37.793 | 39.048 |
| CITY(SFU) | 35.499 | 42.255 | 43.750 |

**Table 3**: PSNR, CSNR and MSE RESULTS for Y, U, V CHANNEL using our proposed methodusing HM 36.0 DIEM and SFU dataset

| VIDEO NAME | PSNR($dB$) | | |
|---|---|---|---|
| | Y | U | V |
| BEES(DIEM) | 32.807 | 37.947 | 38.652 |
| CLOSEUP(DIEM) | 34.170 | 39.688 | 38.726 |
| BUS(SFU) | 35.935 | 40.475 | 42.029 |
| CITY(SFU) | 29.798 | 39.820 | 41.118 |

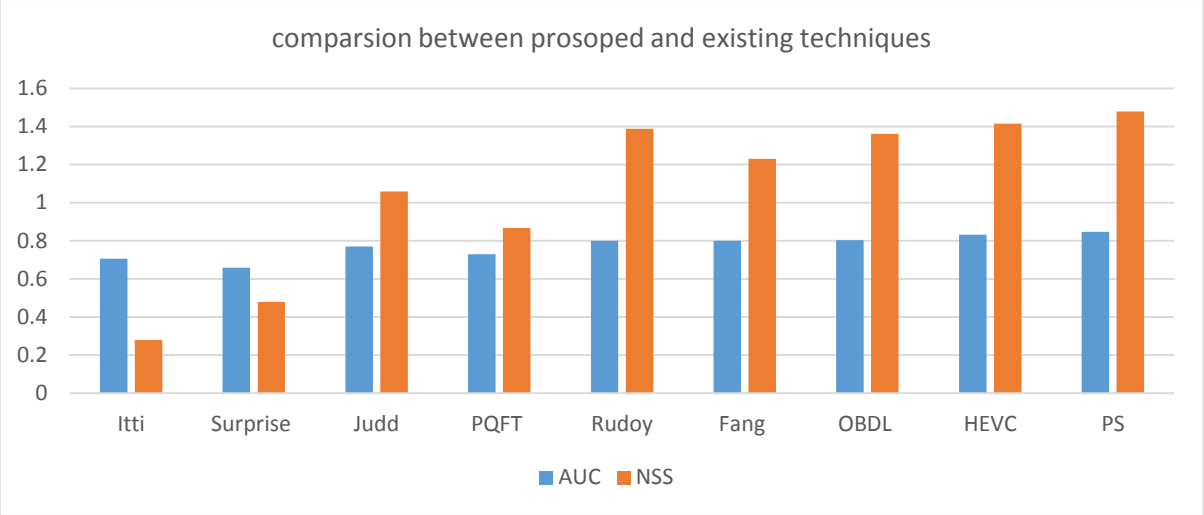


**Figure 3.** Comparison of our proposed model with existing state-of-the art-techniques

**Table 4:** Mean (standard deviation) values for saliency detection accuracy of our and other methods over SFU and DIEM databases

| | SFU | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | our | *Itti* [21] | Surprise [30] | JIDD[31] | PQFT[32] | RUDOY[33] | FANG[34] | OBDL[35] |
| AUC | 0.846(0.07) | 0.705(0.07) | 0.658(0.12) | 0.770(0.07) | 0.729(0.08) | 0.799(0.08) | 0.801(0.07) | 0.802(0.07) |
| NSS | 1.478(0.34) | 0.278(0.36) | 0.479(0.58) | 1.058(0.33) | 0.867(0.45) | 1.388(0.57) | 1.236(0.40) | 1.361(0.57) |
| | DIEM | | | | | | | |
| | our | *Itti* [21] | Surprise [30] | JIDD[31] | PQFT[32] | RUDOY[33] | FANG[34] | OBDL[35] |
| AUC | 0.879(0.07) | 0.775(0.07) | 0.754(0.12) | 0.751(0.09) | 0.795(0.08) | 0.804(0.11) | 0.808(0.09) | 0.790(0.12) |
| NSS | 1.889(0.64) | 0.545(0.67) | 0.935(0.91) | 0.990(0.40) | 1.282(0.75) | 1.488(0.91) | 1.232(0.57) | 1.651(1.01) |

## CONCLUSION

An effective saliency map is extracted from temporal component and spatial component by eliminating camera motion using GME. The proposed saliency model outperforms all the existing techniques. All the videos are compressed to high quality (more than 30 $dB$). The average PSNR for Y, U and V channel using our proposed method is 30.739, 36.488 and 37.35 respectively. Here, we have shown average AUC and NSS comparison with other existing techniques for all 12 videos in **Table 4.** The average AUC and NSS for all 12 videos using our proposed method are 0.846 and 1.478 for SFU dataset and 0.879 and 1.889 for DIEM respectively.

## REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, 1998,"A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach.Intell.*, Vol. 20, No. 11, pp. 1254–1259.

[2] L. Itti, N. Dhavale, and F. Pighin, 2004, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Opt. Sci.Technol.*, Vol. 64, pp. 64–78.

[3] J. Harel, C. Koch, and P. Perona,2006, "Graph-based visual saliency," in *Proc.NIPS*, pp. 545–552.

[4] R. J. Peters and L. Itti, 2007, "Beyond bottom-up: Incorporating taskdependent influences into a computational model of spatial attention,"in *Proc. CVPR*, pp. 1–8.

[5] L. Itti and P. Baldi, 2009,"Bayesian surprise attracts human attention," *Vis.Res.*, vol. 49, no. 10, pp. 1295–1306.

*[6]* L. Zhang, M. H. Tong, and G. W. Cottrell, 2009, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Proc. Annu. Cognit.*

[7] *Sci. Conf.*, pp. 2944–2949.

[8] C. Guo and L. Zhang,2010, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, Vol. 19, No. 1, pp. 185–198.

[9] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan, 2013, "Regularized feature reconstruction for spatio-temporal saliency detection," *IEEE Trans. Image Process.*, Vol. 22, No. 8, pp. 3120–3132.

[10] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, 2013,"A visual-attention model using earth mover's distance-based saliency measurement and nonlinear feature combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.35,No.2,pp.314–328.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, 2009, "Learning to predict where humans look," in *Proc. ICCV*, pp. 2106–2113.

[12] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, 2013, "Learning video saliency from human gaze using candidate selection," in *Proc. CVPR*, pp. 1147–1154

[13] J. Li, Y. Tian, T. Huang, and W. Gao, 2010,"Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, Vol. 90,No. 2, pp. 150–165.

[14] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, 2012, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 22, No. 12, pp. 1649–1668.

[15] Y. F. Ma and H. J. Zhang,2001, "A new perceived motion based shot content representation,".In Proc. IEEE ICIP'01, Vol. 3, pp. 426

[16] Y. F. Ma and H. J. Zhang,2002," A model of motion attention for video skimming," In Proc. IEEE ICIP'02, Vol 1, pp. 129

[17] H. Zen, T. Hasegawa, and S. Ozawa, 1999," Moving object detection from MPEG coded picture," In Proc. IEEE ICIP'99, Vol 4, pp. 25-37

[18] Z. Liu, H. Yan, L. Shen, Y.Wang, and Z. Zhang, 2009," A motion attention model based rate control algorithm for H. 264/AVC," In The 8th IEEE/ACIS International Conference on Computer and Information Science (ICIS'09), pap. 568-573.

[19] Y. Fang, W. Lin, Z. Chen, C. M. Tsai, and C. W. Lin,2014," A video saliency detection model in compressed domain," IEEE Trans. Circuits Syst. Video Technol., pp. 24-27.

[20] K. Muthuswamy and D. Rajan,2013, "Salient motion detection in compressed domain," IEEE Signal Process. Lett.

[21] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk,2009, "Frequency-tuned salient region detection," in Proc. IEEE Comput. Vision Pattern Recog., Miami Beach, FL.

[22] J.W.Woods, 2012, "Multidimensional signal, image, and video processing and coding ," 2nd ed. Academic Press/Elsevier.

[23] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic,2012, "Eye-tracking database for a set of standard video sequences," *IEEE Trans. Image Process.*, Vol. 21, No. 2, pp. 898–903.

[24] Y.-M. Chen and I. V. Baji´c,2010, "Motion vector outlier rejection cascade for global motion estimation," IEEE Signal Process. Lett.,Vol. 17, No. 2, pp. 197–200.

[25] Y. F. Ma and H. J. Zhang,2001" A new perceived motion based shot content representation,"In Proc. IEEE ICIP'01, Vol 3, pp. 426-429.

[26] C. Chamaret, J. C. Chevet, and O. Le Meur, 2010, "Spatio-temporal combination of saliencymaps and eye-tracking assessment of different strategies," IEEE International Conf.on Image Proc., pp. 1077–1080.

[27] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures.Chapman & Hall/CRC, 2007.

[28] H. Hadizadeh, M. J. Enriquez, and I. V. Baji_c,2012," Eye-tracking database for a set ofstandard video sequences ," IEEE Trans. Image Process., pp.898-903.

[29] The dynamic images and eye movements (DIEM) project. http://thediemproject.wordpress.com. [Online]. 5, 20, 70

[30] Borji A, Itti L ,2013," State-of-the-art in visual attention modelling," IEEE Trans Pattern Anal Mach Intell, pp.185–207

[31] Borji A, Sihite DN, Itti L,2013," Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study ," IEEE Trans Image Process 22(1), pp.55–69

[32] Itti L, Baldi P,2005," A principled approach to detecting surprising events in video ," In: IEEE Computer society conference on computer vision and pattern recognition (CVPR'05), Vol 1, pp. 631– 637

[33] L. Itti and P. Baldi,2009, "Bayesian surprise attracts human attention," *Vis. Res.*, Vol. 49, No. 10, pp. 1295–1306,.

[34] T. Judd, K. Ehinger, F. Durand, and A. Torralba, 2009,"Learning to predict where humans look," in *Proc. ICCV*, pp. 2106–2113.

[35] C. Guo and L. Zhang,2010, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, Vol. 19, No. 1, pp. 185–198, 2010.

[36] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor,2013, "Learning video saliency from human gaze using candidate selection," in *Proc.CVPR*, pp. 1147–1154.

[37] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin,2014, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst.Video Technol., Vol. 24, No. 1, pp. 27–38..*