3D MOTION IN VISUAL SALIENCY MODELING

Pengfei Wan [◦], *Yunlong Feng* [*], *Gene Cheung* [#], *Ivan V. Bajić* [$], *Oscar C. Au* [◦], *and Yusheng Ji* [#]

[◦] Hong Kong University of Science and Technology, [*] The Graduate University for Advance Studies,
[#] National Institute of Informatics, [$] Simon Fraser University

ABSTRACT

Visual saliency is a probabilistic estimate of how likely a given spatial area in an image or video is to attract human visual attention relative to other areas. Bottom-up saliency models aggregate low-level image features like luminance and color contrast, flicker, 2D motion, etc. to construct a plausible saliency map. In this paper, we introduce 3D motion (object movements towards or away from the observer) into bottom-up video saliency modeling. Given availability of per-pixel depth maps, we first propose a novel algorithm to estimate 3D motion vectors (3DMVs) for arbitrarily shaped sub-blocks in texture-plus-depth videos. We then derive two feature channels from 3DMVs to be incorporated into a widely accepted bottom-up saliency model. Experiments on subjective quality of Region-of-Interest (ROI) based video coding show that our enriched saliency model with 3DMV channels is more accurate in estimating human visual attention.

Index Terms— 3D motion estimation, visual saliency computation, ROI-based video coding

1. INTRODUCTION

Visual saliency estimates how likely a given local spatial area in an image or video frame is to attract human visual attention relative to other areas. Many models compute saliency maps in a bottom-up manner by aggregating low-level image features, such as luminance and color contrast, flicker, 2D motion, etc. Although the accuracy of different models varies, in general many models predict observers' gaze tendency reasonably well [1]. Accurate saliency maps can be used for Region-of-Interest (ROI) based image and video compression [2], subjective multimedia quality assessment [3], saliency-cognizant error concealment in loss-corrupted video [4], etc.

While 2D motion (object movements along x- or y-dimension) has been previously used as an input feature for saliency map computation (moving objects tend to attract human visual attention [5]), 3D motion—movement along the z-dimension towards or away from the observer—has never been considered in saliency computation. From a biological viewpoint, an object moving towards the observer presents a potential physical threat, and hence should typically trigger immediate attention due to innate survival instinct. One reason why 3D motion has not been considered in visual saliency computation is simply technological: it is difficult to estimate 3D motion in conventional 2D videos composed of texture frames only.

With the advent of depth-sensing cameras such as Microsoft Kinect®, depth video—per-pixel distance between captured objects in the 3D scene and the capturing camera—can now be readily acquired along with texture video (RGB or YUV images) from the same viewpoint. In this paper, we introduce 3D motion into bottom-up video saliency modeling for texture-plus-depth videos. We first estimate 3DMVs for arbitrarily shaped sub-blocks. Then we derive two feature channels from the computed 3DMVs, which are subsequently incorporated into a widely accepted bottom-up saliency model [6, 7]. Extensive experiments on subjective quality of Region-of-Interest (ROI) based video coding show that our enriched saliency model with added 3DMV channels is more accurate in estimating human visual attention.

The outline of the paper is as follows. We first discuss related work in Section 2. Then we present the proposed algorithm to estimate 3DMVs. We discuss how 3DMVs are used for saliency modeling in Section 4. Finally, experiments and conclusion are presented in Sections 5 and 6, respectively.

2. RELATED WORK

Although motion estimation has been extensively studied for texture videos, only a few works study motion estimation for texture-plus-depth videos [8, 9]. However, they are designed for video compression through 2D motion vector sharing [10]. In this paper we propose a joint 3D motion estimation method to recover the physical object motion in 3D space, which is especially useful for high-level video analysis tasks, such as saliency modeling, action recognition and scene understanding.

Generally speaking, there are two classes of saliency modeling approaches for images and videos: bottom-up and top-down. The bottom-up methods [6, 11] are stimuli-driven. They aggregate low-level visual stimuli into a plausible overall visual saliency map. The top-down approaches [12] are semantic-driven; e.g. humans naturally recognize and are attracted to human faces. A recent overview of saliency modeling can be found in [13]. While for simplicity we assume a baseline bottom-up saliency model when incorporating 3D motion in this work, a future extension can involve a hybrid model that combines bottom-up and top-down visual cues.

3. COMPUTING 3D MOTION VECTORS

In this section, we present a novel 3D motion estimation method to estimate the 3D motion of blocks with the help of depth information. While 2D motion estimation methods compute object motion along x- and y-dimension, we conjecture that z-motion deserves a much larger role in visual saliency computation. The reason is biological: an object coming towards a viewer represents a potential physical threat (e.g., a predator), and hence should draw immediate attention due to innate self-preservation instincts.

Our inputs are a texture video (8-bit RGB or YUV) and a depth video of the same resolution taken from the same viewpoint. A $N \times N$ block $\mathbf{B} = \{\mathbf{B^t}, \mathbf{B^d}\}$ refers to a texture block $\mathbf{B^t}$ as well as the corresponding depth block $\mathbf{B^d}$. As a convention, only luma component is used for texture block matching, and the previous frame is used as the reference frame.

3.1. Sub-block Partitioning

Assuming pixels of different objects usually have different motions, a block containing multiple objects may fail to find a good match in the reference frame. Thanks to the depth information, such blocks could be easily divided into two pixel groups, leading to two arbitrarily shaped sub-blocks (e.g., foreground and background sub-blocks). Each sub-block then gets assigned its own 3DMV. By sub-block partitioning, we can improve the accuracy of block matching especially near object boundaries.

To partition current block $\mathbf{B_c} = \{\mathbf{B_c^t}, \mathbf{B_c^d}\}$ into at most two sub-blocks masked by matrix $\mathbf{M} \in \{0,1\}^{N\times N}$, we use only the depth information $\mathbf{B_c^d}$. Specifically, if the standard deviation of values in $\mathbf{B_c^d}$ is smaller than a pre-defined threshold T_s, we return $\mathbf{M} = \mathbf{1}$, indicating that $\mathbf{B_c}$ should not be partitioned. Otherwise, pixels in $\mathbf{B_c^d}$ are divided into two groups (represented by 0s and 1s in $\mathbf{M}$) by the mean depth value in $\mathbf{B_c^d}$. In the end, a morphological closing operation (dilation followed by erosion) is performed on $\mathbf{M}$ so that pixels in one sub-block form a contiguous region. This process is both efficient and effective; see Fig. 1 for an illustration.

Since an unpartitioned block can be viewed as a special case with $\mathbf{M} = \mathbf{1}$, sub-block (with associated $\mathbf{M}$) is the basic processing unit of proposed 3D motion estimation method.

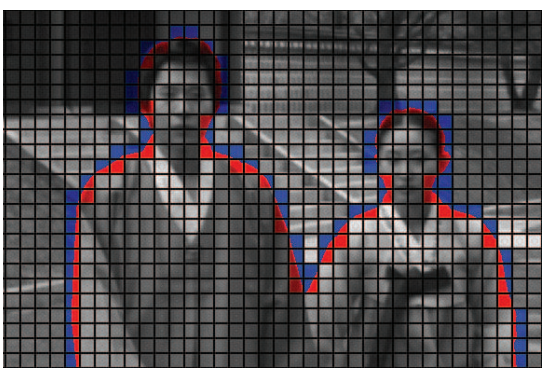


Fig. 1. Arbitrary-shaped sub-block partition with $N = 16$. Sub-blocks with mask $\mathbf{M} \neq \mathbf{1}$ are marked in red and blue.

3.2. 3D Motion Estimation

After partitioning, 3D motion estimation is performed on each sub-block to estimate its 3DMV. To contain complexity, we restrict the block search to a 3D window consisting of a conventional 2D spatial search window and a 1D depth search window. Different from existing methods [8, 9], our method is a joint 3D motion estimation, which uses variable block size adaptive to depth change.

3.2.1. 3DMVs in Physical Units

Our 3D motion estimation is aimed at recovering the true motion in 3D space over a constant frame interval Δt. So the 3DMV $\mathbf{mv} = (mv_x, mv_y, mv_z)$ should be measured in *physical distance*. Although the acquired mv_z is the difference in depth values (in meters) between the current and reference sub-blocks, the acquired x- and y-components mv_x^{pxl} and mv_y^{pxl} are typically in unit of pixels.

The inconsistency of units can be solved by well-accepted pinhole imaging model. In Fig. 2, assuming a 3D point (X_0, Y_0, D_c) moves to $(X_0, Y_0 + \Delta Y, D_c)$, we have the following relation based

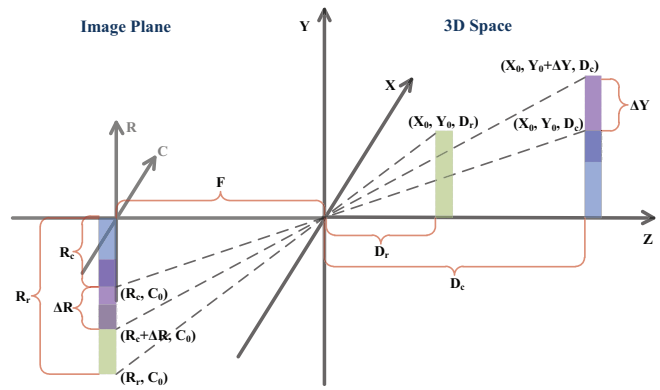


Fig. 2. Perspective pinhole imaging model. A 3D point (X, Y, Z) maps to the image plane indexed by row (axis R) and column (axis C). F is the focal length in unit of pixels, which is given in the camera intrinsic matrix.

on similar triangles:

$$\frac{mv_y}{mv_y^{\text{pxl}}} = \frac{\Delta Y}{\Delta R} = \frac{D_c}{F} \tag{1}$$

Therefore, motion in image plane mv_y^{pxl} can be converted to the physical motion mv_y by:

$$mv_y = \frac{D_c}{F} mv_y^{\text{pxl}} \tag{2}$$

where $D_c = depth(\mathbf{B_c^d})$ is the mean depth of the current sub-block. Similarly, we have $mv_x = \frac{D_c}{F} mv_x^{\text{pxl}}$.

3.2.2. Joint 3D Search

Unlike motion in the x- or y-dimension, z-motion means an object has moved closer to or further away from the camera, resulting in object resizing from reference frame to current frame. Hence, an accurate block matching in 3D motion estimation should have reference blocks with varying size L.

Fortunately, checking all block sizes in reference frame is not necessary. In Fig. 2, assuming a 3D point (X_0, Y_0, D_c) moves to new position (X_0, Y_0, D_r), according to similar triangles we have:

$$\begin{aligned} &\because Y_0 = \frac{R_c}{F} D_c = \frac{R_r}{F} D_r \\ &\therefore \frac{N}{L} = \frac{R_c}{R_r} = \frac{D_r}{D_c} \quad \therefore L = \frac{D_c}{D_r} N \end{aligned} \tag{3}$$

Assume the depth change within frame interval Δt is below a threshold D_0, i.e. the depth of reference block $D_r \in [D_c - D_0, D_c + D_0]$. By replacing D_r with $D_c \pm D_0$, we get the range of L, the size of reference block in integers, below:

$$L \in [N_{\min}, N_{\max}] = \left[round(\frac{N}{1 + \frac{D_0}{D_c}}), round(\frac{N}{1 - \frac{D_0}{D_c}}) \right] \tag{4}$$

Alg. 1 summarizes our joint 3D search for candidate sub-blocks given the current block $\mathbf{B_c} = \{\mathbf{B_c^t}, \mathbf{B_c^d}\}$ located at (r, c). To find the best match, we check reference blocks within a 2D search window in reference depth frame $\mathbf{F_r^d}$ and texture frame $\mathbf{F_r^t}$. Different from conventional methods, the size of reference blocks L varies from

Algorithm 1 Joint Search with 3D Window

Input: $\mathbf{B_c} = \{\mathbf{B_c^t}, \mathbf{B_c^d}\}$ at (r, c), $\mathbf{F_r^d}, \mathbf{F_r^t}$
Output: $\mathcal{S}$

$\mathcal{S} \leftarrow \emptyset$
$D_c \leftarrow depth(\mathbf{B_c^d})$
for $L = N_{\min}$ to $N_{\max}$ **do** ▷ refer to (4)
  for all (i, j) in 2D search window **do**
    $\tilde{\mathbf{B_r^d}} \leftarrow L \times L$ block of $\mathbf{F_r^d}$ at (i, j)
    $\mathbf{B_r^d} \leftarrow rescale(\tilde{\mathbf{B_r^d}})$
    $D_r \leftarrow depth(\mathbf{B_r^d})$
    if $|\frac{N}{L} D_c - D_r| \leq T_d$ **then** ▷ refer to (3)
      $\tilde{\mathbf{B_r^t}} \leftarrow L \times L$ block of $\mathbf{F_r^t}$ at (i, j)
      $\mathbf{B_r^t} \leftarrow rescale(\tilde{\mathbf{B_r^t}})$
      $mv_x \leftarrow \frac{D_c}{F}(c - j)$ ▷ refer to (2)
      $mv_y \leftarrow \frac{D_c}{F}(r - i)$
      $mv_z \leftarrow D_c - D_r$
      $\mathbf{mv} \leftarrow (mv_x, mv_y, mv_z)$
      put $\{\mathbf{B_r^t}, \mathbf{mv}\}$ into $\mathcal{S}$
    end if
  end for
end for
return $\mathcal{S}$

$N_{\min}$ to $N_{\max}$. For sub-block masking, we rescale the reference block from $L \times L$ to $N \times N$. From (3), we know that $\frac{N}{L} D_c$ is the expected depth value for a reference block with size L. So the 1D depth search window rejects any reference blocks with $|\frac{N}{L} D_c - D_r| > T_d$. The survivors are candidate sub-blocks, whose texture information $\mathbf{B_r^t}$ and corresponding 3DMV $\mathbf{mv}$ are stored in set $\mathcal{S}$.

In the special case when the video has been captured at a high frame rate relative to the speed of 3D motion in the scene, we only need to check $L = N$ since $\frac{D_0}{D_c} \to 0$.

3.2.3. *Matching Criterion*

Given the set of candidate sub-blocks $\mathcal{S}$, the 3DMV $\mathbf{mv}^*$ for current sub-block $\mathbf{B_c}$ is the one with the smallest matching error:

$$\mathbf{mv}^*(\mathbf{B_c}) = \underset{\{\mathbf{B_r^t}, \mathbf{mv}\} \in \mathcal{S}}{\operatorname{argmin}} \; err(\mathbf{B_r^t}, \mathbf{mv}) \tag{5}$$

where

$$err(\mathbf{B_r^t}, \mathbf{mv}) = \frac{1}{card(\mathbf{M})} ||\mathbf{B_c^t} - \mathbf{B_r^t}||_1 + \lambda ||\mathbf{mv} - \mathbf{mv_p}||_2 \tag{6}$$

There are two terms in the error function balanced by λ. The first term is the Mean-Absolute-Difference (MAD) for texture blocks, where $card(\mathbf{M})$ is the number of 1's in $\mathbf{B_c}$'s sub-block mask $\mathbf{M}$. The second term $||\mathbf{mv} - \mathbf{mv_p}||_2$ is a regularization term to enforce a piece-wise smooth motion field, where $\mathbf{mv_p}$ is the 3DMV predictor. The predictor is the Laplacian-weighted average of 3DMVs of causal neighboring sub-blocks. In the example in Fig. 3:

$$\mathbf{mv_p} = \frac{\sum_{i=1}^{8} w_i \cdot \mathbf{mv}^*(\mathbf{B_i})}{\sum_{i=1}^{8} w_i} \tag{7}$$
$$w_i = \exp(-\tau |depth(\mathbf{B_c^d}) - depth(\mathbf{B_i^d})|)$$

Our assumption is that 3DMVs of sub-blocks from the same object (neighboring sub-blocks with similar depth) should have strong correlation. The regularization term effectively removes candidate blocks with irregular 3DMVs. The output 3DMV $\mathbf{mv}^*(\mathbf{B_c})$ is proportional to absolute physical velocity in the 3D space.

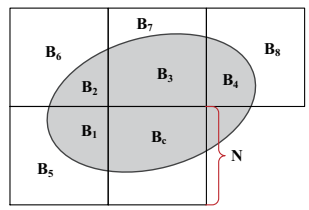


Fig. 3. 3DMV prediction using causal sub-blocks ($\mathbf{B_1}$ to $\mathbf{B_8}$).

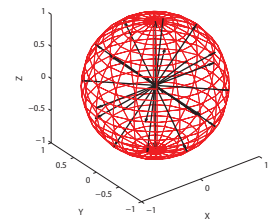


Fig. 4. Quantified 3D directions (black arrows).

4. 3D MOTION IN SALIENCY MODELING

In this section we derive two feature channels from 3DMVs. They are combined with conventional low-level feature channels into an enriched saliency map.

4.1. 3D Motion Magnitude (3DMM)

It is commonly accepted that objects with larger motion draw more attention from observers. Thus, the most straightforward way to compute conspicuity for a given 3DMV (mv_x, mv_y, mv_z) is to compute its Euclidean norm: $\sqrt{mv_x^2 + mv_y^2 + mv_z^2}$.

Euclidean norm treats the vector components equally. However, as discussed in the Introduction, an object moving towards the observer ($mv_z < 0$) should be weighted more heavily. We thus modify it into the following:

$$\text{3DMM} = \sqrt{mv_x^2 + mv_y^2 + \alpha \cdot mv_z^2}$$
$$\alpha = \begin{cases} 1 & \text{if } mv_z \geq 0 \\ 9 & \text{if } mv_z < 0 \end{cases} \tag{8}$$

By calculating 3DMM for each sub-block, we can get a conspicuity map $\mathbf{CM}_{\text{3dmm}}$ for each frame.

4.2. 3D Direction Self-information (3DDS)

Like previous work [14, 15] that assumes "surprise" elements draw more attention, here we assume a 3DMV with an unusual motion direction should be more salient. First, we uniformly divide the 3D motion field into 27 directions (see Fig. 4). For each frame we classify 3DMV directions of all sub-blocks into the 27 bins, and use the normalized histogram as an approximation of probability mass function $\mathbf{Pr}(\cdot)$ of motion directions. By calculating the self-information, we can give higher conspicuity to uncommon directions.

$$\text{3DDS} = -\log(\mathbf{Pr}(\text{bin}(\mathbf{mv}))) \tag{9}$$

Similarly, a conspicuity map $\mathbf{CM}_{\text{3dds}}$ can be constructed for each frame using 3DDS.

4.3. Feature Integration

Itti's model [6, 7] is a well-known framework for bottom-up saliency modeling, where a conspicuity map is calculated for each feature channel. There are several channels in Itti's model for video saliency detection [7]: intensity ($\mathcal{I}$), color ($\mathcal{C}$), orientation ($\mathcal{O}$), 2D motion ($\mathcal{M}$) and temporal flicker ($\mathcal{F}$). The proposed 3DMM and 3DDS serve as additional channels to the existing framework.

In Itti's model, $maxnorm\ \mathcal{N}(\cdot)$ is the most popular approach to fuse conspicuity maps:

$$\mathcal{N}(\mathbf{CM}) = (1 - \overline{m})^2 \mathbf{CM} \tag{10}$$

where $\overline{m}$ is the normalized mean value of local maxima within $\mathbf{CM}$. Like [6], the final saliency map $\mathbf{SM}$ is obtained by combining conspicuity maps of all channels:

$$\begin{aligned} \mathbf{SM}_{\mathrm{Itti}} &= \sum_{i=\{\mathcal{I},\mathcal{C},\mathcal{O},\mathcal{M},\mathcal{F}\}} \mathcal{N}(\mathbf{CM_i}) \\ \mathbf{SM}_{\mathrm{3dmv}} &= \mathcal{N}(\mathbf{SM}_{\mathrm{Itti}}) + \mathcal{N}(\mathbf{CM}_{\mathrm{3dmm}}) + \mathcal{N}(\mathbf{CM}_{\mathrm{3dds}}) \end{aligned} \tag{11}$$

where $\mathbf{SM}_{\mathrm{3dmv}}$ and $\mathbf{SM}_{\mathrm{Itti}}$ are respectively the saliency map with and without proposed 3DMV channels. They are further normalized to $[0, 1]$ for experiments. In this paper, we make use of the implementation of Itti's model for video from [16].

5. EXPERIMENTATION

We verify the effectiveness of our proposed saliency model via extensive subjective experiments. The results show that our model is statistically more accurate in estimating human visual attention.

5.1. ROI-based Video Coding

To test the accuracy of saliency maps, we encoded the texture video with adaptive quality based on saliency value. All frames are intra-coded using H.264/AVC reference software JM version 18.4 [17]. The quantization parameter (QP) of each macro-block is inverse-proportionally determined by its mean saliency value, i.e. small QP (high quality) for high saliency regions.

Four sequences: *lovebird*, *toy_forward*, *toy_f_side*, *toy_f_back* are used for our experiments (see Table. 1), where $lovebird$ is the standard test sequence [18] and the rest are captured using a combination of a RGB camera and a PMD Time-of-Flight (ToF) depth camera [19] with proper view mapping. All sequences have the frame rate 30 fps. Two ROI-encoded videos with the same bit-rate (4 $mbps$) are produced (by $\mathbf{SM}_{\mathrm{Itti}}$ and $\mathbf{SM}_{\mathrm{3dmv}}$ respectively) for each sequence, resulting in $4 \times 2 = 8$ ROI-encoded videos for subjective experiments.

5.2. Subjective Experiments

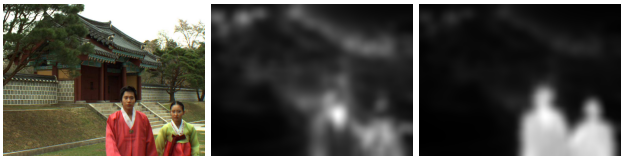
As recommended by ITU-R BT.500 [20], 20 participants (15 male and 5 female, of age 22-34) took part in the experiments. All participants had normal or corrected to normal sight, and were naive about the task of the experiment. A 23-inch LG monitor with resolution 1920×1080 and brightness 250 cd/m^2 was used for display. The ambient light in the room was 250-300 lux. The distance between the participant and monitor was approximately 40 cm.

For each sequence, participants were asked to watch the original sequence first. After that, two ROI-encoded videos (one based on $\mathbf{SM}_{\mathrm{3dmv}}$, the other on $\mathbf{SM}_{\mathrm{Itti}}$) were displayed twice, side by side on the screen in random order. An answer sheet was given to each participant to record the vote on which one is visually closer to the original sequence.

Participants' votes are shown in Table. 1, along with the p-values of the two-sided χ^2-test [21]. The null hypothesis is that votes for $\mathbf{SM}_{\mathrm{3dmv}}$ and $\mathbf{SM}_{\mathrm{Itti}}$ come from distributions with the same mean. Under this hypothesis, the expected number of votes for each case is 10. In our experiments, the extremely small p-values reject the null hypothesis and suggest that ROI-encoded videos based on $\mathbf{SM}_{\mathrm{3dmv}}$ are statistically preferred. The strong preference indicates a higher probability that human gaze falls into high quality (small QP, high saliency) regions in ROI-encoded videos based on $\mathbf{SM}_{\mathrm{3dmv}}$ than those based on $\mathbf{SM}_{\mathrm{Itti}}$. Therefore enriched saliency map $\mathbf{SM}_{\mathrm{3dmv}}$ is more accurate in terms of predicting human gaze, i.e. estimating human visual attention.

Table 1. Subjective Experiment Results

Sequence	*lovebird*	*toy_forward*	*toy_f_side*	*toy_f_back*
Resolution	1024×768	640×480	640×480	640×480
Duration	$2 \times 5s$	$2 \times 4s$	$2 \times 4s$	$2 \times 4s$
$\mathbf{SM}_{\mathrm{Itti}}$	0	1	1	0
$\mathbf{SM}_{\mathrm{3dmv}}$	20	19	19	20
p-value	8×10^{-6}	6×10^{-5}	6×10^{-5}	8×10^{-6}



(a) lovebird

(b) toy_forward

(c) toy_f_side

(d) toy_f_back

Fig. 5. Sample saliency maps of 4 sequences. Columns from left to right: RGB frame, $\mathbf{SM}_{\mathrm{Itti}}$, $\mathbf{SM}_{\mathrm{3dmv}}$. For gray-scale saliency maps, higher intensity means higher saliency value

Sample saliency maps are also shown in Fig. 5. Comparing with $\mathbf{SM}_{\mathrm{Itti}}$, blocks with z-motion towards the camera are effectively detected in $\mathbf{SM}_{\mathrm{3dmv}}$ with proposed 3DMV channels.

6. CONCLUSION

In this paper, we propose a novel method to estimate 3DMVs for texture-plus-depth videos, from which we further derive two feature channels for bottom-up video saliency modeling. Subjective experiments involving ROI-based video coding show that proposed 3DMV significantly improves the accuracy of human attention estimation.

7. REFERENCES

[1] V. A. Mateescu, H. Hadizadeh, and I. V. Bajić, "Evaluation of several visual saliency models in terms of gaze prediction accuracy on video," in *Proc. IEEE QoEMC'12*, 2012.

[2] Chenlei Guo and Liming Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Processing*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[3] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50 –59, Nov. 2011.

[4] H. Hadizadeh, I. V. Bajić, and G. Cheung, "Saliency-cognizant error concealment in loss-corrupted streaming video," in *Proc. IEEE ICME'12*, Melbourne, Australia, Jul. 2012, pp. 73 – 78.

[5] J. Tsotsos, M. Pomplun, Y. Liu, J. Martinez-Trujillo, and Simine E., "Attending to motion: Localizing and classifying motion patterns in image sequences," in *Proc. Intl. Workshop on Biologically Motivated Computer Vision*, 2002.

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

[7] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1304 –1318, Oct. 2004.

[8] B. Kamolrat, W.A.C. Fernando, M. Mrak, and A. Kondoz, "3D motion estimation for depth image coding in 3D video coding," *IEEE Trans. Consumer Electronics*, vol. 55, no. 2, pp. 824 – 830, May. 2009.

[9] Y.-C. Fan, S.-F. Wu, and B.-L. Lin, "Three-dimensional depth map motion estimation and compensation for 3d video compression," *IEEE Trans. Magnetics*, vol. 47, no. 3, pp. 691 –695, Mar. 2011.

[10] S. Grewatsch and E. Miller, "Sharing of motion vectors in 3D video coding," in *Proc. IEEE ICIP'04*, Oct. 2004, vol. 5, pp. 3271 – 3274.

[11] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Processing*, vol. 19, no. 1, pp. 185 –198, Jan. 2010.

[12] A. L. Yarbus, *Eye-Movements and Vision*, Plenum Press, New York, 1967.

[13] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[14] N. D. B Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Neural Information Processing Systems*, 2005.

[15] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Neural Information Processing Systems*, 2005.

[16] J. Harel, "A saliency implementation in MATLAB," *http://www.klab.caltech.edu/ harel/share/gbvs.php*.

[17] JM version 18.4, ," *http://iphome.hhi.de/suehring/tml/*.

[18] ISO/IEC JTC1/SC29/WG11, "Call for proposals on 3D video coding technology, n12036," Mar. 2011.

[19] ," *http://www.pmdtec.com/*.

[20] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures, international telecommunication union," 2002.

[21] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 4 edition, 2007.