3D Motion Estimation for Visual Saliency Modeling

Pengfei Wan *Student Member, IEEE*, Yunlong Feng *Student Member, IEEE*, Gene Cheung *Senior Member, IEEE*, Ivan V. Bajić *Senior Member, IEEE*, Oscar C. Au *Fellow, IEEE*

***Abstract*—Visual saliency is a probabilistic estimate of how likely a spatial area in an image or video frame is to attract human visual attention relative to other areas. When existing bottom-up saliency models aggregate low-level features to construct a plausible saliency map, only 2D motion cues are used as motion features, even though videos typically capture dynamic 3D scenes. In this paper, we introduce 3D motion into bottom-up saliency modeling for texture-plus-depth videos. We first propose an efficient 3D motion estimation algorithm, which computes a 3D motion vector (3DMV) for each sub-block in the frame. Using the computed 3DMVs, we then derive several saliency channels (called 3DMV channels), which are incorporated into a bottom-up saliency model to obtain enhanced saliency maps. Experiments tracking human gaze show that incorporating our 3DMV channels into bottom-up saliency model significantly improves the accuracy of derived saliency maps.**

***Index Terms*—3D motion estimation, visual saliency modeling.**

I. Introduction

Visual saliency estimates how likely a given local spatial area in an image or video frame is to attract human's attention relative to other areas. In the literature, many works [1–4] compute saliency maps in a bottom-up manner by aggregating low-level image (or video) features, such as luminance and color contrast, flicker, 2D motion, etc. Practical applications of derived saliency maps include Region-of-Interest (ROI) based image and video compression [5], subjective multimedia quality assessment [6], saliency-cognizant error concealment in loss-corrupted videos [7], etc.

While 2D motion—planar object movements along horizontal and vertical (x- and y-) dimensions—has been used as a feature for saliency map computation (moving objects tend to attract human's attention [8]), 3D motion—including 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 trigger immediate attention due to innate animal survival instinct. One reason why 3D motion has not been considered is simply technological: it is difficult to estimate 3D motion information in conventional 2D texture videos composed of color frames only.

With the advent of depth-sensing cameras such as Microsoft Kinect®, depth video—per-pixel distance between captured 3D scene and the capturing camera—can now be readily acquired along with texture video from the same viewpoint. Thanks to the availability of depth frames, estimation of 3D motion vectors (3DMVs) becomes possible. In our previous work [9], 3D motion is introduced into saliency modeling for texture-plus-depth videos. We first computed 3DMV for each sub-block composed of neighboring pixels of similar depth values. Using the computed 3DMVs, we next derived two saliency channels—*3D motion magnitude* (3DMM) and *3D direction self-information* (3DDS)—which were then incorporated into a widely-accepted bottom-up saliency model [1, 3]. Although results were encouraging, derivations of 3DMM and 3DDS were ad-hoc with few theoretical justifications.

In this paper, we pursue a more rigorous study and introduce three alternative 3DMV channels—*semi-spherical coordinate motion magnitude* (SCMM), *negative log-likelihood of direction mean* (NLDM) and *3D motion acceleration* (3DMA)—as better saliency indicators based on psychological and statistical theories. Specifically, this work contains three major improvements over [9]: 1) unlike 3DMM which measures motion magnitude in conventional Cartesian coordinate, SCMM measures the 3D motion magnitude in semi-spherical coordinate as suggested by psychological studies [10]; 2) unlike 3DDS with heuristic quantization of motion directions, NLDM statistically models the element of "surprise" for each detected 3D motion direction; 3) besides motion magnitude and direction, we introduce a new channel 3DMA that models *motion acceleration*, which has been previously shown to be a strong feature drawing human's attention [11]. In our experiments, we show that enhanced saliency maps using our new 3DMV channels correspond more closely to our collected human gaze data than conventional saliency maps [3] and those proposed in [9].

The outline of the paper is as follows. We first discuss related work in Section II. Then we introduce our 3D motion estimation algorithm and how 3D motion is used in saliency modeling in Section III and IV. Finally, experiments and conclusion are presented in Sections V and VI, respectively.

II. Related Work

In the literature, only a few works studied motion estimation for texture-plus-depth videos [12, 13]. However, they are straight-forward extensions of 2D motion estimation for video compression [14] and are incapable to recover true 3D motion information. Please see Section III for detailed comparisons.

Generally speaking, there are two classes of saliency modeling approaches for images and videos: bottom-up and top-down. Bottom-up methods[1] [1–4, 15] are stimuli-driven. They aggregate low-level features into a plausible overall visual

Pengfei Wan, Oscar C. Au are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology. Email Address: {leoman, eeau}@ust.hk

Yunlong Feng, Gene Cheung are with the National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan 101–8430. Email Address: {fengyl, cheung}@nii.ac.jp

Ivan V. Bajić is with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A 1A6. Email Address: ibajic@sfu.ca

[1] [15] studied saliency detection using disparity, but for static stereo images.

saliency map. Top-down approaches [16, 17] are semantic-driven; e.g., humans naturally recognize and are attracted to faces. A recent overview of saliency models can be found in [18]. While it is still an open debate which one is more accurate in predicting human's attention, for the sake of low complexity and generality, in this paper we choose the widely-accepted bottom-up model [1, 3] as the baseline model to which 3D motion is incorporated as an additional feature.

III. 3D Motion Estimation

In this section, we present our 3D motion estimation algorithm (first introduced in [9]), where block matching is restricted to a 3D window consisting of a conventional 2D spatial search window and a 1D depth search window.

In texture-plus-depth videos, every texture block $\mathbf{B}^t$ has a corresponding depth block $\mathbf{B}^d$. Because different physical objects in a 3D scene typically have different motion, we first partition any block covering two objects (e.g. foreground and background) into two arbitrarily shaped sub-blocks indicated by a binary *sub-block mask* $\mathbf{M}$. Calculation of $\mathbf{M}(\mathbf{B}^d)$ is efficient: if depth block $\mathbf{B}^d$ is smooth (small variance), we assume all pixels in the block belong to the same object and set $\mathbf{M} = \mathbf{1}$ (i.e. one single sub-block); otherwise, pixels in $\mathbf{B}^d$ are divided into two groups (0s and 1s in $\mathbf{M}$) by the mean value. Finally, morphological closing is performed on $\mathbf{M}$, so that pixels in each group (sub-block) form a contiguous region.

After partitioning, we perform 3D motion estimation on each sub-block to find its corresponding sub-block in reference frame. Let current sub-block be a partition (e.g. 1s in $\mathbf{M}(\mathbf{B}_c^d)$) of $N \times N$ block $\{\mathbf{B}_c^t, \mathbf{B}_c^d\}$ with average depth D_c. According to the pinhole camera model, we know a block of dimension $L \times L$ and of average depth D_r scales to dimension $N \times N$ of average depth D_c, where $L\,D_r = N\,D_c$. By defining δ as the maximum depth change from reference to current frame, reference blocks are restricted to $L \times L$ blocks satisfying: 1) within 2D spatial search window, 2) block size L is integer in $\left[\frac{N D_c}{D_c+\delta}, \frac{N D_c}{D_c-\delta}\right]$, and 3) average depth D_r is close to $N\,D_c/L$.

Each reference block offers a candidate 3DMV for current sub-block, which is calculated by:

$$\mathbf{v} = (\frac{D_c}{F}\Delta x, \frac{D_c}{F}\Delta y, D_c - D_r) \tag{1}$$

where F is the camera focal length; Δ=(Δx,Δy) is 2D spatial offsets. Since F and Δ are both in number of pixels, $\mathbf{v}$ has the same unit as D_c (in physical distance).

Denoting $\mathbb{S}$ as the set of all candidate 3DMVs, the output 3DMV $\mathbf{mv}$ for current sub-block is the one with the smallest matching error:

$$\mathbf{mv} = \arg\min_{\mathbf{v}\in\mathbb{S}} f(\mathbf{v}) \tag{2}$$

where

$$f(\mathbf{v}) = MAD(\mathbf{B}_c^t, \mathbf{B}_r^t, \mathbf{M}) + \lambda||\mathbf{v} - \mathbf{v}_p||_2 \tag{3}$$

There are two terms (balanced by λ) in (3). The first term is the mean absolute difference between $\mathbf{B}_c^t$ and $\mathbf{B}_r^t$ at positions indicated by sub-block mask $\mathbf{M} = \mathbf{M}(\mathbf{B}_c^d)$, where $\mathbf{B}_r^t = \mathbf{B}_r^t(\mathbf{v})$ is the texture reference block (resized from $L\times L$ to $N \times N$) associated with $\mathbf{v}$. The second term $||\mathbf{v} - \mathbf{v}_p||_2$ is a regularization term to enforce a piecewise smooth motion field, where predictor $\mathbf{v}_p$ is calculated from the 3DMVs of causal neighboring sub-blocks.

Our algorithm is different from [12, 13] in three respects. First, instead of square blocks, the basic process unit of our algorithm is arbitrarily shaped sub-blocks (corresponding to objects with different motion). Second, z-motion means object resizing from reference to current frame, so we check reference blocks whose size L spans several values around N. Finally, unlike traditional motion vectors measured in number of pixels, our 3DMV $\mathbf{mv} = (mv_x, mv_y, mv_z)$ is the true object motion measured in *physical distance*.

We note that our 3D motion estimation algorithm is generic: the obtained 3DMVs can be used in other applications other than visual saliency modeling (to be discussed).

IV. Saliency from 3D Motion

We now discuss different saliency channels derived from 3DMVs, including two old channels (3DMM and 3DDS) proposed in [9] and three new channels (SCMM, NLDM and 3DMA). They serve as additional channels to be combined with conventional channels to obtain enhanced saliency maps.

A. 3DMV Channels

1) 3D Motion Magnitude (3DMM): It is commonly accepted that objects with larger velocity draw more attention, so 3DMM is defined as the (weighted) Euclidean norm of $\mathbf{mv}$:

$$\text{3DMM} = \sqrt{mv_x^2 + mv_y^2 + \alpha(mv_z)^2}, \quad \alpha(x) = \begin{cases} x & \text{if } x \geq 0 \\ 3x & \text{if } x < 0 \end{cases} \tag{4}$$

where function α assigns higher weight for sub-blocks moving towards the observer ($mv_z < 0$).

2) 3D Direction Self-information (3DDS): Assuming unusual motion directions are more salient, 3DDS is defined as the self-information of quantized 3D motion direction:

$$\text{3DDS} = -\log(\Pr(\mathcal{Q}(\overline{\mathbf{mv}}))) \tag{5}$$

where unit 3DMV $\overline{\mathbf{mv}} = \mathbf{mv}/||\mathbf{mv}||_2$ represents the 3D motion direction, Pr is the normalized histogram of uniformly-quantized 3D motion directions $\mathcal{Q}(\overline{\mathbf{mv}})$ in current frame.

3) Semi-spherical Coordinate Motion Magnitude (SCMM): SCMM is an improved version of 3DMM. Rooted in animal survival instinct, it is shown in [10] that objects on a "collision path" with the observer are more attention-drawing. Thus we decompose the 3DMV using semi-spherical coordinate centering at camera location $\mathbf{c}$ (instead of Cartesian coordinate in 3DMM). In particular, we first back-project the central pixel of each sub-block into a 3D voxel locating at $\mathbf{n}$. Denoting $\bar{\mathbf{d}} = (\mathbf{c} - \mathbf{n})/||\mathbf{c} - \mathbf{n}||_2$ as the unit projection direction, the motion magnitude along $\bar{\mathbf{d}}$ is simply a dot-product below:

$$mv_d = \mathbf{mv} \cdot \bar{\mathbf{d}} \tag{6}$$

And the motion magnitude orthogonal to $\bar{\mathbf{d}}$ is $mv_o = \sqrt{mv_x^2 + mv_y^2 + mv_z^2 - mv_d^2}$. Since positive mv_d is the motion on "collision path", SCMM is defined as:

$$\text{SCMM} = \sqrt{mv_o^2 + \beta(mv_d)^2}, \quad \beta(x) = \begin{cases} 3x & \text{if } x \geq 0 \\ x & \text{if } x < 0 \end{cases} \tag{7}$$

4) Negative Log-likelihood of Direction Mean (NLDM): NLDM is an improved version of 3DDS. Unlike 3DDS, it is a statistical saliency channel where no quantization of motion directions is involved. In particular, we denote $\mathbf{x}$ as current 3D motion direction and $\mathbf{z_i}$ as other unit 3DMVs in the frame. Assuming 3D motion direction is i.i.d. that follows Gaussian distribution $\mathcal{G}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, the likelihood of $\boldsymbol{\mu} = \mathbf{x}$ is:

$$\prod_i f(\mathbf{z_i}|\boldsymbol{\mu} = \mathbf{x}) = \frac{1}{C}\exp(-\frac{\sum_i(\mathbf{z_i}-\mathbf{x})^\intercal(\mathbf{z_i}-\mathbf{x})}{2\sigma^2}) \quad (8)$$

Because smaller likelihood means higher irregularity of current direction $\mathbf{x}$, NLDM is defined as the negative log-likelihood:

$$\text{NLDM} = \frac{1}{2\sigma^2}\sum_i(\mathbf{z_i}-\mathbf{x})^\intercal(\mathbf{z_i}-\mathbf{x}) + \log(C) \quad (9)$$

which is the sum of squared differences between unit 3DMVs (C and σ are constant). Because the majority of sub-blocks have zero motion, calculation is restricted to non-zero 3DMVs.

5) 3D Motion Acceleration (3DMA): In addition to motion magnitude (included in 3DMM and SCMM) and motion direction (included in 3DDS and NLDM), the change in motion—acceleration—is also a strong attention-drawing feature [11]. Thanks to the per-sub-block 3DMVs obtained in Section III, calculation of acceleration is simply extracting the 3DMV of best-matched reference sub-block $\mathbf{mv}^{\text{ref}}$ from the current one:

$$\mathbf{mv}' = \text{sign}(\mathbf{mv}) \circ (\mathbf{mv} - \mathbf{mv}^{\text{ref}}) \quad (10)$$

where symbol $\circ$ stands for Hadamard product. By emphasizing positive acceleration, we define:

$$\text{3DMA} = ||\beta(\mathbf{mv}')||_2 \quad (11)$$

B. Feature Integration

Itti's model [1, 3] is a well-known framework for bottom-up saliency modeling, where a conspicuity map $\mathbf{CM}$ is calculated for each of the following channels: intensity ($\mathcal{I}$), color ($\mathcal{C}$), orientation ($\mathcal{O}$), 2D motion ($\mathcal{M}$) and temporal flicker ($\mathcal{F}$).

Our 3DMV channels serve as additional channels to Itti's model. For simplicity we denote $\mathcal{A}, \mathcal{B}$ for 3DMM, 3DDS [9]; and $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ for SCMM, NLDM, 3DMA respectively. Computing $\mathbf{CM}$ for a 3DMV channel is easy: for each sub-block in the frame, assign its conspicuity value (calculated by either (4)(5)(7)(9)(11)) to all pixels within it.

In Itti's model, conspicuity maps of different channels are first normalized by $maxnorm$:

$$\mathcal{N}(\mathbf{CM}) = (1-m)^2\overline{\mathbf{CM}} \quad (12)$$

where m is the mean value of local maxima within normalized conspicuity map $\overline{\mathbf{CM}}$. Itti's $maxnorm$ weighs different channels in a content-adaptive manner without any prior of relative importance. Like [1], the final saliency map $\mathbf{SM}$ is obtained by combining conspicuity maps of several channels:

$$\mathbf{SM} = \sum_{i=\mathbb{C}} w_i \cdot \mathcal{N}(\mathbf{CM_i}) \quad (13)$$

where $\mathbb{C}$ is the set of channels used in the saliency map, please refer to Table. I for details. Weight w_i enables us to tune the relative importance of difference channels, whose value is 1 for conventional channels and κ for 3DMV channels. The value of κ depends on how much z-motion there is in the video. As in [9], we fix $\kappa = 2$ in our experiments.

TABLE I
CHANNELS USED IN SALIENCY MAP

Saliency Map		$\mathbb{C}$
$\mathbf{SM}_{\text{Itti}}$	[3]	$\mathcal{I}, \mathcal{C}, \mathcal{O}, \mathcal{M}, \mathcal{F}$
$\mathbf{SM}_{\text{old}}$	[9]	$\mathcal{I}, \mathcal{C}, \mathcal{O}, \mathcal{M}, \mathcal{F}, \mathcal{A}, \mathcal{B}$
$\mathbf{SM}_{\text{new}}$		$\mathcal{I}, \mathcal{C}, \mathcal{O}, \mathcal{M}, \mathcal{F}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$

V. EXPERIMENTATION

In this section, we evaluate and compare the accuracy of derived saliency maps using collected human gaze data. Four test sequences (see Fig. 1) are used for our experiments, where `lovebird` is from the standard MPEG test sequence [19]; `jumprope` is available online[2] with associated human gaze data; `toy_fs` and `toy_fb` are captured using a combination of a RGB camera and a PMD® Time-of-Flight depth camera with proper view mapping.

A. Human Gaze Data Collection

We conducted subjective tests to collect human gaze locations during viewing of test sequences. As recommended by ITU-R BT.500 [20], 24 participants (15 male and 9 female, of age 22-30) took part in the tests. All participants had normal or corrected-to-normal vision, and were naïve about the task of the experiment. A 24-inch Dell LCD monitor with resolution 1920×1200 and default brightness $180\ cd/m^2$ was used for display. The ambient light in the room was 250-300 lux. Tobii® X-60 gaze-tracker was used to detect and collect participants' gaze locations, with sampling rate at 60 gaze points per second (pps). The angle of gaze-tracker was approximately 22 degrees to the ground and the distance between participants and gaze-tracker was around $50\ cm$.

Participants were asked to watch the test sequences (texture videos at $30\ fps$) sequentially in random order with gaze-tracker on. A 5-second black screen was inserted between sequences for rest. By repeating this process 10 times, we got approximately $24 \times 10 \times 60\ pps/30\ fps = 480$ gaze points per frame. Acquired points form a gaze density map[3] for each frame, which serves as ground-truth saliency map $\mathbf{SM}_{\text{gt}}$.

B. Performance Evaluation

In order to compare the accuracy of different saliency maps, we interpret saliency maps as 2D probability distributions of human attention and calculate the Kullback-Leibler divergence (K-L divergence) with respect to the ground-truth distribution.

$$\text{D}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j}\ln\left(\frac{\mathbf{P}(i,j)}{\mathbf{Q}(i,j)}\right)\mathbf{P}(i,j) \quad (14)$$

where ground-truth distribution $\mathbf{P} = \mathbf{SM}_{\text{gt}}/\sum_{i,j}\mathbf{SM}_{\text{gt}}(i,j)$, and the approximated distribution $\mathbf{Q} = \mathbf{SM}/\sum_{i,j}\mathbf{SM}(i,j)$ is calculated from different saliency maps shown in Table. I.

[2] http://www.irccyn.ec-nantes.fr/spip.php?article555

[3] http://www.youtube.com/user/leomanUST/videos

TABLE II
AVERAGE K-L DIVERGENCE VALUE

Sequence	lovebird	toy_fs	toy_fb	jumprope
Resolution	1024 × 768	640 × 480	640 × 480	720 × 576
Frames	100	75	75	100
D_{Itti}	3.0227	2.3745	2.3011	4.0494
D_{old}	2.4973	1.9904	2.0351	3.5413
D_{new}	2.4124	1.8480	1.8734	3.5209
$\frac{D_{old}-D_{new}}{D_{Itti}-D_{new}}$	13.91%	27.05%	37.81%	3.86%

(a) lovebird (b) toy_fs (c) toy_fb (d) jumprope

Fig. 1. Sample texture frames for 4 test sequences.

K-L divergence results are plotted frame-by-frame in Fig. 2; corresponding average values are shown in Table. II. It is clear that $D_{new} < D_{old} < D_{Itti}$, which implies: 1) incorporation of 3DMV channels ($\mathcal{A}, \mathcal{B}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$) improves the accuracy of derived saliency maps; and 2) the enhanced saliency maps using our new 3DMV channels ($\mathbf{SM}_{new}$) are more accurate than those using old 3DMV channels ($\mathbf{SM}_{old}$).

Note that above observations hold true not only in sequences where z-motion dominates conventional video features, but also in jumprope[4] where there is little z-motion. Therefore incorporating 3DMV channels into bottom-up saliency model always improves the accuracy of derived saliency maps, no matter how much z-motion there is in the scene.

Improvement of $\mathbf{SM}_{new}$ over $\mathbf{SM}_{old}$ is also shown in the last row of Table. II. In jumprope, the improvement (3.86%) is trivial because 3D motion reduces to 2D motion when z-motion is small. In lovebird, the improvement (13.91%) is non-negligible, but smaller than those in toy_fs and toy_fb because of the presence of human faces.

VI. CONCLUSION

In this paper, we first present our 3D motion estimation algorithm for texture-plus-depth videos. Using the computed 3DMVs, we then propose several saliency channels to improve an existing bottom-up saliency model. Experiments tracking human gaze show that the incorporation of 3DMV channels significantly improves the accuracy of derived saliency maps.

REFERENCES

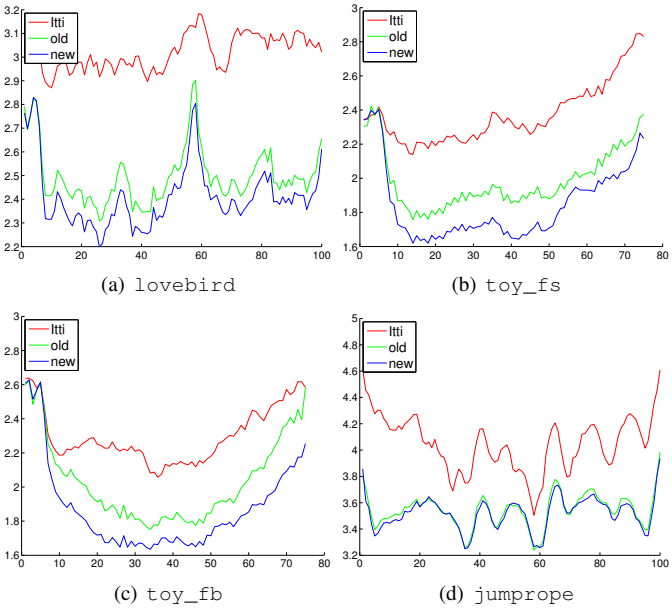
[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[2] N. Bruce and J. Tsotsos, "Spatiotemporal saliency: Towards a hierarchical representation of visual saliency," in *Attention in Cognitive Systems*. Springer Berlin Heidelberg, 2009, vol. 5395, pp. 98–111.

[3] 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.

[4] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features." *Vision Research*, vol. 47, no. 19, pp. 2483–98, 2007.

[4]depth frames are not available for jumprope, so we assume $mv_z = 0$.



(a) lovebird (b) toy_fs

(c) toy_fb (d) jumprope

Fig. 2. K-L divergence plot of 4 sequences. X-axis is the frame index and Y-axis is the K-L divergence value. Smaller value means higher accuracy.

[5] 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.

[6] 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.

[7] 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.

[8] J. Tsotsos, M. Pomplun, Y. Liu, J. Martinez-Trujillo, and E. Simine, "Attending to motion: Localizing and classifying motion patterns in image sequences," in *Biologically Motivated Computer Vision*. Springer Berlin Heidelberg, 2002, vol. 2525, pp. 439–452.

[9] P. Wan, Y. Feng, G. Cheung, I. V. Bajić, O. C. Au, and Y. Ji, "3d motion in visual saliency modeling," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.

[10] J. Y. Lin, S. Franconeri, and J. T. Enns, "Objects on a collision path with the observer demand attention," *Psychological Science*, vol. 19, no. 7, pp. 686–692, 2008.

[11] C. Howard and A. Holcombe, "Unexpected changes in direction of motion attract attention," *Attention, Perception and Psychophysics*, vol. 72, no. 8, pp. 2087–2095, 2010.

[12] B. Kamolrat, W. 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.

[13] 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.

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

[15] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 454–461.

[16] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205 – 231, 2005.

[17] A. Borji, M. Ahmadabadi, and B. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Machine Vision and Applications*, vol. 22, no. 1, pp. 61–76, 2011.

[18] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

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

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