# SEMIAUTOMATIC VISUAL-ATTENTION MODELING AND ITS APPLICATION TO VIDEO COMPRESSION

*Yury Gitman*[*]*, Mikhail Erofeev*[*]*, Dmitriy Vatolin*[*]*, Bolshakov Andrey*[†]*, Fedorov Alexey*[*]

[*] Lomonosov Moscow State University
[†] Institute for Information Transmission Problems

## ABSTRACT

This research aims to sufficiently increase the quality of visual-attention modeling to enable practical applications. We found that automatic models are significantly worse at predicting attention than even single-observer eye tracking. We propose a semiautomatic approach that requires eye tracking of only one observer and is based on time consistency of the observer's attention.

Our comparisons showed the high objective quality of our proposed approach relative to automatic methods and to the results of single-observer eye tracking with no postprocessing. We demonstrated the practical applicability of our proposed concept to the task of saliency-based video compression.

***Index Terms***— Saliency, Visual attention, Eye-tracking, Saliency-aware compression, H.264

## 1. INTRODUCTION

Modeling of visual saliency is a promising approach to improving the quality of many existing applications, such as image and video compression [1], description [2], quality measurement [3], and retargeting [4]. But each of these applications requires a model of visual attention to allow high-quality prediction of saliency.

Unfortunately, a recent comparison [5] revealed that most of the existing models of visual saliency fail to work as well as a simple model that prefers the center of the image. But such a center-prior model is entirely independent from the video content. In fact, even optimal blending of the best model and the center-prior model shows only a 0.037 AUROC (area under receiver operating characteristic) gain over center-prior. To compare, the center-prior model shows a 0.28 AUROC gain versus salt-and-pepper noise.

Significantly higher quality could be achieved through eye tracking of multiple observers, but the associated costs in time and money are extremely high, making it impractical.

We propose a trade-off between these two approaches: a semi-automatic visual-attention model (SAVAM). We use fixation points from just one human observer and apply automatic postprocessing. This postprocessing enables us to improve the robustness of the initial fixation points. Our method is inspired by the ability of human short-term memory to preserve information about a scene in time and use that information to interact with the environment, particularly for controlling future eye movements [6]. Such a trade-off seems to be reasonable because data from only one observer (not the usual tens of observers) is needed, and our proposed postprocessing method can significantly improve this data.

This work was partially supported by the Intel/Cisco Video Aware Wireless Networking (VAWN) Program.

In the objective comparison described in Section 4.2, we show that our algorithm outperforms state-of-the-art automatic visual-attention models and increases the similarity score [5] of single-observer eye tracking by 0.012.

The high quality of the proposed approach makes it suitable for practical applications; in particular, we used it in Section 5 to perform saliency-aware video compression. We achieved up 23 % lower bitrate than x264 encoder while keeping the same quality of salient region.

## 2. RELATED WORK

### 2.1. Models of visual attention

To the best of our knowledge no other research attempts to construct saliency maps semiautomatically. Therefore the most related efforts involve entirely automatic models of visual attention. Currently, two approaches predict visual attention: bottom up and top down [7].

The bottom-up approach assumes that attention is driven by the properties of an image. In [4] the saliency of the point is considered to be the uniqueness of a small surrounding area. The authors of [8] use the same definition of saliency, but they also perform postprocessing on the basis of pixel reciprocity and association of pixels into objects. In [9] *saliency* refers to the uniqueness of some of the image frequencies and is extracted in the Fourier domain. This idea is expanded to the case of video in [1] by using a multiscale pyramid of quaternion Fourier transforms for the initial image and motion-strength map. The authors of [10] propose a general algorithm to extract saliency from local image features. The feature map is transformed into a Markov chain with the edges marked using a normalized measure of distinctiveness and the spatial distance between nodes. The saliency map is the equilibrium distribution obtained using the random-walk algorithm.

The top-down approach assumes that attention is mostly driven by the viewer's goals and experience; In our estimation, the most remarkable model of top-down attention is described in [11]. Here, the authors use face, person and car detection together with multiple bottom-up features to train a per-pixel SVM classifier. They then consider the distance to the SVM hyperplane to be the saliency value. Although their proposed approach obviously cannot consider complex spatial relationships, it nevertheless demonstrates high scores in different comparisons [5, 8].

Although Yarbus in his work [12] described the important role of top-down mechanisms in determining eye movements, these mechanisms remain poorly studied; corresponding models are thus able at this point to produce only comparable results relative to bottom-up ones.

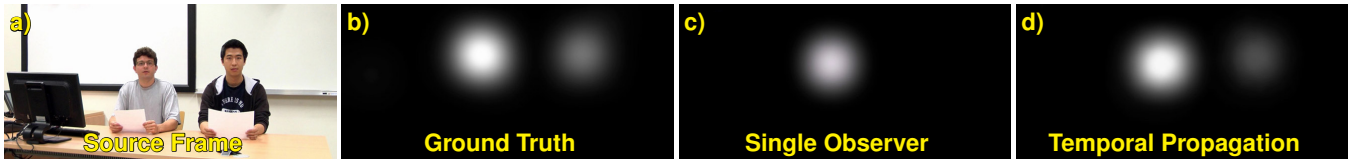


**Fig. 1**. Example of temporal propagation. **a)** Initial frame with multimodal saliency distribution. **b)** Ground truth from 50 free-viewing observers **c)** Saliency map constructed using gazes from a single observer. **d)** Saliency map from (c) after temporal propagation. The second focus of attention appeared because the observer looked at the second character in one of the surrounding frames.

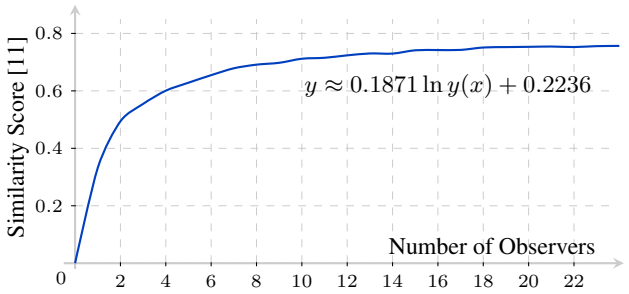


**Fig. 2**. Performance of $x$ observers to predict ground-truth saliency. The gazes of the first two observers have the greatest effect.

### 2.2. Saliency-based compression

The main idea of saliency-based compression is bit allocation in favor of salient regions. There are several implementations of this idea. We propose classification according to the following criteria:

- Model of visual attention underlying the method
- Reference encoder: MPEG-1 [13], MPEG-4 [1, 13], or H.264 [1, 14–18]
- Method of bit-allocation control: implicit [1, 13, 16] (video preprocessing before encoding; e.g., non-uniform blur) or explicit (modifying internal encoder data; e.g., setting saliency-specific individual quantization-parameter (QP) values for macroblocks) [14, 15, 17, 18]
- Evaluation methodology: Two different strategies exist; researchers can claim that videos encoded using their methods have lower bit rates than the reference video at the same visual quality [1, 13–16], or they can conclude that their proposed encoders can provide better visual quality than a reference at the same bit rate [17, 18].
- Method of visual-quality measurement: objective [15] or subjective [18]

## 3. DATABASE CREATION

Our research required high-quality gaze maps for a set of high-definition videos. Concerning increasing popularity of stereoscopic devices we created a gaze-map database for S3D full-HD videos. It consists of 43 sequences (approximately 13 minutes, or 19,760 frames) from well-known films and scientific databases that we received from the Laboratory for Image and Video Engineering of the University of Texas at Austin, the Video Quality Experts Group [19], and NTT Corporation.

We collected eye-tracking data from 50 people (mostly between 19 and 24 years of age) during task-free viewing, excepting the case of a special calibration pattern. We used a video-based eye-tracking system, the SMI iViewXTM Hi-Speed 1250, with a 500 Hz frequency for eye registration in binocular mode, as well as spatial resolution of up to one angular minute. To reduce inter-video influence we inserted cross-fade by adding a black frame between adjacent scenes.

Similarly to related works (e.g., [11, 20]) we used our collected data to create ground-truth saliency maps. We estimated the final ground-truth saliency map as a Gaussian mixture with centers at the fixation points. We chose a standard deviation for the Gaussians equal to 30 (considering the distance to the screen and its resolution, this value matches two angular degrees, which is known to be the sector of sharp vision).

All collected and computed data, including source videos and fixation points before any postprocessing, are available for download from http://compression.ru/video/savam/

## 4. SEMI-AUTOMATIC VISUAL ATTENTION MODEL

Most of the time the distribution of visual attention is strongly non-uniform, at least in the case of artistic content This distribution hints at the fact that gazes from just a few observers can produce near-ground-truth saliency maps. In [5] the exact dependence of similarity to ground truth (saliency map for all observers) on a number of observers was obtained empirically. We garnered a similar result for videos in our database. Figure 2 shows this result. Each point is the similarity score [5] between the saliency map for a chosen number of observers and the saliency map for the other observers. Each point was averaged over all frames and five different groups of observers.

At the same time, according to [5] existing models of visual attention offer little improvement over the center-prior model; our results (Section 4.2) indicate that not one can compete with eye-tracking, even for a single observer. Because we intend to apply our saliency framework to video compression, we require high-quality saliency maps.

In accordance with the above results, the only way to achieve such high quality is by using the eye-tracking procedure. But this approach is also unreasonable because of its laboriousness. One possible solution is a semiautomatic approach that uses fixation points from just one observer together with some postprocessing on the basis of temporal consistency of attention.

### 4.1. Temporal propagation

Our short-term memory retains a representation of our environment for some time [6]. In fact, an observer's next eye movement may be determined by short-term memory of the scene as much as by the current perception of it. This behavior can be viewed as temporal consistency of attention, i.e. objects that are salient in a certain frame are assumed to be salient in neighboring frames. This leads us to the idea of bidirectional temporal saliency propagation:

$$\mathbf{R} = \beta \mathbf{P}^{+} + (1 - \beta)\mathbf{P}^{-}, \tag{1}$$

where $\mathbf{R}$ is the result of the propagation, and $\mathbf{P}^{+}$ and $\mathbf{P}^{-}$ are forward and backward terms, respectively, defined as follows (depending on

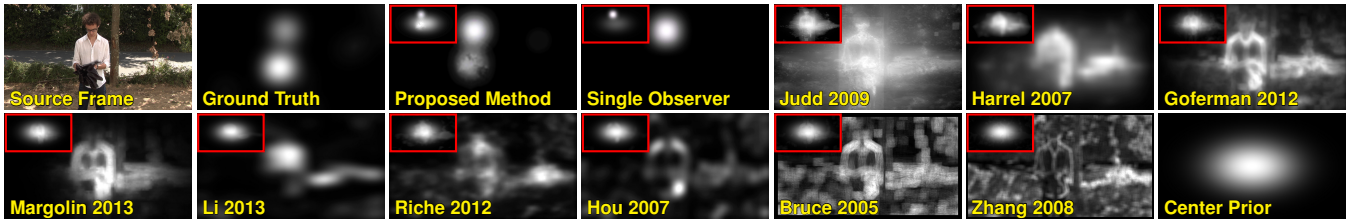


**Fig. 3**. Saliency maps predicted by different methods. The icons in the corner are the same images prepared for comparison (see Section 4.2). Histograms for all images are normalized for the sake of visibility. Also, the weakness of automatic methods is clearly visible in comparison with eye-tracking for even a single observer.

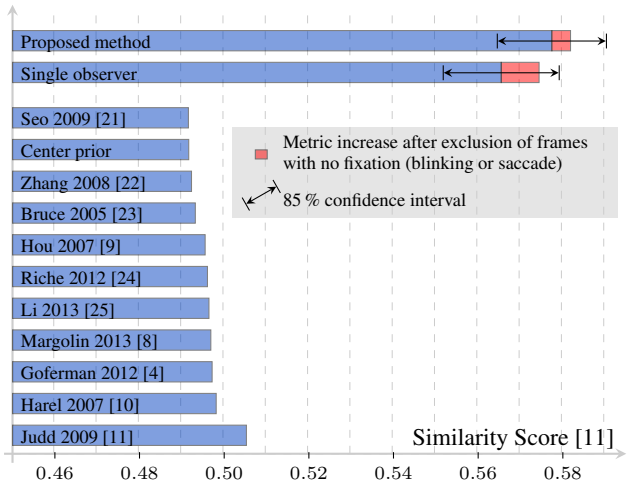


**Fig. 4**. Objective evaluation of our temporal propagation technique compared with other state-of-the-art saliency models, as well as and with the mean result of a single observer where no postprocessing has been applied.

the sign):

$$\mathbf{P}_t^{\pm}(x,y) = \alpha \mathbf{P}_{t\mp1}^{\pm}(x + v_x^{\pm}(x,y), y + v_y^{\pm}(x,y)) + (1-\alpha)\mathbf{S}_t(x,y). \quad (2)$$

Here, $\mathbf{S}$ is source sequence of saliency maps and $\vec{v}^{\pm}(x,y)$ is a motion vector field from $\mathbf{S}_{t\mp1}$ to $\mathbf{S}_t$. In our implementation, $\alpha = 7/10$ and $\beta = 1/2$. Vectors $\vec{v}^{\pm}(x,y)$ are computed using the motion-estimation algorithm described in [26].

This propagation technique is especially helpful for scenes with multiple saliency foci. Figure 1 shows one example.

### 4.2. Objective evaluation

We performed a quantitative evaluation of our proposed technique compared with multiple state-of-the-art saliency-prediction methods. The test videos and the ground-truth fixations were from our database described in Section 3. For the sake of fairness we exclude any fixations we used in our ground-truth method.

We wanted the results of the evaluation to be independent of the blending methods with center-prior as well as level and gamma correction, because these transformations are known to be able to significantly improve quality of the predicted saliency; at the same time, choosing them optimally to fit a specific task is easy. Therefore, we applied blending with center-prior, and gamma and level correction to each method with parameters adjusted for the best similarity-score value. Formally, these transformations can be written in the following way:

$$\mathbf{R} = (1-\eta)\cdot \mathrm{lv}(\mathbf{S}, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma) + \eta \mathbf{CP}, \quad (3)$$

$$\alpha_1, \alpha_2, \beta_1, \beta_2, \eta \in [0;1], \gamma \in \mathbb{R}^+ \quad (4)$$

Here $\mathbf{S}$ is a source saliency map, $\mathbf{R}$ is the resulting map, and lv function is levels and gamma correction.

$$\mathrm{lv}(\mathbf{S}, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma) = \frac{(\mathbf{S}-\alpha_1)^\gamma (\beta_2 - \beta_1)}{(\alpha_2 - \alpha_1)^\gamma} + \beta_1, \quad (5)$$

$$\mathbf{CP} = \mathrm{lv}(e^{-[(x-x_c)/\sigma_x]^2 - [(y-y_c)/\sigma_y]^2}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}). \quad (6)$$

The value of the lv function is computed in saturation arithmetic, so it is confined to the range of $\mathbf{S}$. The parameters $\sigma_x$, $\sigma_y$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\gamma}$ are chosen for the best match between $\mathbf{CP}$ and ground truth; $x_c$, and $y_c$ are the coordinates of the image center.

For each method $\alpha_1$, $\alpha_2$, $\beta_1$, $\beta_2$, $\eta$, $\gamma$ are optimized for the best similarity score using the interior-point algorithm. To avoid running into the local extremum, we perform optimization 100 times using randomly selected initial points and then chose the best one. We investigated the structure of the extremum distribution in our sampling. The hyperplane at $\eta = 1$ contains numerous trivial extrema, which appear in 28 % to 83 % of runs, depending on the method that we were adjusting. Interestingly, the best extremum is very probable among the remaining ones, and it appears in 57 % to 100 % of nontrivial runs. Each of the remaining points appears only once and is likely to have a low probability. The smaller the optimal value of $\eta$ is, the greater the probability of getting the best extremum. For methods with optimal $\eta < 8/10$, the best extremum appears in 88 % to 100 % of nontrivial runs.

Figures 3 and 4 show the results of the comparison. In contrast to [5] we found the results of single-observer eye tracking to be significantly better than those of the automatic saliency models. We believe our results are fairer, however, since the authors of [5] do not add the center-prior model to the eye-tracking results for the single observer.

The proposed temporal propagation technique increases the similarity score of single-observer eye tracking by 0.012.

A small part of this increase can be explained by frames that lack fixation data because of blinking or saccades. We additionally measured the score of our method only for frames with fixations. We found that the contribution of the excluded frames was only 38 % of overall quality improvement.

## 5. COMPRESSION

The semiautomatic approach we propose enables us to obtain saliency maps with significantly higher quality than those from automatic methods. We focused on improving video compression performance,


**Fig. 5**. Compression results of the proposed pipeline and x264 encoder for the same bit rate (1500 kbps). Quality differences between salient and non-salient regions are clearly visible. Degradation of quality for the proposed method in non-salient region has no significant effect.

in particular, we choose H.264 [27] as the most widely used video compression standard and x264 [28] as the most popular video encoder.

Estimated saliency maps are downscaled to match the dimensions of a macroblock grid (during our experiments we used a default macroblock of $16 \times 16$ pixels). Where $Q \in \mathbb{R}^+$ is the map of macroblock's QP values selected by encoder, and $S : \mathbb{R}^2 \to [0; 1]$ is a downscaled saliency map for the current frame, new QP values can be computed using the following equation:

$$Q' = \max(Q - \psi \cdot (S - \mathbb{E}S), 0). \tag{7}$$

Thus we reduce the QP value for macroblocks containing salient regions, and vice versa. The parameter $\psi$ is selected by the user and controls bitrate distribution between salient and non-salient regions: the greater the value, the more bits for salient areas.

We propose the following pipeline to implement this idea:

1. Run x264 with the following arguments `--qcomp 0 --pass 1 --bitrate` $\phi$, where $\phi$ is the target bitrate
2. Read QP values from `.mbtree` file produced by encoder and modify them in accordance with Equation 7
3. Run x264 with the following arguments `--qcomp 0 --pass 2 --bitrate` $\phi$

To relate the increase of similarity score reported in Section 4.2 and the increase of rate-distortion ratio reported below we execute the proposed pipeline for saliency maps of one human observer, two human observers, ground-truth (50 observers), center prior model, the best automatic model from our comparison [11], our proposed method and the same pipeline with Step 2 omitted.

Before measurements we optimally blended saliency models with center prior while keeping default values for levels and gamma correction, because the exact scheme described in Section 4.2 led us to unpredictable change in compression quality, that is clearly explained by the fact that the optimized function was different from the one used for measurement.

In the case of saliency-aware compression common metrics, e.g., PSNR and SSIM [29], fail to work correctly. To investigate the performance of saliency-aware encoding we used EWSSIM metric defined similarly to eye-tracking weighted PSNR [14]:

$$\mathrm{EWSSIM}(\mathbf{A}, \mathbf{B}, \mathbf{S}) = \frac{\sum_{i,j} \mathbf{S}_{i,j} \cdot \mathrm{SSIM}(\mathbf{A}, \mathbf{B})_{i,j}}{\sum_{i,j} \mathbf{S}_{i,j}}, \tag{8}$$

where $\mathbf{A}$ and $\mathbf{B}$ are source and reference frames, $\mathbf{S}$ is the ground-truth saliency for the reference frame, and $\mathbf{SSIM}(\mathbf{A}, \mathbf{B})$ is SSIM index map.

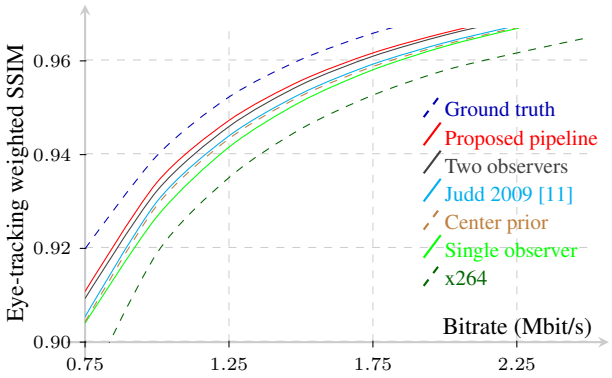


**Fig. 6**. Objective evaluation of our compression pipeline. We used saliency maps obtained with different visual attention models for saliency-aware x264-based video compression. By expending fewer bits on the non-salient area, we achieved a quality increase in the salient region up to 0.022 EWSSIM for the same bit rate.

Figure 6 shows the rate-distortion curves for proposed pipeline using different saliency models with $\psi = 50$ and for non saliency-aware x264 encoder. The objective measurements revealed that for the compression purposes the proposed saliency model outperform two human observers and all automatic models while using gazes from the only one observer. The proposed method achieves up to 0.022 EWSSIM increase over non saliency-aware x264 encoding for the same bitrate. Figure 5 shows example frames for subjectively estimating the quality difference.

## 6. CONCLUSION

In this paper we introduce a novel method for saliency-map estimation using postprocessing of eye-tracking data for a single observer. During our objective comparison, we showed that our method significantly outperforms other visual-attention models and saliency maps obtained from a single observer.

We also used the proposed method to design a saliency-aware video-compression framework. This framework enables us to surpass the performance of two human observers and to achieve a quality of the salient regions that is better than that of an x264 encoder by 0.022 EWSSIM yielding the same bit rate (or a bitrate that is lower than that of x264 encoder by 23 % yielding the same quality).

Additionally, eye-tracking dataset collected for this research is available to the scientific community.

**Acknowledgment:** This work was partially supported by the Intel/ Cisco Video Aware Wireless Networking (VAWN) Program.

## 7. REFERENCES


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

[2] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros, "Data-driven visual similarity for cross-domain image matching," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 6, pp. 154, 2011.

[3] Dubravko Culibrk, Milan Mirkovic, Vladimir Zlokolica, Maja Pokric, Vladimir Crnojevic, and Dragan Kukolj, "Salient motion features for video quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 4, pp. 948–958, 2011.

[4] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 10, pp. 1915–1926, 2012.

[5] Tilke Judd, Frédo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," Tech. Rep., Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 2012.

[6] María Pilar Aivar, Mary M. Hayhoe, Christopher L. Chizk, and Ryan E. B. Mruczek, "Spatial memory and saccadic targeting in a natural task," *Journal of Vision*, vol. 5, no. 3, 2005.

[7] Michael Land and Benjamin Tatler, "How our eyes question the world," in *Looking and Acting: Vision and Eye Movements in Natural Behaviour*. Oxford University Press, 2009.

[8] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal, "Saliency for image manipulation," *The Visual Computer*, pp. 1–12, 2013.

[9] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[10] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, vol. 19, pp. 545–552.

[11] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *International Conference on Computer Vision (ICCV)*, 2009, pp. 2106–2113.

[12] Alfred Yarbus, "Eye movements during perception of complex objects," in *Eye Movements and Vision*, pp. 171–211. Plenum Press, 1967.

[13] Laurent Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 10, pp. 1304–1318, 2004.

[14] Zhicheng Li, Shiyin Qin, and Laurent Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, vol. 29, no. 1, pp. 1–14, 2011.

[15] Rupesh Gupta, Meera Thapar Khanna, and Santanu Chaudhury, "Visual saliency guided video compression algorithm," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1006–1022, 2013.

[16] Shao-Ping Lu and Song-Hai Zhang, "Saliency-based fidelity adaptation preprocessing for video coding," *Journal of Computer Science and Technology*, vol. 26, no. 1, pp. 195–202, 2011.

[17] H. Hadizadeh and I.V. Bajic, "Saliency-preserving video compression," in *International Conference on Multimedia & Expo (ICME)*, 2011, pp. 1–6.

[18] Hadi Hadizadeh, *Visual Saliency in Video Compression and Transmission*, Ph.D. thesis, School of Engineering Science, Simon Frazer University, 2013.

[19] Matthieu Urvoy, Marcus Barkowsky, Romain Cousseau, Yao Koudota, Vincent Ricorde, Patrick Le Callet, Jesús Gutiérrez, and Narciso García, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 109–114.

[20] Qi Zhao and Christof Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, 2011.

[21] Hae Jong Seo and Peyman Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *IEEE International Workshop on Computer Vision and Pattern Recognition*, 2009, pp. 45–52.

[22] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[23] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 155–162.

[24] Nicolas Riche, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, "RARE: A new bottom-up saliency model," in *International Conference on Image Processing (ICIP)*, 2012, pp. 641–644.

[25] Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 4, pp. 996–1010, 2013.

[26] K. Simonyan, S. Grishin, D. Vatolin, and D. Popov, "Fast video super-resolution via classification," in *International Conference on Image Processing (ICIP)*, 2008, pp. 349–352.

[27] ITU-T recommendation, "H.264: "advanced video coding for generic audiovisual services"," *ISO/IEC*, vol. 14496, 2003.

[28] "x264 software video encoder," http://www.videolan.org/developers/x264.html.

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.