A Study on Interest Point Guided Visual Saliency

Xiang Zhang, Shiqi Wang, Siwei Ma and Wen Gao
Institute of Digital Media, Peking University, Beijing, China
Cooperative Medianet Innovation Center, Shanghai, China

***Abstract*—Visual attention is one of the most critical characteristics of human visual system (HVS), which infers the attractive regions in a visual scene. It has been an active research topic over the past decades and many proposed models of visual attention have demonstrated successful applications in a wide range of fields including computer vision and image processing. On the other hand, interest point detection is another hot topic that leads practical contributions to the real-time applications such as visual retrieval and augmented reality. In this paper, we try to investigate the relationship between the interest point and the visual attention. An informative analysis is reported by comparing the performance of different interest point models in predicting the visual fixation. It is found that the blob based interest point model generally outperforms the corner based model. Furthermore, we propose a mixture strategy by integrating all the interest point algorithms, and the experimental results indicate that this proposed method is competitive with some state-of-the-art algorithms.**

I. Introduction

Human eyes will always receive and process huge amount of visual elements every second, and some key objects or targets that are striking to human visual system (HVS) would naturally stand out from other negligible background. It is referred to as the visual saliency mechanism in visual perception where the "saliency" is a general term indicating the interesting regions and always it will be the foreground of a particular scene. Similarly, the machines "read" the input visual contents as a rich data stream. Due to the disaster of huge data size, it is important to extract "core data" corresponding to the "salient regions" from the raw data by mimicking the visual saliency mechanism of HVS. Visual saliency is essential in the field of computer vision for cognitive tasks and also demonstrates its efficiency in image quality assessment [1].

In the last decades, many successful models aiming at accurately predict the visual saliency have been studied. One major taxonomy tries to classify all the models into three categories, which are top-down, bottom-up and combined methods respectively according to whether the characteristics of HVS are explicitly utilized in the scheme [2]. The top-down methods always have some complex and cognitive process that incorporating psychophysical and neurophysiological prior knowledge. These kind of schemes are typically slow and application-specific, such as the human face recognition. On the contrary, the bottom-up methods work in a general way by utilizing some simple and efficient techniques. In [3], [4], visual information/saliency is estimated by the statistics of sparse primitives.

An interest point is a clear, well-defined, mathematically well-founded position in an image space and can be detected robustly with illuminance variations as well as geometrical changes including translation, rotation, scaling etc. Interest point detection techniques are critical in many real-time applications including visual retrieval and augmented reality, and can be categorized into corner based methods and blob based methods. A corner can be defined as the intersection of two edges and always considered as an fixation point that will attract human attention. A blob region in an image typically contains different properties, such as brightness or color, compared to the surrounding areas.

In this paper, we investigate the relationship between the interest point and the visual saliency and try to find an effective bottom-up way in predicting visual saliency. The motivation is from the well-known *foveation effect* in visual perception, where the resolution in the retina rapidly decreases with the increasing distance to the fixation centra. It assumes that the response (or sensitivity) of HVS on different image regions is distinct as the human eyes always focus on some interest regions. When one fixates at a point, the region around it is sampled with the highest resolution, while the remote region would be felt like blurred. It is because of the nonuniform distributions of cone receptors and ganglion cells in the retina.

The contribution of this paper is twofold. First, we compare the performance of several popular interest point models in predicting visual saliency. Second, a mixture model is proposed by incorporating multiple interest point algorithms and demonstrates competitive performance with other models.

The remaining of this paper is organized as follows. In Section II we introduce some state-of-the-art techniques of interest point detection, then the visual saliency map is generated by incorporating different models of interest point technique. The comparison results of saliency performance are provided in Section III. Section IV concludes this paper.

II. Interest Point Guided Visual Saliency

A. Interest Point Detection

An interest point is a clear, well-defined, mathematically well-founded position in an image space and can be detected robustly with illuminance variations as well as geometrical changes including translation, rotation, scaling etc. In this paper we divide the interest point detection algorithms into two categories, corner based methods and blob based methods according to its methodology.

1) Corner Methods: A corner can be defined as the intersection of two edges or a point for which there are two dominant edge directions in a local neighbourhood of the point. Without loss of generality, let I denote the input image.

Considering a local patch (u, v) and a corresponding patch with a small shifting by (x, y). The weighted sum of squared difference between the two patches can be written as follows,

$$S(x,y) \approx \begin{pmatrix} x & y \end{pmatrix} A \begin{pmatrix} x \\ y \end{pmatrix}, \quad (1)$$

$$A = \sum_u \sum_v w(u,v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = \begin{bmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{bmatrix}, \quad (2)$$

where I_x and I_y are the partial derivatives of image I. The notation $\langle \bullet \rangle$ denotes the summation operator over u and v. In [5], corner is detected by computing the $min(\lambda_1, \lambda_2)$, where λ_1 and λ_2 are two eigenvalues of A. To avoid the computational expenses in solving the eigenvalue, famous Harris Corner method [6] is proposed by using the following function M_c for simplification,

$$M_c = \lambda_1 \lambda_2 - \kappa(\lambda_1 + \lambda_2)^2 = \det(A) - \kappa \cdot trace^2(A), \quad (3)$$

where the computing of eigenvalue is replaced by evaluating the determinant and trace of the matrix A.

Features from accelerated segment test (FAST) [7] is a high-efficiency corner detection method, which is faster than many other well-known feature extraction methods and suitable for real-time image/video processing applications. FAST corner detector uses a circle of 16 pixels (a Bresenham circle of radius 3) to classify whether a candidate point p is actually a corner. If a set of N contiguous pixels in the circle are all brighter or darker than the intensity of candidate pixel p (denoted by I_p) plus a threshold value t, then p is classified as corner. More recently, the robust BRIEF [8], BRISK [9] and ORB [10] descriptors are proposed for more fast and accurate performance in feature detection and matching as the extensions of the FAST version.

2) Blob Methods: In the field of computer vision, blob detection refers to mathematical method aiming at detecting regions in an image that differ in properties, such as brightness or color, compared to areas surrounding those regions. Blob detector has been applied in many feature detection techniques including Scale Invariant Feature Transform (SIFT) [11] and Speeded Up Robust Features (SURF) [12].

Laplacian of Gaussian (LoG) is one of the most common blob detectors. Given an input image $f(x, y)$, this image is convolved by a Gaussian kernel g as follows,

$$L(x,y,t) = g(x,y,t) * f(x,y). \quad (4)$$

where t is the scale parameter. Then the Laplacian operator is applied to the blurred image, and the extreme points with maximum or minimum values are detected in the multi-scale LoG space. The LoG operator is effective and efficient in extracting the feature points.

However the LoG has a disadvantage of big computational complexity induced by the second derivative in Laplacian operator. To overcome this, an approximation method called Difference of Gaussian (DoG) is proposed by computing the difference between two adjacent Gaussian smoothed images in scale space as follows,

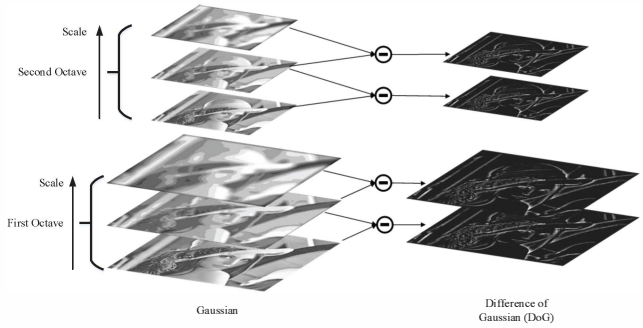

Fig. 1. Illustration of the Difference of Gaussian (DoG) [11].

$$\begin{aligned} D(x,y,t) &= L(x,y,t+1) - L(x,y,t) \\ &= (g(x,y,t+1) - g(x,y,t)) * f(x,y). \end{aligned} \quad (5)$$

This process, which is applied in the famous SIFT descriptor, is schematically illustrated in Fig. 1.

For SURF descriptor, the Determinant of Hessian (DoH) operator is used for extracting the feature points. The Hessian Matrix of an image is defined as,

$$HL = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}, \quad (6)$$

thus the DoH can be calculated as follows,

$$DoH(x,y,t) = \det HL(x,y,t) = L_{xx} L_{yy} - L_{xy}^2. \quad (7)$$

Maximally Stable Extremal Regions (MSER) [13] is another method of blob detection in images by finding correspondences between image elements from two images with different viewpoints. This method of extracting a comprehensive number of corresponding image elements contributes to the wide-baseline matching, and it has led to better stereo matching and object recognition algorithms.

B. Interest Point Guided Visual Saliency

Given the extracted interest points, we need to generate the corresponding saliency map. Intuitively, the salient value of a point is dependent to the distance from interest points. If a point is far away from all the other interest points, a small saliency value should be assigned to this point. On the contrary, the region with many interest points is confident to be the visual fixation. Generally, suppose point p be an interest point, the saliency value of another point p' is dependent to the distance between them. The farther the distance is, the saliency value will be smaller. We utilize the normalized Gaussian function to build this model as follows,

$$s(p'|p) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{d(p,p')^2}{2\sigma^2}}, \quad (8)$$

where the $s(p'|p)$ indicates the saliency value of the point p' given point p. The distance $d(p, p')$ represents the Euclidean distance between them. The parameter σ is the variance of the

(a) Controller (b) Bicycle (c) Lawn

(d) Smart car (e) Children (f) Stuff

Fig. 2. Example images in Bruce's database [14]. Images in first row have single fixation target and images in second row contain multiple targets.

Gaussian distribution that controls the change rate of visual importance around the fixation point. Based on Eqn. (8), the salient map can be obtained given the position of all the interest points. To alleviate the mistakes caused by some wrong detected noise points, the salient map is calculated in a simpler but more robust way. Let $P_{x,y}$ be a map indicating whether a point at position (x, y) is an interest point or not,

$$P_{x,y} = \begin{cases} 1, & \text{point at } (x,y) \text{ is an interest point;} \\ 0, & \text{otherwise.} \end{cases} . \quad (9)$$

Then the saliency map is generated as follows,

$$\mathbf{S} = \mathbf{G} * \mathbf{P}, \quad (10)$$

where $\mathbf{G}$ is a $w \times w$ Gaussian kernel with the variance σ, $\mathbf{P}$ is the indicating map defined in Eqn. (9) with the identical size of input image. The $\mathbf{S}$ represents the saliency map generated by convolving $\mathbf{G}$ over $\mathbf{P}$.

III. Experiments and Analysis

In order to evaluate the performance of different interest point models on visual saliency, extensive simulations are conducted with the following seven popular interest point detection algorithms including four corner based methods (Harris [6], FAST [7], ORB [10], BRISK [9]) and three blob based methods (SIFT [11], SURF [12], MSER [13]).

A. Database and Parameter Settings

The experiments are conducted on the famous eye tracking database developed by Bruce and Tsotsos in 2005 [14]. The Bruce's database consists of 120 color images, all of which are with the same resolution of 681×511 including different scenes. Some examples of the dataset are demonstrated in Fig. 2 where images in first row have single fixation target and images in second row contain multiple targets.

Each algorithm of interest point detection has its own parameters that control the number of detected points. It is unfair if different methods generate distinct number of point when comparing the performance. To solve this problem, we introduce a ranking strategy to make sure every methods are with the identical ability of extracting feature points. All the detected points are reordered by the *response value*, which indicates the confidence of a point being a interest point. And the first K points with maximum *response value* are drawn out for further evaluation. In this paper, K is fixed as 500.

The parameters w and σ of the Gaussian kernel $\mathbf{G}$ are uniformly governed by one parameter β as follows,

$$\begin{aligned} w &= round(\beta \times W) \times 4, \\ \sigma &= \beta \times W, \end{aligned} \quad (11)$$

where W indicates the width of input image. We compare the generated saliency maps by several interest point models under different parameter β in Fig. 3. The test image is the *Smart Car* demonstrated in Fig. 2(d) which contains multiple interest objects. It can be seen that the saliency map is disturbed by many noise points when the β is small ($\beta = 0.01$). Then the saliency map becomes more robust and tends to peak at some concentrated areas with the increasing value of β. However the β could not be too large as the true interest regions would be over-blurred and vanished from the saliency map. In this paper β ranges from 0.01 to 0.08 by the step of 0.01.

B. Performance Comparison of Single Models

We utilize the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) score as the evaluation metrics, which are the most popular measurements of the visual saliency. ROC is widely used for evaluating the performance of a binary classification system with a variable threshold. Thus the predicted saliency map is treated as a binary classifier on each pixel of the image. Pixels with larger saliency value then a given threshold are classified as salient pixels and vice versa. By using the human fixations as the ground truth, the ROC curve can be drawn as the false positive rate versus true positive rate by sweeping over all the thresholds, and the area under the ROC curve, indicating how well the saliency model predicts the visual saliency of human eyes, is defined as the AUC score. As stated in [15] & [16], human fixations have strong center-bias that may affect the performance of the saliency algorithm. To remove the center-bias effect, the positive sample set is composed of the fixation points of all subjects on this image, whereas the negative sample set consists of the union of all fixation points across all images from the database except for the positive samples.

The comparison results over seven feature point models are provided in Tab. I. The corresponding curves of AUC in terms of β are drawn in Fig. 4, where the optimal parameter of each algorithm is marked as a solid dot. Note that the dotted line indicates the corner based method and the dashed line indicates the blob based method. It is obvious from the Fig. 4 that the optimal parameter is either 0.04 or 0.05 for the seven methods. Thus we can infer that the setting of $\beta = 0.04 \sim 0.05$ is generally the best choice in most cases. Second the blob methods outperform the corner methods according to both the table and the plots, and the SURF and MSER are the two champion algorithms in the competition.

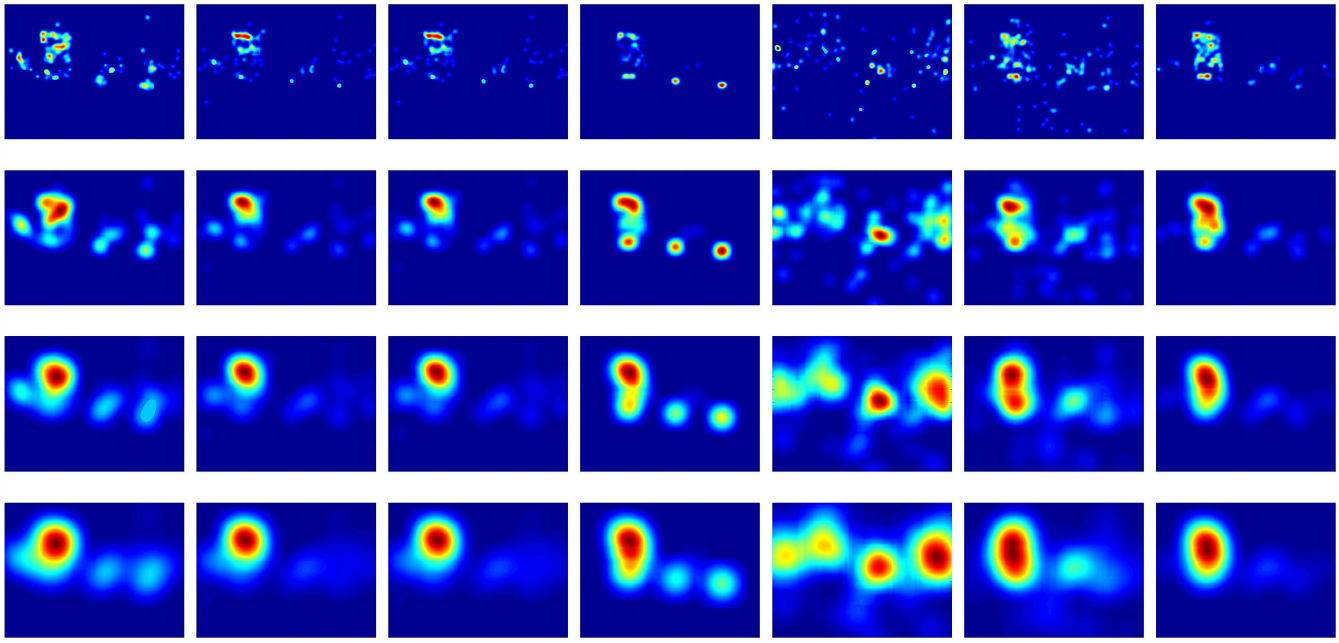

Fig. 3. Illustrations of saliency map generated by different interest point models under different β. The saliency maps are generated by Harris, FAST, ORB, BRISK, SIFT, SURF and MSER respectively from left to right. The parameter β is set as 0.01, 0.03, 0.05 and 0.07 respectively from top to bottom. The test image is the *Smart Car* shown in Fig. 2(d).

TABLE I

THE PERFORMANCE COMPARISON OF VISUAL SALIENCY UNDER SEVEN SINGLE INTEREST POINT MODELS AND THREE MIXTURE MODELS IN TERMS OF AUC VALUES ACROSS DIFFERENT PARAMETER β. THE LAST TWO ROWS INDICATE THE AVERAGE AUC AND MAXIMUM AUC VALUES OVER ALL β RESPECTIVELY. THE FIRST FOUR, MIDDLE THREE AND LAST THREE COLUMNS ARE THE CORNER METHODS, BLOB METHODS AND THE MIXTURE METHODS RESPECTIVELY. THE TOP THREE MODELS AMONG ALL THE SEVEN SINGLE MODELS ARE HIGHLIGHTED BY BOLDFACE IN THE LAST TWO ROWS.

β	**Harris**	**FAST**	**ORB**	**BRISK**	**SIFT**	**SURF**	**MSER**	**Corner**	**Blob**	**ALL**
0.01	0.5939	0.5978	0.5758	0.5861	0.5927	0.6129	0.5769	0.6276	0.6334	0.6463
0.02	0.6389	0.6438	0.6235	0.6246	0.6382	0.6598	0.6419	0.6615	0.6709	0.6762
0.03	0.6544	0.6579	0.6402	0.6361	0.6540	0.6722	0.6650	0.6697	0.6815	0.6833
0.04	0.6585	0.6595	0.6451	0.6395	0.6582	0.6742	0.6715	0.6699	0.6832	0.6827
0.05	0.6587	0.6579	0.6452	0.6386	0.6585	0.6727	0.6712	0.6660	0.6808	0.6788
0.06	0.6571	0.6539	0.6428	0.6366	0.6570	0.6704	0.6697	0.6620	0.6783	0.6739
0.07	0.6545	0.6496	0.6385	0.6329	0.6544	0.6671	0.6666	0.6572	0.6745	0.6686
0.08	0.6502	0.6446	0.6338	0.6282	0.6500	0.6632	0.6626	0.6522	0.6694	0.6627
AUC (avg. β)	**0.6458**	0.6456	0.6306	0.6278	0.6454	**0.6616**	**0.6532**	0.6583	0.6715	0.6716
AUC (opt. β)	0.6587	**0.6595**	0.6452	0.6395	0.6585	**0.6742**	**0.6715**	0.6699	0.6832	0.6833

C. Performance Comparison of Mixture Models

According to the performance of single models, it can be concluded that the different model is adaptive in extracting some specific kinds of interest points. The higher performance of a single model indicates the stronger generalization or robustness ability. Therefore, we wonder what if some of the methods are combined as a hybrid model.

Let $\mathbf{P}^i$ indicate the point map of method i defined in Eqn. (9), where $i \in \{$Harris, FAST, ORB, BRISK, SIFT, SURF, MSER$\}$. Thus the mixed map $\mathbf{P}^{MIX}$ can be obtained,

$$P_{x,y}^{MIX} = \begin{cases} 1, & \exists\ i,\ P_{x,y}^{i} = 1 \\ 0, & \text{otherwise.} \end{cases} . \tag{12}$$

Then the saliency map can be obtained in the same way. In the experiments, we generate three kinds of mixture models which uses corner methods only, blob methods only and all methods respectively. The corresponding AUC curves and the detailed results are given in Fig. 4 and Tab. I. It is obvious that the performance of the mixture model by all methods is much higher than the single models indicating that the compound model can significantly improve the accuracy of visual map. The mixed blob model leads more contributions to the overall performance gain than the mixed corner model.

We also compare the proposed mixture model with some state-of-the-art salient models including: the Itti-Koch saliency

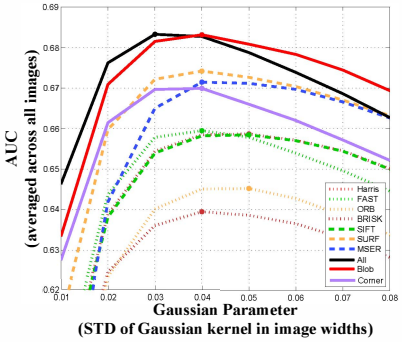

Fig. 4. The AUC curves of different interest point models in terms of β. For each algorithm, the optimal β is labeled as a dot on the plot. The dotted line is used for the corner based methods while the dashed line is for the blob based methods. The solid lines indicate the three mixture methods: the black line is mixed by all seven models, the red one is mixed by three blob methods and the purple one is mixed by four corner methods.

model (Itti) [17], Dynamic Visual Attention model (DVA) [18], Graph-Based visual saliency (GBVS) [19], Attention based on Information in maximization (AIM) [20], and Saliency Using Natural image statistic (SUN) [21]. The results are listed in Tab. II from which we can see that the mixture model is competitive with the existing saliency models and even outperforms some of them.

TABLE II
THE PERFORMANCE COMPARISON BETWEEN THE PROPOSED MIXTURE MODEL AND SEVERAL STATE-OF-THE-ART MODELS.

	AUC (opt. β)	AUC (avg. β)
DVA	0.6216	0.6213
Itti	0.6524	0.6517
GBVS	0.6782	0.6760
Prop.	**0.6833**	**0.6716**
SUN	0.6872	0.6849
AIM	0.7000	0.7000

IV. Conclusions

In this paper, we investigate the relationship between interest point and visual saliency. It is assumed that the region with many interest points has more confidence to be the fixation of human eyes. Seven state-of-the-art interest point algorithms are compared in this work. We conclude that the blob based methods can achieve better performance compared to the corner based methods in predicting the visual saliency. Furthermore we propose a mixture model by integrating all the interest point algorithms and experimental results have shown that this hybrid strategy achieves competitive results with some state-of-the-art saliency models.

Acknowledgment

This work was supported in part by the Major State Basic Research Development Program of China (2015CB351800) and in part by the National Science Foundation (61322106, 61390515 and 61210005). This work was also granted by Cooperative Medianet Innovation Center.

References

[1] X. Zhang, S. Wang, S. Ma, and W. Gao, "Towards accurate and efficient image quality assessment with interest points," in *Multimedia Big Data (BigMM), IEEE International Conference on*, 2015.

[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.

[3] X. Zhang, S. Wang, S. Ma, S. Liu, and W. Gao, "Entropy of primitive: A top-down methodology for evaluating the perceptual visual information," in *Visual Communications and Image Processing (VCIP), 2013*. IEEE, 2013, pp. 1–6.

[4] X. Zhang, S. Wang, S. Ma, and W. Gao, "Towards accurate visual information estimation with entropy of primitive," in *Circuits and Systems (ISCAS), IEEE International Symposium on*, 2015.

[5] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[6] C. Harris and M. Stephens, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

[7] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 105–119, Jan 2010.

[8] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 778–792.

[9] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2548–2555.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. IEEE, 1999, pp. 1150–1157.

[12] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[14] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[15] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[16] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 194–201, 2012.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 2009, pp. 681–688.

[19] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[20] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of vision*, vol. 9, no. 3, p. 5, 2009.

[21] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, p. 32, 2008.