2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Combined Model for Scan Path in Pedestrian Searching

Lijuan Duan, Zeming Zhao, Wei Ma*, Jili Gu, Zhen Yang
College of Computer Science and Technology
Beijing University of Technology, China
{ljduan, mawei, yangzhen}@bjut.edu.cn
{zhaozeming, gujili}@emails.bjut.edu.cn

Yuanhua Qiao
College of Applied Science
Beijing University of Technology, China
qiaoyuanhua@bjut.edu.cn

***Abstract*—Target searching, i.e. fast locating target objects in images or videos, has attracted much attention in computer vision. A comprehensive understanding of factors influencing human visual searching is essential to design target searching algorithms for computer vision systems. In this paper, we propose a combined model to generate scan paths for computer vision to follow to search targets in images. The model explores and integrates three factors influencing human vision searching, top-down target information, spatial context and bottom-up visual saliency, respectively. The effectiveness of the combined model is evaluated by comparing the generated scan paths with human vision fixation sequences to locate targets in the same images. The evaluation strategy is also used to learn the optimal weighting coefficients of the factors through linear search. In the meanwhile, the performances of every single one of the factors and their arbitrary combinations are examined. Through plenty of experiments, we prove that the top-down target information is the most important factor influencing the accuracy of target searching. The effects from the bottom-up visual saliency are limited. Any combinations of the three factors have better performances than each single component factor. The scan paths obtained by the proposed model are optimal, since they are most similar to the human vision fixation sequences.**

Keywords—visual attention; bottom-up visual saliency; top-down target information; spatial context

I. INTRODUCTION

Human visual attention, one of the most important mechanisms in biological vision systems [1], [3], [4], guides us to fast locate a specific kind of targets in images. A comprehensive understanding of factors influencing human visual searching is essential to design computer vision systems. In this paper, we explore three factors, bottom-up visual saliency, top-down target information and spatial context, which influence human vision systems to search targets (pedestrians) in images. The factors have been experimentally evaluated, separately or integratedly in literatures [5], [6], [7]. The paper presents a combined model which integrates the three factors with optimal weights, to guide target searching for computer vision systems. The weights are learned by linear search [2]. The performance of the combined model on the generation of scan paths is evaluated by comparing with human vision scan paths.

Psychological studies show that at each moment, humans are attracted to salient parts in images [6], [8], [9]. The bottom-up saliency clue is considered to have influences on computer visual searching, which has been experimentally proved by Itti et al. [10]. On the other hand, during visual searching, humans not only fixate on a target, but also scan regions or objects with similar shapes to the target [11], [12]. For example, during searching for a pedestrian, objects of a rectangular shape, or with a circle on the top would attract attention. The spatial context information provides rich cues to target positions for human vision [13], [14], [15]. It is widely used in object detection [14] and recognition [16].

Based on the above facts, the paper experimentally explores each factor and presents a method to combine them for efficient target searching in images. The proposed method is given in section II. Section III

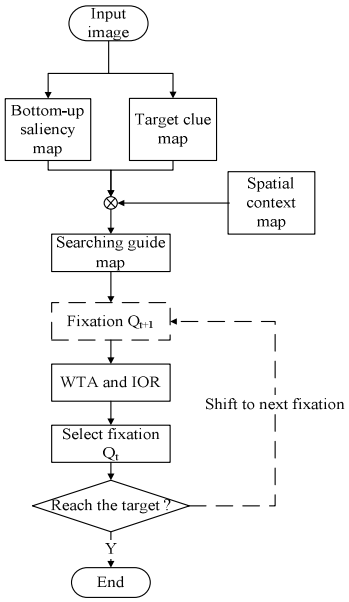


Fig. 1. The workflow of scan path generation. The saliency map and target map are computed based on the input image. The searching guide map is obtained by combining the spatial context map. At each round of fixation choosing, the strategies of WTA and IOR are used.

This research is partially sponsored by Natural Science Foundation of China (Nos.61003105, 61175115 and 61370113), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (CIT&TCD201304035), Jing-Hua Talents Project of Beijing University of Technology (2014-JH-L06), and Ri-Xin Talents Project of Beijing University of Technology (2014-RX-L06), and the International Communication Ability Development Plan for Young Teachers of Beijing University of Technology (No.2014-16).

(a) (b)

Fig. 2. Example of the Bottom-up Saliency Map: (a) is the original image; (b) is the saliency map. The lighter regions in (b) have higher probabilities to be fixated on.

evaluates the performance of our model. Conclusions are given in section IV.

II. OUR METHOD

The workflow of our method is given in Fig. 1. Firstly, for each image, we compute its bottom-up saliency map and target clue map. The spatial context map is learned from a database of images. Secondly, a searching guide map is obtained by combining the three maps. Then fixation regions in the image are sequentially chosen by using the strategies of winner-takes-all (WTA) and inhibition-of-return (IOR) [10] based on the searching guide map. The sequential fixations form a scan path searching for pedestrians in the image. The first fixation is initialized at the image center, which is consistent with the center bias in oculomotor behavior.

A. Computation of bottom-up visual saliency

Saliency, bottom-up visual clue, indicates regions attracting human attention. Although it is independent of the search task [17], it can improve the performance of pedestrian detection [10].

We compute the saliency map by using the spatially weighted dissimilarity method recently proposed by Duan et al. [18]. The method integrates dissimilarity, spatial distance and central bias. The image is regularly divided into patches of a fixed size. The saliency of image patch p_i is given by:

$$S(p_i) = w_1(p_i, C) \sum_{j=1}^{L} \{ w_2(p_i, p_j) . D(p_i, p_j) \} \quad (1)$$

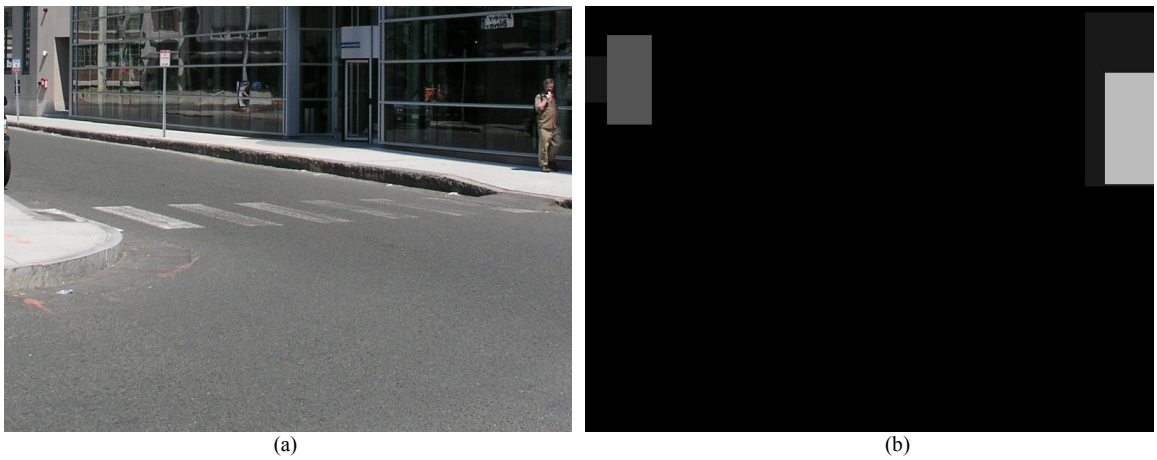
Here, $w_1(p_i, C)$ is the central bias term, which is inversely proportional to the distance between p_i and the image center C. The remaining parts in the right of (1) define the global dissimilarity. L is the total number of image patches. $w_2(p_i, p_j)$ is the inverse of the spatial distance between p_i and p_j. $D(p_i, p_j)$is the dissimilarity between p_i and p_j.

From (1), we can see that with the increasing of the spatial distance between p_i and p_j, the influence on $S(p_i)$ from $D(p_i, p_j)$ is decreasing. On the other hand, the less the distance between p_i and the image center C, the larger $S(p_i)$.

Examples are shown in Fig. 2, (a) is the original image, (b) is the saliency map.

B. Computation of top-down target clues

Target information serves as a referent for searching a target in the image. A large number of psychology experiments show that during target searching, humans



(a) (b)

Fig. 3. Example of the Target Clue Map. The rectangle regions indicate all the target-like items.

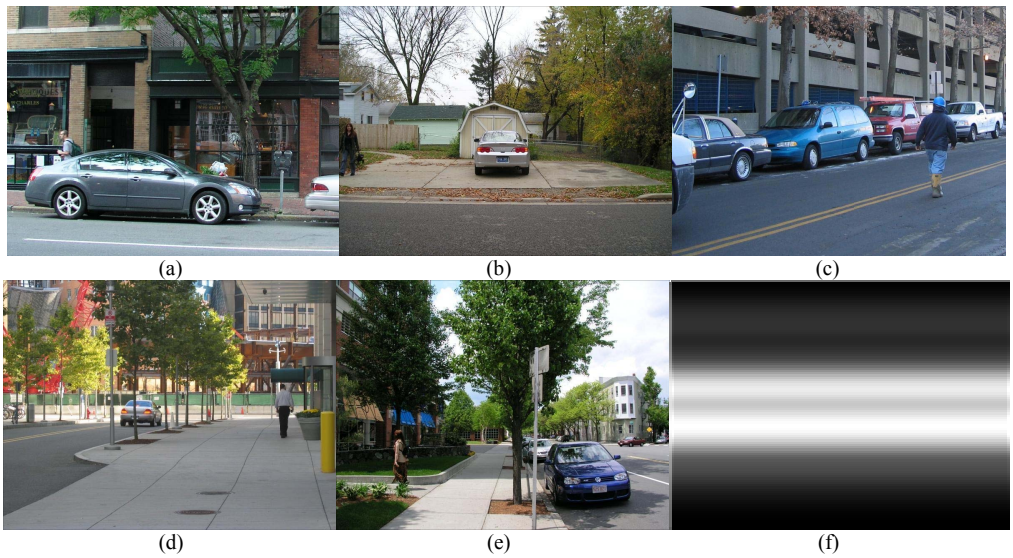


Fig. 4. (a) to (e) are image samples containing pedestrian, (f) is the statistical probability map of spatial context, which indicates that pedestrians are more likely to appear in the middle of images.

would not only fixate on the real target, but also scan regions or objects with similar shapes to the target. Therefore, we compute the target clue map by using the computational model proposed by [19], which is capable of indicating all the regions with similar shapes to the target in the image. Some examples are shown in Fig. 3.

C. Computation of spatial context

Spatial context provides a holistic description of the relationship between the target and its background scenes. It is a useful prior for human vision to fast locate the target in the image. To compute a spatial context map for pedestrians, we borrow the context oracle maps marked by different participants in [20], each of which indicates the region of pedestrians in different images. Our spatial context map is the average of all the context oracle maps. As shown in Fig. 4, (a) to (e) are samples of images containing pedestrians. (f) is the spatial context map. The pixel brightness in (f) denotes the probability of the existence of pedestrians at that position.

D. Combination of three factors

We generate a guide map to guide searching pedestrians. The searching guide map is obtained by linearly weighting the three maps. Given a pixel q_i, its value in the guide map is computed as:

$$M(q_i) = K_1 M_S(q_i) + K_2 M_T(q_i) + K_3 M_C(q_i) \quad (2)$$

Where $M_S(q_i)$, $M_T(q_i)$ and $M_C(q_i)$, represent the pixel values in the saliency map, the target clue map and the spatial context map, respectively. M will change during the fixation selection. The initial searching guide map is denoted as M_0. A memory map, having the same size with M, denoted as R, is created for intermediate steps. The values of the pixels in R are initialized to be zero. A scan path is generated by sequentially choosing fixations. The order of the fixations is decided by applying WTA and IOR to M. In each round of searching a fixation, the region around the pixel with the maximum value (the "winner") in M is selected as the current fixation. Then the pixels in the current fixation region in M are copied to R. Before selecting next fixation, the IOR strategy is used to suppress pixels in M around the current fixation by setting $M = M_0 - \zeta * R$, to avoid them being selected again right away. $\zeta = 0.8$, is a constant forgetting factor. Pixels belonging to earlier fixations in M are suppressed less. After several rounds, the values of those pixels in the early fixations would be close to those of pixels in M_0. Therefore, they become candidates again. The weighting coefficients K_1, K_2, and K_3 are constrained by:

$$K_1 + K_2 + K_3 = 1 \quad (3)$$

The optimal weighting coefficients are computed by minimizing the differences between the generated scan paths and human vision fixation sequences. The computation will be explained in Section III.

III. The Performance of Our Model

We evaluate the effectiveness of the combined model by comparing the generated scan path with human vision fixation sequences in a public database collected by Ehinger et al. [20]. The database contains 912 street images, half with and half without pedestrians. The eye movements were recorded as observers searched for

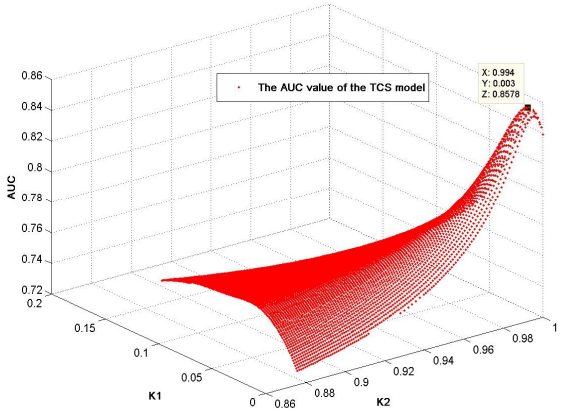


Fig. 5. The linear search of parameters in TCS model. We fix the weighting coefficients of target map and spatial context map, so the saliency map coefficient can be obtained according to (3).

TABLE I
AUC IN TEST DATABASE BY DIFFERENT MODEL

Models	Pedestrian-present (AUC)	Pedestrian-absent (AUC)
C	0.72	0.75
S	0.70	0.73
T	0.83	0.77
C+S	0.75	0.77
T+C	0.84	0.80
T+S	0.84	0.79
T+C+S	0.85	0.81

pedestrians in these images. The 912 images are divided into a training set and a testing set. The testing set consists of 200 images, half with and half without pedestrians. The rest is training set, which is used to learn the parameters in (2) by linear search. In this section, we first describe the parameter learning. Then, we evaluate our model by comparing with the other combined models.

A. Parameters Selection

We use linear search to learn the parameters in our model in (2). The linear search is done by fixing all but one free parameter in (2). The ROC curves are employed to measure the performance of the TCS model with the selected parameters. These ROC curves are drawn based on the false alarm rate and detection rate between generated fixations and recorded human eye movements. The optimal value of the free parameter is selected to maximize the area under ROC curve (AUC) value from 200 different discrete values over a predefined range (0 to 1 for parameters in (2)).

The parameter learning result is shown in Fig. 5. We denote our model which combines Target clues, Context information, and Saliency map, as TCS for short. From the figure, we can see that TCS model achieves the best performance when $K_1 = 0.003$, $K_2 = 0.994$, $K_3 = 0.003$. The optimal weights tell that the top-down target information plays a dominant role in generating scan paths.

B. Evaluation

We evaluate our TCS model by comparing it with several different models, including C (involving only context information), S (involving only saliency map), T (involving only target clues), CS (linear combination of context information and saliency map), TC (linear combination of target clues and context information), and TS model (linear combination of target clues and saliency map) in locating targets in the test dataset. The optimal parameters in CS, TC, and TS are also learned by linear search as done for TCS. We compute the ROC curves of each model to evaluate their performance. As given in Table I, the combined models CS, TC, TS, and TCS perform better than model C, S, and T which involve only one factor. TCS outperforms the other models both in the image set with and without pedestrians. In TCS model, the top-down target information is the most important factor influencing target searching paths in images with pedestrians. However, its influence in images without pedestrians is limited.

We further compare the scan paths generated by the proposed model with ground truth of human fixation sequences in the test dataset. Each scan path is composed of four fixations. Fig. 6 presents the generated scan paths of TCS model (in red), S (in purple), C (in blue), and T (in cyan), as well as human fixation sequences (in green). From this figure, we can see that the scan paths obtained by the proposed model are most similar to the real human eye movements. Quantitative comparison results are given in Table II. We evaluate the differences between the generated scan paths with those of humans using

TABLE II
HAUSDORFF DISTANCE OF SCAN PATHS BETWEEN DIFFERENT MODEL AND HUMAN

	Pedestrian-present	Pedestrian-absent
Human fixations	229.7589	205.8550
S model with two fixations	284.9030	281.5046
S model with three fixations	280.5515	267.6675
S model with four fixations	278.0492	255.8902
The average of S model	281.1679	268.3541
C model with two fixations	338.1721	299.2809
C model with three fixations	345.2276	304.5828
C model with four fixations	350.2776	309.0133
The average of C model	344.5591	304.2923
T model with two fixations	284.1035	273.4091
T model with three fixations	304.2821	269.1187
T model with four fixations	319.3544	267.9660
The average of T model	302.58	270.1916
TCS model with two fixations	273.2361	269.2430
TCS model with three fixations	276.3487	259.4316
TCS model with four fixations	285.4296	250.5578
The average of TCS model	278.3381	259.7441

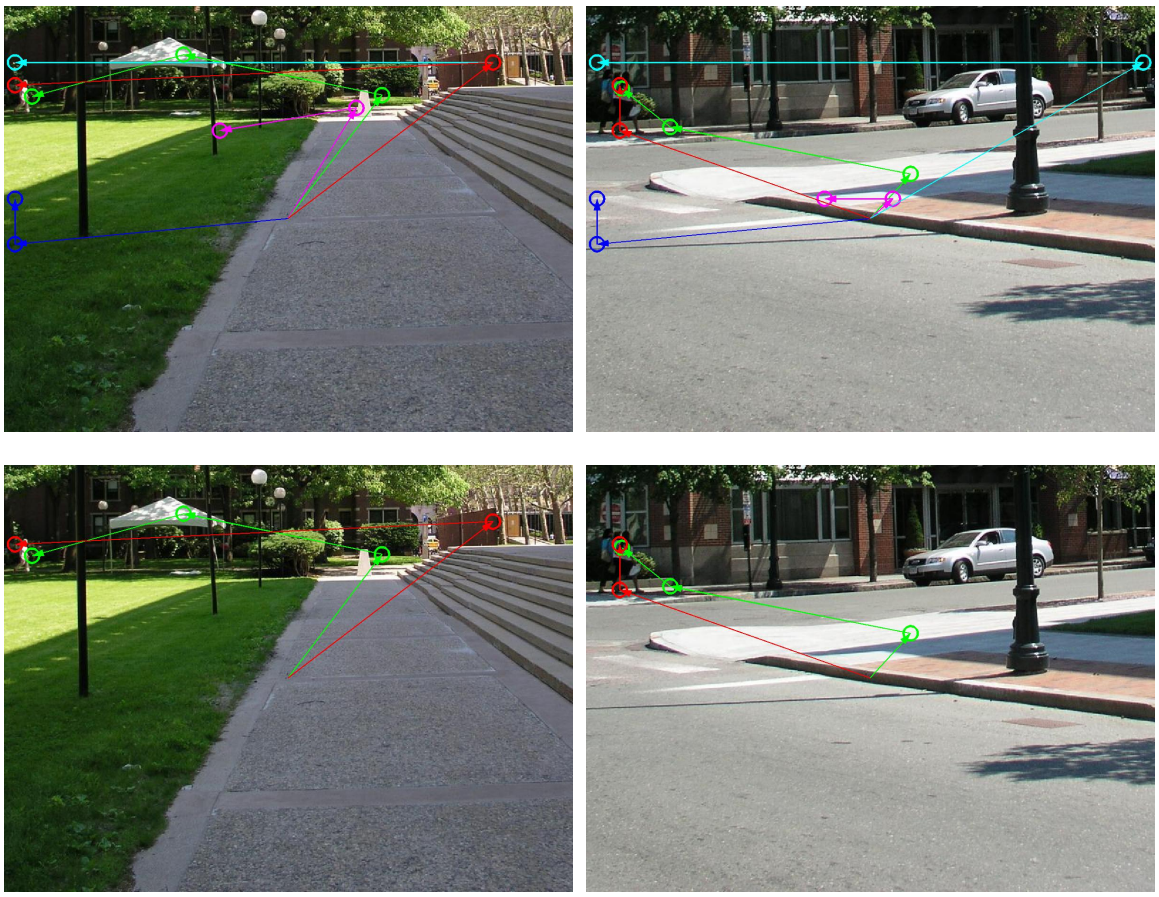


Fig. 6. Scan paths generated by T, C, S, and TCS model and human fixation sequences, marked in cyan, blue, purple, red and green, respectively. The second row just compares the TCS model with human eye movements for a striking contrast.

Hausdorff distance (H-Distance). Hausdorff distance describes the similarity of two point sets by computing the maximal value of all the minimal distances between two sets of scan paths,

$$H(A,B) = max(h(A,B), h(B,A)) \quad (4)$$

in which,

$$h(A,B) = max_{a \in B}\, min_{b \in A} \|a - b\| \quad (5)$$

$$h(B,A) = max_{b \in B}\, min_{a \in A} \|b - a\| \quad (6)$$

A denotes a generated scan path in a test image. B is the path of human vision in the image. a and b are the fixation pointes in A and B, respectively.

As shown in Table II, we compute the Hausdorff distance for the first four fixations since the observers used 3.5 fixations to reach the target averagely. Compared with the models S, C, and T, TCS model performs best in both pedestrian present and pedestrian absent datasets. Therefore, we can say that the proposed model is closest to the human vision systems than the other three.

IV. Conclusion and Discussion

In this paper, we presented a combined model to guide pedestrian searching in images. The model integrates three types of information, including top-down target information, spatial context and bottom-up visual saliency. The effectiveness of the combined model is verified by comparing its generated scan paths with human vision fixation sequences in the same images. The optimal parameters of the combined model are learned by linear search using the evaluation strategy. We also evaluated the power of the three factors in predicting targets in images, separately. Results indicated that the top-down target information performs best in images containing targets. Any integration of the factors performs better than a single component factor. The presented model, combining all the three factors with optimal parameters, has the best performance in both pedestrian present images and pedestrian absent images.

References

[1] J. R. Antes, "Time course of picture viewing," Journal of Experimental Psychology, 1974, no. 103, pp. 62-70.

[2] P. Kohli, H. Nickisch, C. Rother, C. Phemann, "User-centric learning and evaluation of interactive segmentation systems," International journal of computer vision, 2012, vol.100, pp. 261-274.

[3] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," Cognitive Psychology, 1982, no. 14, pp. 143-177.

[4] R. Carmi, and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," Vision Research, 2006, no. 46, pp. 4333-4345.

[5] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," Psychological Review, 2006, no. 113, pp. 766-786.

[6] L. Itti, and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," Vision Research, 2000, no. 40, pp. 1489-1506.

[7] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, SUN: "Top-down saliency using natural statistics," Visual Cognition, 2009, no. 17, pp. 979-1003.

[8] L. Itti, and C. Koch, "Computational modeling of visual attention," Nature Reviews Neuroscience, 2001, no. 2, pp. 194-203.

[9] C. Koch, and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," Human Neurobiology, 1985, no. 4, pp. 219-227.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," Pattern Analysis and Machine Intelligence, 1998, no. 20(11), pp.1254-1259.

[11] C. Chi, L. Qing, J. Miao, and X. Chen, "Evaluation of the Impetuses of Scan Path in Real Scene Searching," ACCV Workshops 1, 2010, vol. 6468, pp. 450-459.

[12] W. Einhäuser, U. Rutishauser, and C. Koch, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," Journal of Vision, 2008, no. 8(2), pp. 1-19.

[13] M. P. Eckstein, B. A. Drescher, and S. S. Shimozaki, "Attentional cues in real scenes, saccadic targeting and Bayesian priors," Psychological Science, 2006, no. 17, pp. 973-980.

[14] A. Oliva, and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," International Journal of Computer Vision, 2001.

[15] A. Torralba, K. P. Murphy, and R. Fergus, "Small codes and large databases of images for object recognition," Anchorage, AK, 2008.

[16] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing Visual scenes using transformed objects and parts," International Journal of Computer Vision, 2008, no. 77(1-3), pp. 291-330.

[17] G. Malcolm, and J. Henderson, "Combining top-down processes to guide eye movements during real-world scene search," Journal of Vision, 2010, 10.

[18] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'11), Colorado Springs, 2011, pp. 473-480.

[19] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Conference on Computer Vision and Pattern Recognition, 2005, no. 2, pp. 886-893.

[20] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," Visual Cognition, 2009, no. 17, pp. 945-978.

[21] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," Journal of Vision, 2008, no. 8(14):18, pp. 1–26.

[22] M. B. Neider, and G. J. Zelinsky, "Scene context guides eye movements during visual search," Vision Research, 2006, no. 46, pp. 614-621.

[23] A. Treisman, and G. Gelade, "A feature integration theory of attention," Cognitive Psychology, 1980, no. 12, pp. 97-136.