Supervised saliency maps for first-person videos based on sparse coding

Yujie Li*, Atsunori Kanemura*,†,1, Hideki Asoh*, Taiki Miyanishi†, Motoaki Kawanabe†
*National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
Email: {yujie-li, atsu-kan, h.asoh}@aist.go.jp
†Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan
Email: {atsu-kan, miyanishi, kawanabe}@atr.jp

***Abstract*—Specifying attentive regions in first-person vision (FPV) plays an important role to find meaningful objects in our daily life. Saliency detection is a major technique to locate such attentive regions. However, even though the FPV captured from the user perspective is always associated with his/her actions, existing saliency detection methods are bottom-up, and they cannot incorporate the information about the actions of the user. Since people look at the target of their actions, saliency detection algorithms for FPV should take into account which objects are more likely to be manipulated by the user. In this paper, we propose a supervised saliency detection method that uses human gaze information when the user performs actions as supervised signals. Our proposed method is based on sparse coding (dictionary learning) with a supervised saliency dictionary. Experiments using a real-world gaze dataset show that our proposed approach outperforms a state-of-the-art saliency detection algorithm based on sparse coding.**

***Index Terms*—Gaze prediction, first-person vision (FPV), egocentric vision, saliency detection, sparse modeling.**

I. Introduction

The ability to predict where people look at (i.e., gaze positions) in a scene is useful to understand their daily living and offers many applications in such as graphics design and robotic vision [1], [2]. In first-person vision (FPV), human gaze has a strong association with the user's actions since the gaze point tends to fall on the object that is going to be or currently being manipulated by the user [3]. For example, if the user is going to take a piece of bread, he/she must look at the bread to know where to reach out. This is a prominent feature of FPV that differentiate it from the third-person vision. However, most existing saliency detection algorithms are designed to be bottom-up without considering gaze/action association, not for FPV captured in daily living [4]–[10].

Fig. 1 contrasts how saliency is different between bottom-up and FPV (action-guided). The red circles denote the targets of actions whereas the blue rectangles denote objects with high bottom-up saliency. Fig. 1(a) is an FPV frame when the user is about to take a piece of bread. Bottom-up saliency tells that the orange plastic film is most distinct from other regions, whereas the user's action is to grab the bread, not to pinch the orange film. Fig. 1(b) is about the action of taking a paper bowl and shows that a bottom-up saliency map indicates the oatmeal box on the desk instead of the target paper bowl since the color of the paper bowls is the same with the background and not visually stand out. These observations pose limitations on finding "salient" regions only from video frames in FPV. Thus, traditional saliency detection methods, which have been developed without FPV in mind, are not suitable for detecting saliency in FPV, which shares the user's eyesight and is associated with the user's actions.

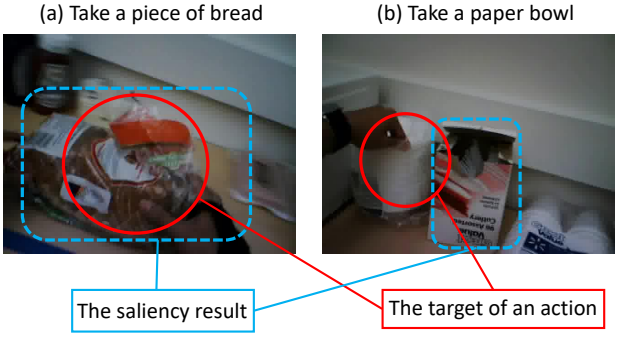


Fig. 1. Illustration of saliency results and target action.

For detecting saliency in images and videos, many saliency models have been proposed. Li et al. [11] have proposed a visual saliency detection algorithm, dense and sparse reconstruction (DSR), from the perspective of reconstruction errors. The image boundaries are first extracted via superpixels as likely cues for background templates, from which dense and sparse appearance models are constructed. Li et al. [12] have built a dictionary-based framework that constructs saliency and non-saliency dictionaries from stacked feature vectors and detects saliency with a weighted sparse coding framework, which is called the weighted sparse coding framework (WSCF). Their article reports that WSCF performs favorably against then-state-of-the-art methods in terms of precision and recall. Li et al. [13] have extended sparse coding based method by introducing l_1-norm as sparsity constraint. Moreover, saliency detection methods based on deep neural networks have been proposed recently [14]–[16].

In this paper, we propose a saliency detection method for predicting human gaze in FPV. We employ sparse coding

This study was supported in part by the New Energy and Industrial Technology Development Organization (NEDO), Japan, JST CREST JPMJCR15E2, and JSPS KAKENHI 18K18083.

[1]Now with LeapMind Inc., Tokyo, Japan (email: atsu-kan@leapmind.io).

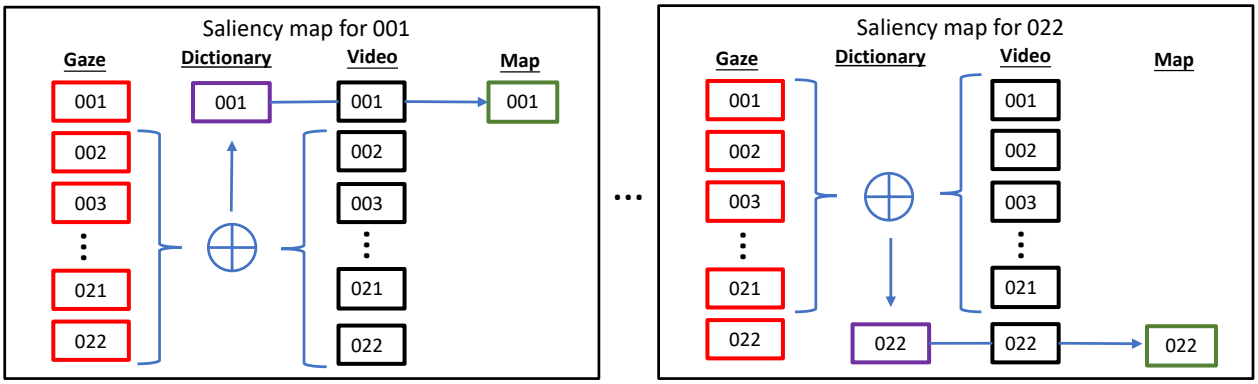


Fig. 2. Framework of supervised saliency mapping.

(dictionary learning) and build a supervised dictionary that incorporates human gaze during actions. Therefore, our proposed method is not bottom-up and we call it supervised gaze prediction based on sparse coding (SGP). We use the weighted l_1-norm as our sparsity measure to attain robustness. For evaluation, we use an FPV and gaze dataset collected with eye tracking glasses in real-world environments, where the subjects perform many actions. Experimental results show that our proposed method improves the gaze prediction performance in the FPV compared to the existing state-of-the-art sparse coding based saliency method.

II. Framework and Formulation

We use supervised saliency maps to predict the gaze point of FPVs. We use a dataset consisting of gaze and video. The GTEA Gaze dataset [17] contains both gaze and FPV video for 17 people. The gaze and video data are labeled G001 and V001, respectively, for person 001.

The framework of supervised saliency mapping is illustrated in Fig. 2. When predicting saliency maps for person 001, we use the gaze and video information from the other people to learn the supervised saliency dictionary, which we denote D001. Then, the learned dictionary D001 is employed to predict saliency maps for video V001.

As a basis of our formulation, we use sparse coding with a weighted l_1-norm [12], which consists of two separate stages of saliency dictionary updates and gaze saliency mapping.

A. Features for Video

We divide a video frame into R superpixels and extract a feature vector for each superpixel. We use coupled RGB and Lab color spaces as color descriptors that can improve the accuracy of saliency maps [18]. Two feature matrices are generated for all superpixels: An averaged feature matrix $\mathbf{F}_{\mathrm{a}} = \mathbb{R}^{C \times R}$ and a color histogram feature matrix $\mathbf{F}_{\mathrm{h}} = \mathbb{R}^{C' \times R}$, where R is the number of superpixels, C is the averaged feature dimensionality, and C' is the color histogram feature dimensionality. By concatenating them, the video feature matrix $\mathbf{F} = [\mathbf{F}_{\mathrm{a}}, \mathbf{F}_{\mathrm{h}}]$ is generated. Note that the rth column of $\mathbf{F}$ is a feature vector for superpixel r: $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_r, \ldots, \mathbf{f}_R]$.

The averaged feature $\mathbf{F}_{\mathrm{a}}$ performs well when the scene is composed of objects with simple colors and textures but is less robust when the foreground and background contain complex textures. This is because averaging over all pixels loses information that characterizes color variations within each superpixel. The color histogram $\mathbf{F}_{\mathrm{h}}$ is suitable for handling scenarios where the scene contains highly textured objects. Thus, these two features complement each other.

B. Sparse Modeling for Gaze Prediction

Our proposed sparse coding based gaze prediction framework calculates saliency from the feature matrix by monitoring the reconstruction errors from a saliency dictionary. We stand on existing studies that show non-saliency regions can be represented by a sparsely coded dictionary [11], [12]. We use the error measure to refine the foreground superpixels and to identify foreground saliency ones.

Saliency detection based on sparse coding [11] identifies salient regions as those having high reconstruction errors with a background templates dictionary. The dictionary $\mathbf{D} \in \mathbb{R}^{\tilde{C} \times K}$ comprises K bases (or atoms) representing feature vectors for background superpixels. The sparse reconstruction error for superpixel $r \in \{1, \ldots, R\}$ is defined to be

$$\epsilon_r^* = \|\mathbf{f}_r - \mathbf{D}\mathbf{h}_r^*\|_2^2, \tag{1}$$

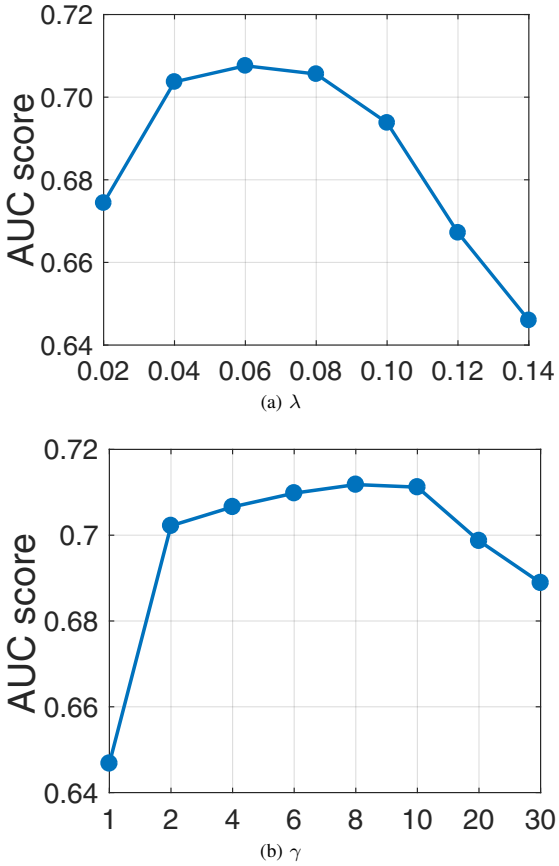
where sparse coefficients $\mathbf{h}_r^* \in \mathbb{R}^K$ are found by

$$\mathbf{h}_r^* = \underset{\mathbf{h}}{\operatorname{argmin}} \|\mathbf{f}_r - \mathbf{D}\mathbf{h}\|_2^2 + \lambda \|\mathbf{h}\|_1. \tag{2}$$

Here, $\lambda > 0$ is a regularization parameter, which determines a tradeoff between the approximation error and the sparsity constraint. Thanks to the sparsity induced from the l_1-norm, the sparse reconstruction errors are robust to complicated background [11].

For saliency detection, we adopt a diagonal matrix $\mathbf{W}_r$ that contains weights for each atom for superpixel r [19] and modify the sparse coding scheme to be

$$\mathbf{h}_r^* = \underset{\mathbf{h}}{\operatorname{argmin}} \|\mathbf{f}_r - \mathbf{D}\mathbf{h}_r\|_2^2 + \lambda \|\mathbf{W}_r \mathbf{h}_r\|_1. \tag{3}$$



(a) λ (b) γ

Fig. 3. AUC scores of SGP with varying (a) λ and (b) γ.

The diagonal elements of $\mathbf{W}_r$ compute the similarities between superpixel $\mathbf{f}_r$ to all the atoms in dictionary $\mathbf{D}$:

$$\mathbf{W}_r = \text{diag}(\exp(\|\mathbf{f}_r - \mathbf{d}_1\|_2^2), \ldots, \exp(\|\mathbf{f}_r - \mathbf{d}_K\|_2^2)). \quad (4)$$

The weight matrix for saliency detection is designed to be inversely proportional to the similarity between the feature vector $\mathbf{f}_r$ and the dictionary $\mathbf{D}$. Namely, if the $\mathbf{f}_r$ is similar to some template in D, the weight should be small and vice versa.

C. Supervised Saliency Dictionary

To predict gaze in FPV videos captured from a user performing actions, our approach uses a supervised saliency dictionary $\mathbf{D}$ built from gaze positions and video frames of the other users. The dictionary $\mathbf{D}$ is constructed by gaze frames, starting from an initial dictionary and repeatedly refining it [12]. The initial dictionary is generated from the other videos (except the one needs to predict). To construct the initial dictionary, the feature vectors of superpixels that contain the other user's gaze is used as an atom. In the refinement stage, the dictionary is updated to be a set of the feature vectors whose Sal values are higher than the mean value of Sal, there Sal is defined in Section II-E.

D. Object-Biased Gaussian Model for Center Prior

It is known that human gaze has the center bias—humans look at peripheral areas less frequently than central areas. To incorporate the center bias into our model, we use an object-biased Gaussian model [11], [20], [21] to determine a center prior. For each superpixel r with its coordinates (x, y), we define our center prior to be

$$Sal^+(r) = \exp\left[-\left(\frac{(x_r - x_{\text{obj}})^2}{2\sigma_{\text{x}}^2} + \frac{(y_r - y_{\text{obj}})^2}{2\sigma_{\text{y}}^2}\right)\right], \quad (5)$$

where σ_{x} and σ_{y} are 25% of the height and width of an image, respectively, and x_{obj} and y_{obj} are the object center derived from the pixel error as follows.

$$x_{\text{obj}} = \sum_{r=1}^{R} \omega_r x_r, \quad y_{\text{obj}} = \sum_{r=1}^{R} \omega_r y_r, \quad (6)$$

where ω_r are weights defined by normalizing the reconstruction errors ϵ_r^* in (1) as

$$\omega_r = \frac{\epsilon_r^*}{\sum_{r'=1}^{R} \epsilon_{r'}^*}. \quad (7)$$

E. Supervised Saliency Prediction

We compute the saliency value $Sal(r)$ for superpixel r as follows [12].

$$Sal(r) = Sal^+(r) \cdot Sal^*(r), \quad (8)$$

where $Sal^+(r)$ is the object-bias center prior defined in (5) and $Sal^*(r) = \exp(-\gamma\epsilon_r^*)$ depends on the reconstruction error in (1). The parameter γ determines a tradeoff between the object-bias and the spare reconstruction error.

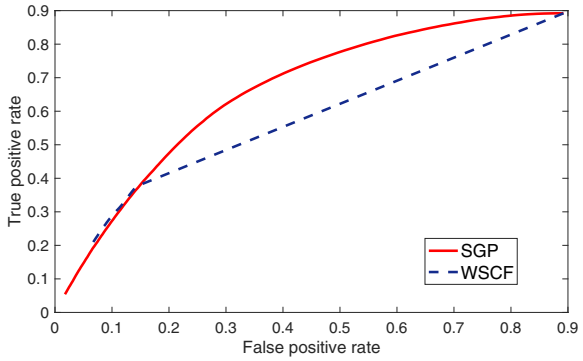
Our proposed algorithm, supervised gaze prediction based on sparse coding (SGP), is shown in Algorithm 1. There are two main stages of this algorithm: a) Supervised dictionary learning stage, which uses the gaze information to learn a supervised dictionary; and b) Gaze saliency mapping stage, which obtains the saliency map of gaze points using the learned dictionary.

III. Experiments

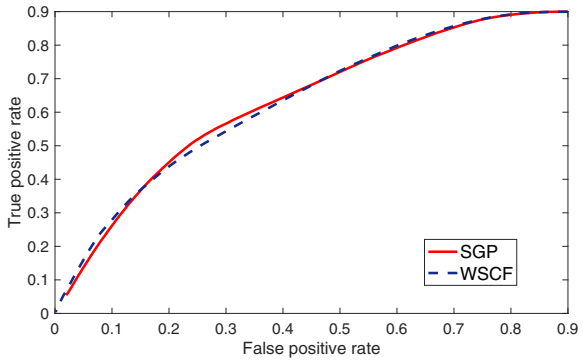
We used the GTEA Gaze dataset [17][1], which has recorded FPV videos together with gaze points obtained from eye-tracking glasses. The original motivation of collecting this dataset is to understand the relationship between human activities and gaze. There are 17 FPVs in the dataset, each captures one person performing many actions. For example, video 001 is a FPV of a user cooking sandwiches and contains 30 sessions, each of which is associated with an action such as "take bread" or "take knife."

We compared the results by our proposed algorithm, SGP, with the state-of-the-art saliency detection method based on sparse modeling [12]. We used the receiver operating characteristic (ROC) curve and the area under curve (AUC) to measure the consistency between predicted gaze maps and the ground truth gaze points, which are widely used for evaluation in the saliency detection literature [22].

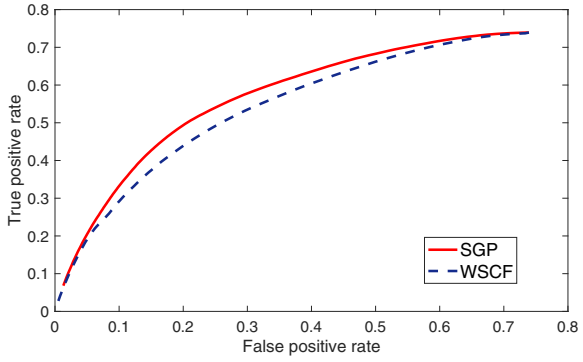
[1] http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/GTEA_Gaze.html.



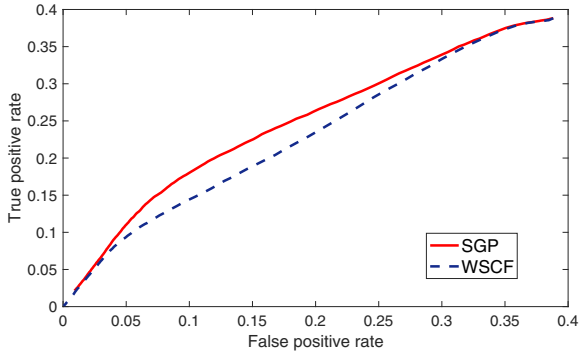
(a) video 001



(b) video 002



(c) video 003



(d) video 005

Fig. 4. ROC curves for video 001 to 005.

Algorithm 1 Supervised gaze prediction based on sparse coding (SGP)

Input: Video and gaze data from the target user $\mathcal{V}_\mathrm{t}$ and $\mathcal{G}_\mathrm{t}$ and from the other supervising users $\mathcal{V}_\mathrm{s}$ and $\mathcal{G}_\mathrm{s}$.

 # Supervised dictionary construction:
Compute the averaged feature matrix $\mathbf{F}_\mathrm{a}$ and the color histogram feature matrix $\mathbf{F}_\mathrm{h}$ for supervisor videos $\mathcal{V}_\mathrm{s}$.
Built an initial saliency dictionary $\mathbf{D}$ from the video features and gaze positions $\mathcal{G}_\mathrm{s}$.
 # Gaze saliency mapping:
for each frame in video $\mathcal{V}_\mathrm{t}$ **do**
    Compute the averaged feature matrix $\mathbf{F}_\mathrm{a}$ and the color histogram feature matrix $\mathbf{F}_\mathrm{h}$.
    Calculate the object-biased Gaussian model based center prior by (5).
    Update the saliency dictionary $\mathbf{D}$ by selecting feature vectors whose saliency values were larger than the average: $\mathbf{D} \leftarrow \{\mathbf{f}_r \mid Sal(r) > \mathrm{mean}(Sal(r))\}$.
    Obtain the saliency values by (8).
end for

TABLE I
AUC SCORE COMPARISON.

No.	WSCF	SGP	Dev.(%)
001	0.499	**0.583**	16.83
002	**0.571**	0.565	−1.05
003	0.397	**0.414**	4.28
005	0.085	**0.095**	11.76
006	0.420	**0.469**	11.67
007	0.630	**0.653**	3.65
008	0.325	**0.402**	23.69
010	0.253	**0.259**	2.37
012	0.439	**0.440**	0.23
013	0.233	**0.239**	2.58
014	**0.691**	0.667	−3.47
016	0.569	**0.586**	2.99
017	0.572	**0.583**	1.92
018	0.579	**0.581**	0.35
020	0.297	**0.354**	19.19
021	0.603	**0.683**	13.27
022	0.439	**0.508**	15.72

A. Degree of Trade-off Parameters

We conducted an experiment to evaluate the effects of changing the sparsity parameter λ and the trade-off parameter γ using session 1 in video 001. Fig. 3 shows the AUC scores for different values of λ and γ. From Fig. 3(a), we can observe that l_1-norm sparsity controls the tradeoff between precision and recall well and the AUC values are robust when λ is between 0.04 and 0.08. Therefore we used $\lambda = 0.06$ for all

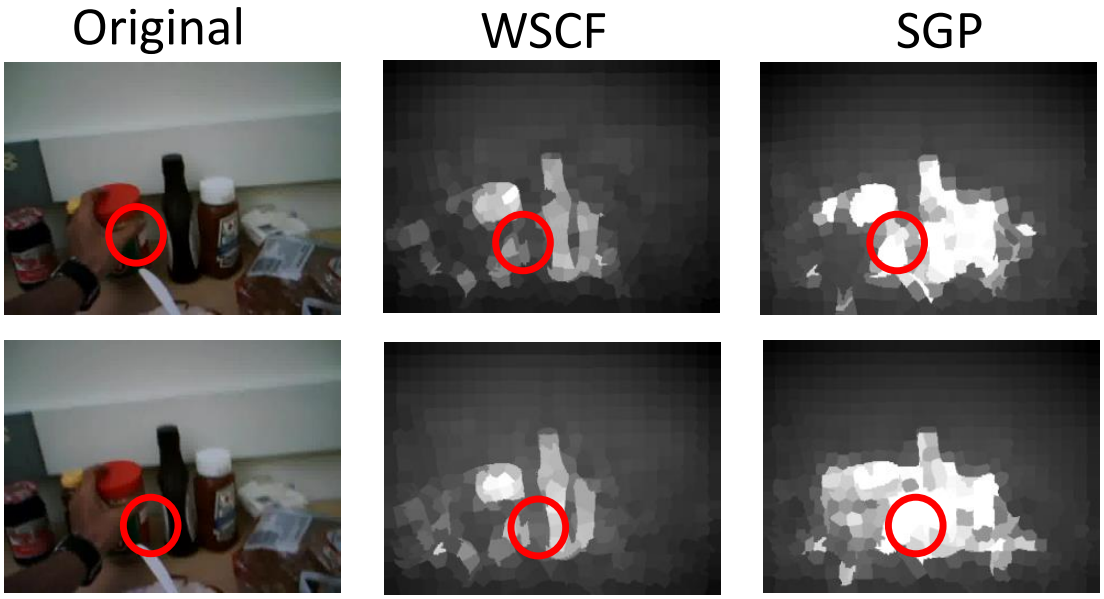


Fig. 5. Gaze prediction by WSCF and SGP with the original frame. The red circles are the gaze position of the user.

the experiments. Fig. 3(b) shows the AUC scores for different values of γ, from which we can observe that the AUC values are robust when γ is between 2 and 10. Therefore we used $\gamma = 8$ for all the experiments.

B. Detection Performance

Fig. 5 shows saliency detection results for two frames in video 001. Although WSCF did not detect saliency at the true gaze position marked by the red circle, the proposed algorithm, SGP, assigns high saliency to there. The user's gaze was on seemingly unimportant positions, which were in fact not salient based on the classical definition; but our FPV saliency did not miss that gaze.

IV. Conclusion

We proposed a novel supervised gaze prediction method based on sparse coding, which compared favorably to the existing sparse coding based method. The novel technical element was the introduction of gaze positions for building the supervised dictionary to take into account gaze/action association in FPV.

Our proposed idea can be extended in several ways. Since the human gaze behaves differently for performing different actions, it will be important to discover when the supervised framework is most useful. If we specify in more detail the type of actions, for example cooking actions, we may incorporate domain knowledge like food detection with convolutional neural networks [23].

References

[1] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 329–341, 2013.

[2] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, 2010.

[3] M. F. Land, "The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations," *Exp. Brain Res.*, vol. 159, no. 2, pp. 151–160, 2004.

[4] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2007, pp. 1–6.

[5] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Res.*, vol. 40, no. 10, pp. 1489–1506, 2000.

[6] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 155–162.

[7] Z. Li, "A saliency map in primary visual cortex," *Trends Cogn. Sci.*, vol. 6, no. 1, pp. 9–16, 2002.

[8] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 241–248.

[9] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends Cogn. Sci.*, vol. 7, no. 11, pp. 498–504, 2003.

[10] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231, 2009.

[11] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 2976–2983.

[12] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 5216–5223.

[13] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe, "A sparse coding framework for gaze prediction in egocentric video," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 1313–1317.

[14] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Int. Conf. Pattern Recognit. (ICPR)*, 2016, pp. 3488–3493.

[15] Z. Wang, P. Jiang, and F. Wang, "Dense residual pyramid networks for salient object detection," in *Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 606–621.
[16] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4446–4456, 2017.
[17] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 314–327.
[18] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 478–485.
[19] J. Huang, S. Ma, and C.H. Zhang, "Adaptive lasso for sparse high-dimensional regression models," *Stat. Sin.*, pp. 1603–1618, 2008.
[20] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Scand. Conf. Image Anal. (SCIA)*, 2011, pp. 666–675.
[21] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.
[22] X. Li, Y. Li, C. Shen, A. Dick, and A. van den Hengel, "Contextual hypergraph modeling for salient object detection," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 3328–3335.
[23] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *ACM Int. Conf. Multimed. (ACM MM)*, 2014, pp. 1085–1088.