A SPARSE CODING FRAMEWORK FOR GAZE PREDICTION IN EGOCENTRIC VIDEO

Yujie Li[1], Atsunori Kanemura[1,2], Hideki Asoh[1], Taiki Miyanishi[2], Motoaki Kawanabe[2]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
{yujie-li, atsu-kan, h.asoh}@aist.go.jp
[2]Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan
{atsu-kan, miyanishi, kawanabe}@atr.jp

ABSTRACT

To efficiently process and understand a large amount of incoming visual information from first-person perspective (i.e. egocentric vision), predicting human gaze is important. However, even though people continuously gaze in noisy environments, most existing gaze prediction methods mainly use image saliency, which is sensitive to noise in the real-world. To address this issue, we propose a sparse coding-based saliency detection method for gaze prediction. Our model uses a cost function with the l_0 norm as a sparse constraint that can control the area of visual saliency in response to the contents of egocentric vision in intuitive and consistent ways. Moreover, we use canonical correlation analysis (CCA) to combine different types of features for reducing noise and the computational complexity. We also utilize the temporal continuity of image frames when defining our saliency. Experiments using a real-world gaze dataset show that our proposed approach outperforms the state-of-the-art algorithms on gaze prediction in egocentric videos.

Index Terms— Gaze prediction, egocentric video, saliency detection, sparse modeling, canonical correlation analysis (CCA).

1. INTRODUCTION

Human gaze is an important function for a person to sense the visual world through a series of fixations [1]. Predicting such human gaze is useful for many applications that need to extract the most significant regions in an image or video frame, such as surveillance and robotic vision. In this paper, we tackle the problem of egocentric gaze prediction, which is the task of predicting the user's perspective gaze given a video from the first-person perspective (i.e. egocentric vision). The egocentric vision is different from the third-person vision in that it is noisier and more unstable because of the head shakes of the user, and thus special considerations are required to design algorithms for it.

This study was supported in part by the New Energy and Industrial Technology Development Organization (NEDO), Japan, JST CREST JPMJCR15E2, JST SICORP, and JSPS KAKENHI 26730130.

To predict human gaze, many works have proposed visual saliency detection algorithms; however, these are designed for still and third-person images and thus sensitive to noise and instability in egocentric videos. Itti et al. [2] have proposed a saliency model based on the feature integration theory that extracts early visual features (e.g., colors, orientations, and edges) and fuses them into a saliency map. Harel et al. [3] have proposed a bottom-up visual saliency model called graph-based visual saliency (GBVS), which forms activation maps on certain feature channels, normalizes them in a way that highlights outstanding areas, and combines with other maps. Li et al. [4] 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). However, even though a person continuously gazes in the noisy environment, most existing gaze prediction methods use image saliency, which is sensitive to noise in the real world. Moreover, Li et al. [4] uses the sparsity with l_1 norm for building saliency dictionaries, which is not flexible to estimate the area of visual saliency and thus the regularization parameter needs subject-wise tuning.

In this paper, we propose a sparse coding-based gaze prediction method addressing these issues based on the following techniques. First, we use averaged features of neighborhood frames for smoothing gaze movements as dynamic egocentric video features. Second, we use canonical correlation analysis (CCA) to combine different type of features (static and dynamic gaze features) for reducing noise in egocentric vision and the computational complexity. Third, we use a sparse coding-based gaze prediction framework with the l_0 norm as sparsity, which can flexibly control the area of visual saliency in response to the content of egocentric vision.

For evaluation, we use a gaze dataset collected with eye tracking glasses in real-world environments. The experimental results show that our proposed method improves the gaze prediction performance in the egocentric vision compared to the existing saliency-based methods.

2. FRAMEWORK AND FORMULATION

To predict gaze in egocentric videos captured from real-world noisy environments, our approach uses saliency dictionaries built for superpixels. We build saliency dictionaries though sparse modeling with the color histogram of the frame image and the averaged saliency map of two neighbors frames in videos.

2.1. Feature Extraction by CCA

Color is the most intuitive feature to distinguish different regions. We use coupled RGB and Lab color spaces as color descriptors that can improve the accuracy of saliency maps [5]. We generate two feature matrices for all superpixels: An averaged feature matrix $\mathbf{F}_{\mathrm{a}} = \mathbb{R}^{C\times N}$ and a color histogram feature matrix $\mathbf{F}_{\mathrm{h}} = \mathbb{R}^{C'\times N}$, where N is the number of superpixels, C is the averaged feature dimensionality, and C' is the color histogram feature dimensionality.

The averaged feature $\mathbf{F}_{\mathrm{a}}$ performs well when the scene is composed of objects with simple color and textures but is less robust when the foreground and background contain highly 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.

Moreover, to combine these different types of features, we extract common features between the feature matrices $\mathbf{F}_{\mathrm{a}}$ and $\mathbf{F}_{\mathrm{h}}$ by canonical correlation analysis (CCA), which finds linear projections of matrices $\mathbf{F}_{\mathrm{a}}$ and $\mathbf{F}_{\mathrm{h}}$ maximizing the correlation with each other (Fig. 1). Features generated by CCA can reduce the computational complexity since the dimensionality of $\mathbf{F}_{\mathrm{a}}$ and $\mathbf{F}_{\mathrm{h}}$ becomes $\widetilde{C} = \min(\mathrm{rank}(\mathbf{F}_{\mathrm{a}}), \mathrm{rank}(\mathbf{F}_{\mathrm{h}}))$, which is less than C and C'. CCA is formulated as the following minimization problem.

$$\min_{\mathbf{P}_{\mathrm{a}},\mathbf{P}_{\mathrm{h}}} \|\mathbf{P}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}} - \mathbf{P}_{\mathrm{h}}\mathbf{F}_{\mathrm{h}}\|_{\mathrm{F}}^2, \tag{1}$$

where $\mathbf{P}_{\mathrm{a}} \in \mathbb{R}^{\widetilde{C}\times C}$ and $\mathbf{P}_{\mathrm{h}} \in \mathbb{R}^{\widetilde{C}\times C'}$ are linear projectors from the averaged feature domain and the color histogram domain, respectively, to the common space with the same dimension. The detailed derivation can be found in [6]. We estimate optimal projection matrices $\mathbf{P}_{\mathrm{a}}$ and $\mathbf{P}_{\mathrm{h}}$ by solving an eigenvalue problem. Once the projection matrices $\mathbf{P}_{\mathrm{a}}$ and $\mathbf{P}_{\mathrm{h}}$ have been learned, we can utilize these matrices for projecting data vectors from averaging domain and color histogram domain into the $\widetilde{C}$-dimensional common space where the relevant pairs of information get close [7, 8].

The common space averaged features $\mathbf{F}_{\mathrm{a}}^{\mathrm{C}} = \mathbf{P}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}}$ and the common space color histogram features $\mathbf{F}_{\mathrm{h}}^{\mathrm{C}} = \mathbf{P}_{\mathrm{h}}\mathbf{F}_{\mathrm{h}}$ are concatenated to be the video feature matrix $\mathbf{F} = [\mathbf{F}_{\mathrm{a}}^{\mathrm{C}}, \mathbf{F}_{\mathrm{h}}^{\mathrm{C}}]$.

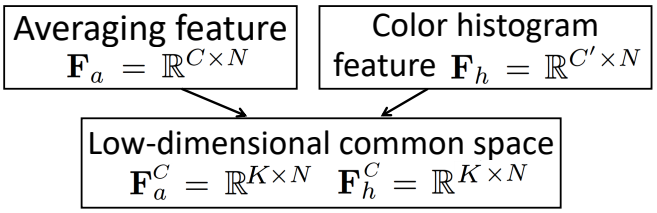

Fig. 1. Graphical model for canonical correlation analysis.

2.2. 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 [4, 9]. We use the error measure to refine the foreground superpixels and to identify foreground saliency ones.

Saliency detection based on sparse coding [9] identifies salient regions as those having high reconstruction errors with background templates dictionary. The dictionary $\mathbf{D} \in \mathbb{R}^{\widetilde{C}\times K}$ comprises K bases (or atoms) representing feature vectors for background superpixels. The sparse reconstruction error for superpixel $r \in \{1, \ldots, N\}$ is defined to be

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

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

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

Here, $\lambda > 0$ is a regularization parameter. Thanks to the sparsity induced from the l_1 norm, the sparse reconstruction errors are robust to complicated background [9].

We propose the l_0 norm to be a better sparsity measure for detecting saliency, and define our sparse reconstruction error to be

$$\epsilon_r^\star = \|\mathbf{f}_r - \mathbf{D}\mathbf{h}_r^\star\|_2^2, \tag{4}$$

$$\mathbf{h}_r^\star = \underset{\mathbf{h}}{\mathrm{argmin}} \|\mathbf{f}_r - \mathbf{D}\mathbf{h}\|_2^2 \quad \text{s.t. } \|\mathbf{h}\|_0 \le l. \tag{5}$$

The l_1 norm optimization problem [10] has been developed as a relaxation of the original l_0 problem, and it is known that l_1 solutions are not as sparse as l_0 solutions thus they may not induce adequate sparsity when applied to certain applications [11, 12]. Although the l_0 norm optimization problem is generally NP-hard [13], there are practical methods to obtain approximate solutions such as matching pursuit (MP) [14] or orthogonal matching pursuit (OMP) [15]. In this paper, we employ OMP to solve the sparse coding problem.

The sparsity parameter l in (5) has an intuitive interpretation as the number of atoms and can be adjusted to measure the area of gaze prediction. Once we fix the l parameter, we will have consistent results for different videos. In contrast, the l_1 norm does not count the number of atoms and the l_1 solutions are subject to atom count variability over different videos even with a fixed parameter λ.

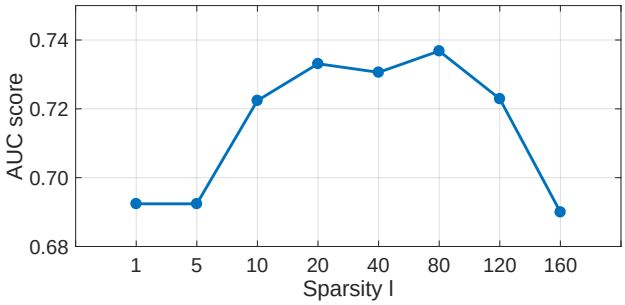

Fig. 2. The AUC scores of GPSC with different values of the l_0 sparsity.

Then we compute the saliency value $Sal(r)$ for superpixel r from the reconstruction error as follows.

$$Sal(r) = Sal^{+}(r) \cdot Sal^{*}(\epsilon_r^{\star}), \tag{6}$$

where $Sal^{+}(r)$ is the object-bias center prior defined in [16] and $Sal^{*}(\epsilon_r^{\star}) = \exp(-\epsilon_r^{\star})$ depends on the dictionary type [4]. We average the saliency values from two temporally adjacent frames to obtain final gaze predictions.

The saliency dictionary $\mathbf{D}$ is constructed by starting from an initial dictionary and repeatedly refining it [4]. The initial dictionary is a set of feature vectors from superpixels in non-salient regions. Non-salient regions are defined to be those whose boundary connectivity scores are non-zero. The connectivity score for superpixel r measures how r is connected to neighbor superpixels sharing the boundary with r; the connectivity score is high if the neighbors belong to other regions (i.e. many boundaries) and low if the neighbors are from the same region (i.e. no boundaries) [17]. In the refinement state, the dictionary is updated to be a set of feature vectors whose Sal values are higher than the mean value of Sal.

2.3. Gaze Prediction Algorithm

We present our proposed algorithm: Gaze prediction based on sparse coding (GPSC) in Algorithm 1 below.

Algorithm 1 Gaze prediction based on sparse coding (GPSC)

1: Compute the averaged feature matrix $\mathbf{F}_{\mathrm{a}}$ and the color histogram feature matrix $\mathbf{F}_{\mathrm{h}}$ for frame i.
2: Obtain $\mathbf{F}_{\mathrm{a}}^{\mathrm{c}}$ and $\mathbf{F}_{\mathrm{h}}^{\mathrm{c}}$ by projecting to the common CCA space.
3: Built an initial saliency dictionary $\mathbf{D}$.
4: **repeat**
5: Calculate the saliency values by (6).
6: Update the saliency dictionary $\mathbf{D}$ by selecting feature vectors whose saliency values were larger than the average: $\mathbf{D} \leftarrow \{\mathbf{f}_r \mid Sal(r) > \mathrm{mean}(Sal(r))\}$.
7: **until** convergence
8: From Sal for all the frames, obtain gaze prediction by averaging the saliency values from two adjacent frames.

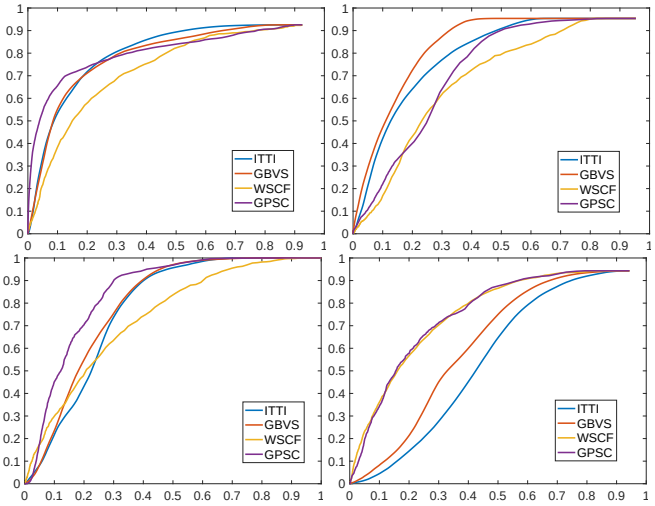

Fig. 3. ROC curves for sessions 1–4 in video 001.

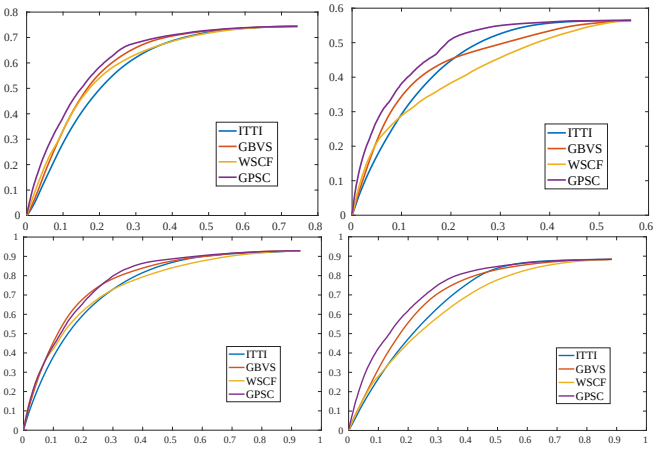

Fig. 4. ROC curves for videos 012, 013, 014, and 016.

3. EXPERIMENTS

We use the GTEA Gaze dataset [18], which has recorded the egocentric video together with gaze points obtained from eye-tracking glasses, which are used as the ground truth for gaze prediction. There are 17 egocentric videos in the dataset. Video 001 captures a person cooking sandwiches and contains 30 sessions, each of which is associated with an action such as "take bread" or "take knife."

We compare the results by our proposed algorithm GPSC with three competing methods: two traditional image saliency-based methods, ITTI [2] and GBVS [3], and a l_1 sparse modeling method WSCF [4]. We use the receiver operating characteristic (ROC) curve and the area under curve (AUC) to measure the consistency between a predicted gaze map and the ground truth gaze points, which are widely used in the saliency detection literature [19].

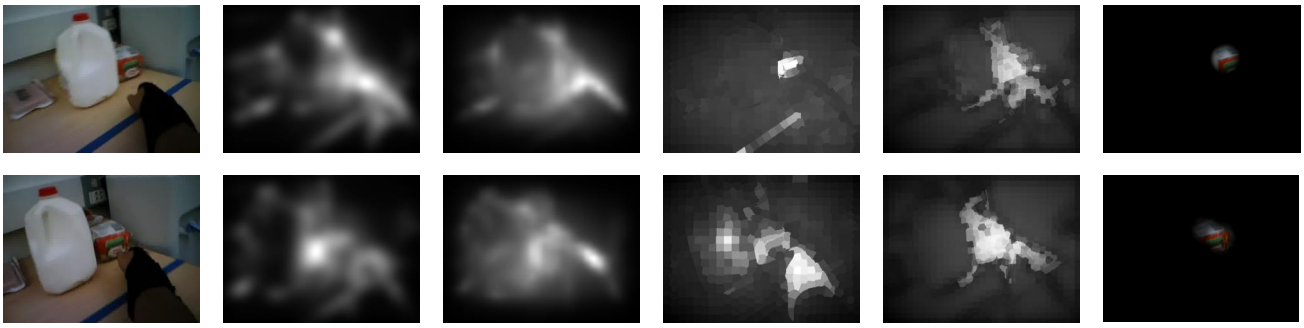

Fig. 5. Gaze prediction by different methods, from left to right: Original frame, ITTI, GBVS, WSCF, GPSC, Ground Truth.

Table 1. The AUC scores by different methods in diffident videos.

No.	002	003	005	006	007	008	010	012	013	014	016	017	018	020	021	022	Ave.
ITTI	0.606	0.418	0.099	0.466	0.664	0.425	0.276	0.426	0.250	0.690	0.592	0.555	0.581	0.350	0.680	0.513	0.474
GBVS	0.598	0.428	0.086	**0.549**	0.697	0.399	0.276	0.444	0.251	0.717	0.614	0.583	**0.604**	0.371	**0.715**	**0.545**	0.492
WSCF	0.571	0.397	0.085	0.420	0.630	0.325	0.253	0.439	0.233	0.691	0.569	0.572	0.579	0.297	0.603	0.439	0.444
GPSC	**0.641**	**0.432**	**0.113**	0.503	**0.714**	**0.444**	**0.297**	**0.458**	**0.271**	**0.719**	**0.644**	**0.596**	0.603	**0.379**	0.695	0.538	**0.503**

Table 2. The AUC scores by different methods in video 001.

No.	Session name	ITTI	GBVS	WSCF	GPSC
1	take bread	0.733	0.722	0.662	**0.737**
2	take PlateBowl	0.747	**0.789**	0.644	0.682
3	take knife	0.773	0.791	0.737	**0.845**
4	take bread	0.503	0.568	0.701	**0.701**
5	take peanut	0.842	0.810	0.772	**0.859**
6	open peanut	0.630	0.678	**0.830**	0.781
7	scoop peanut	0.554	0.590	0.652	**0.662**
8	spread peanut	0.579	**0.702**	0.590	0.612
9	scoop peanut	0.714	0.769	0.683	**0.795**
10	spread peanut	0.568	**0.632**	0.568	0.552
11	close peanut	0.520	0.501	**0.640**	0.631
12	put peanut	0.621	**0.729**	0.686	0.715
13	take jam	0.775	0.718	0.684	**0.792**
14	open jam	0.866	0.855	0.771	**0.894**
15	spread peanut	0.661	0.752	0.594	**0.778**
16	scoop jam	0.891	**0.906**	0.765	0.898
17	scoop jam	0.882	0.893	0.766	**0.906**
18	close jam	0.870	**0.896**	0.765	0.884
19	put jam	**0.735**	0.666	0.550	0.560
20	spread jam	0.857	0.869	0.758	**0.872**
21	sandwich bread	0.622	0.546	0.551	**0.661**
22	take PlateBowl	0.808	**0.833**	0.745	0.773
23	take cereal	0.736	0.726	0.744	**0.865**
24	pour cereal	0.790	0.867	0.800	**0.871**
25	put cereal	0.671	0.698	**0.714**	0.674
26	take milk	0.593	**0.617**	0.513	0.593
27	open milk	0.855	**0.864**	0.805	0.833
28	pour milk	0.852	**0.879**	0.744	0.873
29	close milk	0.725	**0.761**	0.690	0.715
30	put milk	0.727	0.746	0.681	**0.749**
mean		0.723	0.746	0.693	**0.759**

3.1. Degree of sparsity

We conduct an experiment to evaluate the effects of changing the l sparsity parameter using session 1 in video 001.

Fig. 2 shows the AUC scores for different values of l. From Fig. 2, we can observe that l_0 sparsity controls the trade-off between precision and recall well and the AUC values are robust when l is between 20 and 80. Therefore we use $l = 50$ for all the experiments.

3.2. Detection Performance

Fig. 5 shows results for frame 1 (top) and frame 2 (bottom) in session 1 of video 001. The proposed algorithm GPSC achieves more accurate gaze prediction than WSCF and the gaze area by GPSC is more compact than traditional methods, ITTI and GBVS.

Our proposed algorithm quantitatively outperformed the other three methods. Fig. 3 shows the ROC curves of sessions 1–4 in video 001 and Fig. 4 shows the ROC curves of videos 012–016. The curves of GPSC are mostly placed to top-left to the other methods. The AUC scores for the 30 sessions from video 001 is shown in Table 2. Table 1 shows the AUC scores averaged over all the sessions for each video. GPSC has the highest performance and it has improved the sparse coding based baseline WSCF.

4. CONCLUSION

We have proposed a novel gaze prediction method based on sparse coding, which compared favorably to three existing methods. Novel technical elements include CCA projection to a common space, the use of the l_0 norm as a sparsity measure, and the consideration of temporal continuity. We expect further improvement by using multi-modal information integrating audio and motion to pure video information [7, 20].

5. REFERENCES

[1] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3216–3223.

[2] 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.

[3] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[4] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5216–5223.

[5] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 478–485.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[7] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe, "Extracting key frames from first-person videos in the common space of multiple sensors," in *IEEE International Conference on Image Processing (ICIP)*, 2017.

[8] T. Miyanishi, J. Hirayama, Q. Kong, T. Maekawa, H. Moriya, and T. Suyama, "Egocentric video search via physical interactions," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[9] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 2976–2983.

[10] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[11] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "L1/2 regularization," *Science China Information Sciences*, vol. 53, no. 6, pp. 1159–1169, Jun 2010.

[12] L. Niu, R. Zhou, Y. Tian, Z. Qi, and P. Zhang, "Non-smooth penalized clustering via regularized sparse regression," *IEEE Transactions on Cybernetics*, vol. 47, no. 6, pp. 1423–1433, 2017.

[13] L. Chaari, H. Batatia, N. Dobigeon, and J.-Y. Tourneret, "A hierarchical sparsity-smoothness bayesian model for ℓ_0+ ℓ_1+ ℓ_2 regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1901–1905.

[14] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[15] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44.

[16] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation.," in *Scandinavian Conference on Image Analysis (SCIA)*, 2011, pp. 666–675.

[17] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[18] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 314–327.

[19] X. Li, Y. Li, C. Shen, A. Dick, and A. Van Den Hengel, "Contextual hypergraph modeling for salient object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3328–3335.

[20] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe, "Key frame extraction from first-person video with multi-sensor integration," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2017.