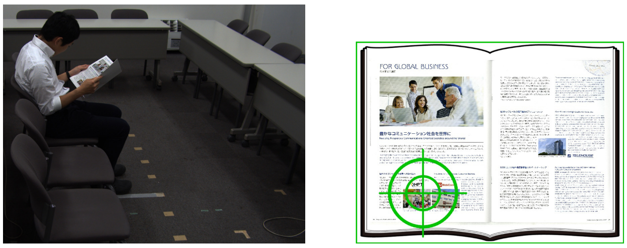
2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Which Content in a Booklet is he/she Reading? Reading Content Estimation using an Indoor Surveillance Camera

Yasutomo Kawanishi*†, Hiroshi Murase*†, Jianfeng Xu†, Kazuyuki Tasaka† and Hiromasa Yanagihara†
*Graduate School of Informatics, Nagoya University, Nagoya, Japan. Email: kawanishi@i.nagoya-u.ac.jp
†KDDI Research, Saitama, Japan



(i) A person reading a booklet. (ii) Example of a result.

Fig. 1: Reading Contents Estimation

***Abstract*—In this paper, we propose a method for estimating reading content in a booklet using an image captured by an indoor surveillance camera. Here, we assume that a reading content can be specified by estimating followings; what booklet, which page of the booklet, and which region in the page. We propose a reading booklet/page estimation method based on image search, and a reading region estimation method focusing on the body pose of the reader. We evaluated the method as a 44 classes classification problem, which consists of eleven pages of booklets and four regions in each pages. We achieved 25.6% in accuracy of the reading content estimation.**

I. Introduction

Booklets or magazines are usually put in a waiting room such as lounge and lobby. People who are waiting for someone usually read such booklets to pass the time. Knowing what are they interested in is important for providing appropriate service for them. It can be realized by knowing which contents in a booklet are they reading. Since surveillance cameras are often installed in such kind of rooms, in this research, we aim to estimate reading contents using an indoor surveillance camera.

In this paper, our goal is to estimate the content in a booklet that a person is reading, by using an indoor surveillance camera. In general, since it is sometimes difficult to simultaneously observe the reading content and the reader, it is difficult to estimate the reading contents directly. We consider that the estimation can be realized by the followings independent procedures; estimating which page is the reader reading and which region in a page is the reader reading.

We show the situation where we assume and an example of the estimation results in Figure 1. To estimate the reading page of the booklet, we need to observe at least a part of the page. We assume the situation as Figure 1 (i) that a camera is mounted on a wall to observe a wide area. In Figure 1 (ii), the estimated page are shown and the estimated reading content is marked with the green circle. In this case, the reading content is successfully estimated.

Under the assumption that all of the pages of the booklets are known, the estimation problem of which page of a booklet can be considered as a specific object retrieval problem. Therefore, we follow the approach for object retrieval by using image features.

In general, to know the reading region of a booklet reader, gaze estimation is required. In this research, we tackle gaze estimation by using a surveillance camera mounted on a wall.

Generally, when we are reading a booklet, we usually hold the booklet to read it easily. At that time, since the face of the reader and the booklet is close, the relative angle of the face and the booklet (i.e. the pose of the reader) is slightly change according to the reading regions even though head position of the reader is fixed. Therefore, by focusing on the readers pose, we propose a reading position estimation from the pose of a reader. For the purpose of estimating reading content in a booklet, since estimating the coordinate of the reading position in continuous value is not necessary, we consider reading position estimation as a classification problem by dividing a page into several regions.

Our contributions in this paper are as follows:

- We separate the reading content estimation process into two individual processes; one is reading page estimation and the other is reading region estimation. The combination is one of the contributions of the paper.
- We propose a method to estimate the reading region of a person from two dimensional coordinates of his/her body joints.

The rest of this paper is organized as follows: In Section II, a brief survey are provided. In Section III, the details of our proposed method are introduced. Experimental results are reported in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

This research is strongly related to image based object retrieval and gaze estimation. In this section, we briefly review these researches.

978-1-5386-3787-6/18/$31.00 ©2018 IEEE

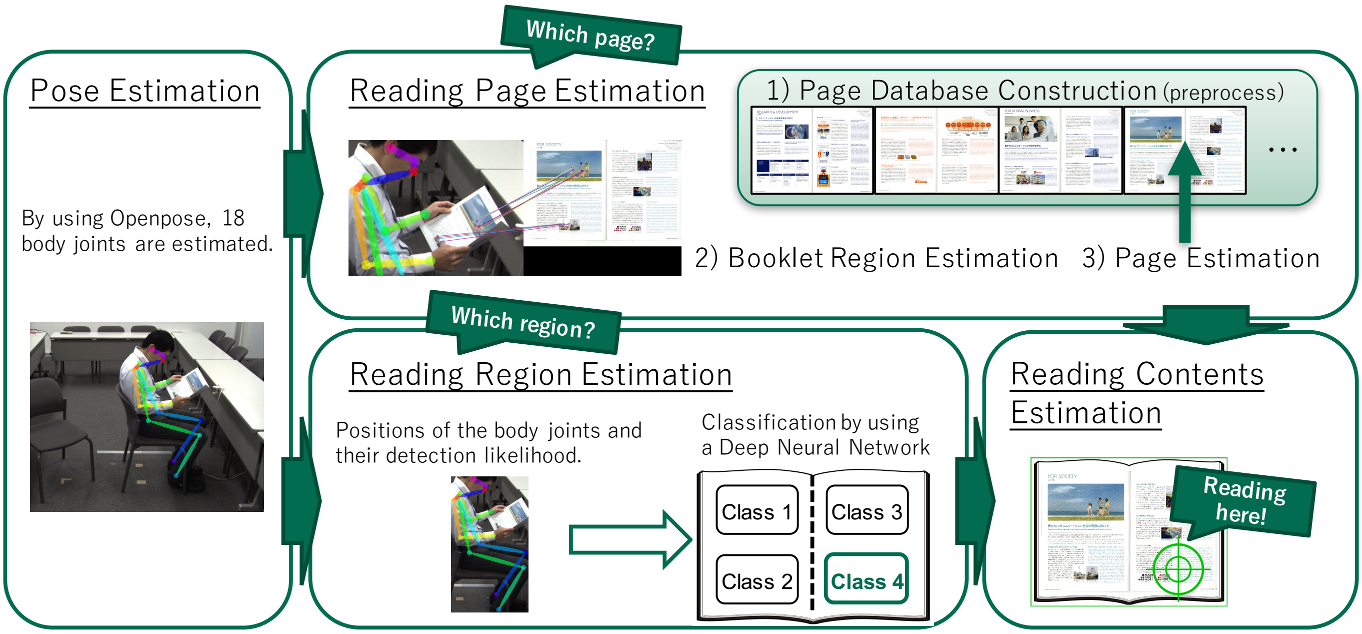

Pose Estimation
By using Openpose, 18 body joints are estimated.
Which page?
Reading Page Estimation
1) Page Database Construction (preprocess)
...
2) Booklet Region Estimation
3) Page Estimation
Which region?
Reading Region Estimation
Positions of the body joints and their detection likelihood.
Classification by using a Deep Neural Network
Class 1
Class 3
Class 2
Class 4
Reading Contents Estimation
Reading here!

Fig. 2: The process flow of the proposed method.

A. Image-based Object Retrieval

Generally, the specific object retrieval is achieved by following steps; keypoints are detected from an query image, features are calculated around the keypoints [1], [2], [3], [4], [5], image similarities are calculated between the query and each image in the dataset by using the keypoints and features, and finally, the image which is best matched with the query is selected from the database [6], [7], [8]. When focusing on a specific target, there are several specific methods. For document retrieval, Locally Likely Arrangement Hashing (LLAH) [9], which is a feature for describing document image feature has been proposed.

Most image features require high resolution input image. However, we use a surveillance camera mounted on a wall which observes a wide area. Since the size of target booklets is small in a captured image, it is difficult to use such kind of features.

B. Gaze Estimation

If we can use an eye tracker, we can easily obtain gaze of a specific reader [10]. However, it is difficult to use such devices since we want to estimate gaze of arbitrary persons. If the captured image is high resolution, we can estimate gazes from an image [11], [12], [13], [14], [15]. Since the size of the target reader is small in a captured image, it is still difficult to observe the face of the reader in high resolution from the front.

There are several researches to estimate gaze directions by using a surveillance camera [16], [17]. Gaze targets in the research are relatively larger than our target, we need more precise gaze estimation.

III. Reading Contents Estimation

A. Overview

Since the reading contents of a person can be specified by the reading page and the reading region as described in the previous section, we can estimate the reading contents by estimating following two points:

- Which page in a booklet is the reader reading?
- Which region in a page is the reader reading?

According to the idea, the proposed reading content estimation method consists of the following two parts:

- Reading Page Estimation: Estimating the reading page of the target person.
- Reading Region Estimation: Estimating the region in the booklet coordinate.

The process flow is shown in Figure 2. We will describe the details in the following subsections.

B. Reading Page Estimation

Since the number of booklets in a waiting room is limited, we can know all pages of the booklets previously. Therefore, we construct a page database which consists of images of all pages beforehand. By searching the most similar page to the page that a person is reading from the database, we can realize reading page estimation.

Since directly estimating the similar page from a captured image with cluttered background is difficult, first we estimate the position of a booklet in the image, and then, we estimate the similar page from the estimated booklet position. By focusing on the positional relation between the booklet and the reader, we estimate the position of the booklet from the positions of face and arms.

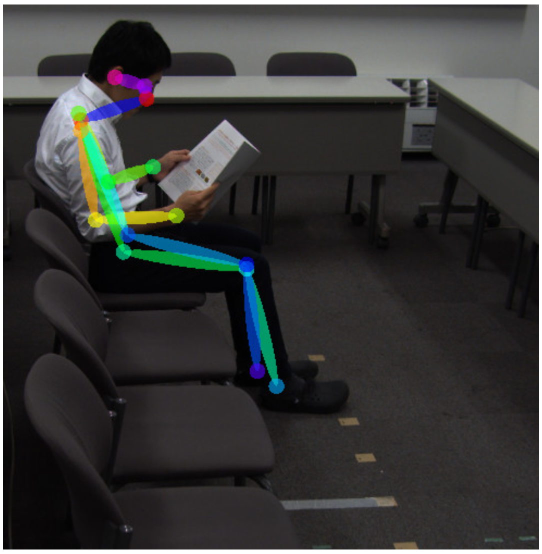


Fig. 3: A pose estimation result.

1) *Page Database Construction:* First of all, we need to prepare a set of images $I = \{i_1, i_2, \dots, i_N\}$ of facing pages of booklets. N denotes the number of facing pages. In case we have multiple booklets, the N denotes the total number of pages. Image feature keypoints $\{\mathbf{k}_n^j\}_{j=1}^{J_n}$ are detected from each image $i_n \in I$, and then, local features $\mathbf{f}_n^j$ are extracted from each keypoint $\mathbf{k}_n^j$. Pairs of an image and its local features $\{(\mathbf{k}_n^j, \mathbf{f}_n^j)\}_{j=1}^{J_n}$ are stored in the page database. In this paper, we use Accelarated KAZE (AKAZE) features [18] for keypoint detection and feature description.

2) *Booklet Region Estimation:* Generally, when we are reading a booklet, we hold the booklet with both hands to keep it easy to read. When we observe a person reading a booklet, the booklet exists around the hands of the person in the observed image. Focus on this fact, we estimate the booklet position from the pose of the person.

To estimate pose of a person, we use a method proposed by Cao et al. [19][1], which can estimate body parts fast and accurately only from an image. By the method, we can obtain 18 body joint positions and their confidences from an input image I^c (Figure 3).

A minimum rectangle region which contains positions of head parts (eyes, nose, and ears)and arms (elbow and wrist joints) estimated from the image I^c is determined. We consider the rectangle expanded to 2 times wider to the wrist side as the booklet region.

3) *Page Estimation:* We crop the booklet region estimated by the method described in the previous section, we named it as a booklet region image I^p.

Keypoints are detected and local features are extracted from the booklet region image I^p, a set of keypoints and features $\{(\mathbf{k}_j^p, \mathbf{f}_j^p)\}_{j=1}^{J^p}$ is obtained.

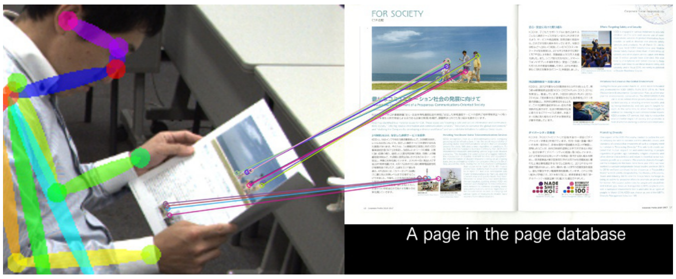


Fig. 4: Keypoint matching to find the reading page.

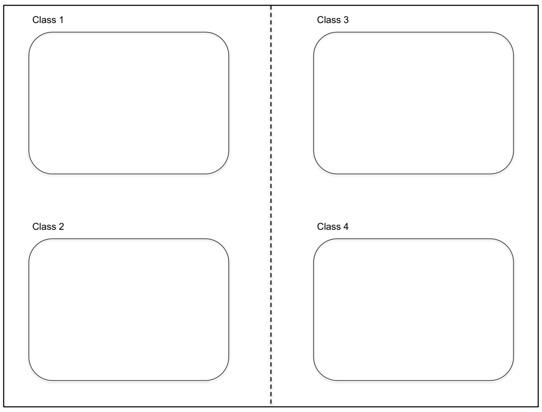


Fig. 5: Positions of the four classes.

By matching keypoints $\{(\mathbf{k}_j^p, \mathbf{f}_j^p)\}_{j=1}^{J^p}$ extracted from the booklet region I^p with each keypoint set $\{(\mathbf{k}_j^n, \mathbf{f}_j^n)\}_{j=1}^{J^n}$ stored in the page database, we select the best match page as the reading page. An example is shown in Figure 4.

C. *Reading Region Estimation*

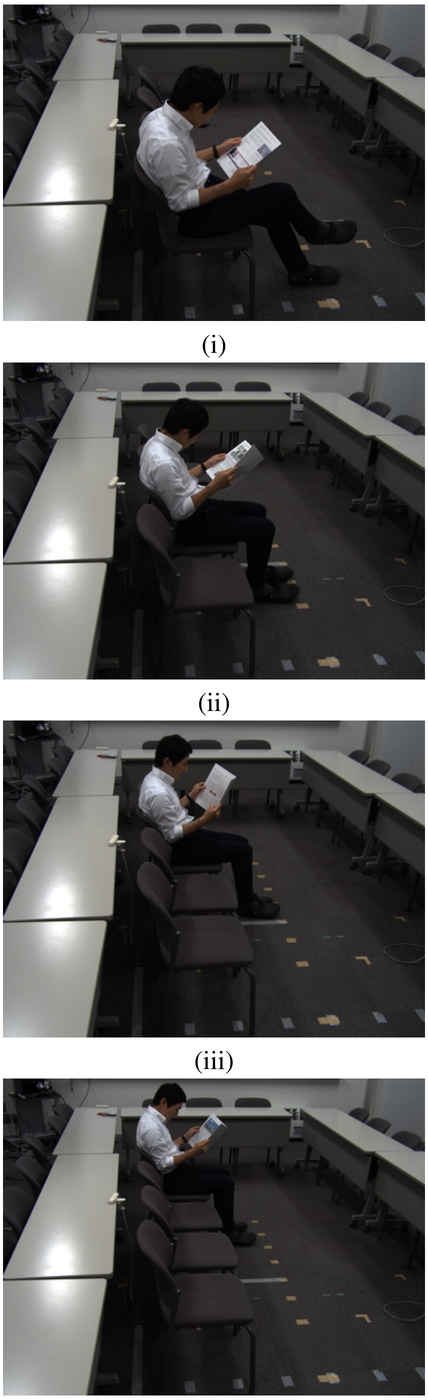
While a person is reading a booklet, pose of the person changes slightly depending on the reading region in a page of the booklet. Therefore, we propose a method to estimate which region is the person reading based on the pose of the person.

In this paper, we follow a classification approach to estimate the reading region by dividing a page of a booklet into four classes; upper-left, lower-left, upper-right, and lower-right as shown in Figure 5.

We build a classifier for reading region estimation. For the input of the classifier, we use position of body joints of the person. Here, we also obtain positions of 18 body joints using a method proposed by Cao et al. [19] as same as section III B 3). As input of the reading region classifier, positions of the 18 body joints and their detection confidence are concatenated into a feature vector $\mathbf{v}$, which is a 54 dimensional vector. Here, we build the reading region classifier $f(\mathbf{v})$, which outputs a vector of scores for the four classes. Note that $\mathbf{v}$ is an output of the deep learning model used in the Cao's method [19].

To make it robust to size, shift, and distances caused by person's position against the camera, positions of 18 body joints are normalized.

[1] OpenPose: https://github.com/CMU-Perceptual-Computing-Lab/openpose



(i)

(ii)

(iii)

(iv)

Fig. 6: Different settings (chair positions): distance variation between the camera and subjects.

For the normalization, we first calculate the bounding box of 18 body joints, and then normalize the width and height of the bounding box into 1.

As a ground truth, we give annotations of reading regions (4 classes) for all images.

TABLE I: The result of reading contents estimation for the chair position (i) in Figure 6.

Evaluation		Accuracy (%)
Overall	Contents (44 classes)	49.0
	Contents (16 classes)	56.9
Individual	Page (11 classes)	64.9
	Page (4 classes)	75.0
	Region (four classes)	76.0

TABLE II: The result of reading contents estimation for all the chair position.

Evaluation		Accuracy (%)
Overall	Contents (44 classesh)	25.6
	Contents (16 classes)	37.9
Individual	Page (11 classes)	41.1
	Page (4 classes)	56.5
	Region (4 classes)	66.3

In this paper, we use a deep learning model consists of 4 fully connected layers as the classifier $f(\mathbf{v})$. We empirically tuned the number of units in each layer, and set 128 units for each hidden layer. To avoid over fitting, the weight parameters for each layer are regularized by L_2-norm, and Dropout [20] is performed in the training phase.

IV. Evaluation

A. Dataset

We construct a dataset for evaluating the proposed method.

First of all, we construct a page database. We scanned eleven pages of a booklet by a scanner, and stored them into the page database.

Then, we collect a set of images of a person reading a booklet in an experimental environment. We use a camera (Grasshopper 3) to capture the scene.

To evaluate the robustness against distances between a person and the camera, we prepare four different settings as showed in Figure 6. For these four settings, we put four chairs at different distances from the camera in a room.

Each subject sat down at each chair, opened a specific page, and read each position of four classes as showed in Figure 5. Four pages are selected from eleven pages in the page database to make the subject read. We prepared three subjects. Each combination of chair positions and reading regions, we captured 200 images. In total, we captured $3 \times 4 \times 4 \times 200 = 9,600$ images.

B. Evaluation Method

To evaluate estimation accuracy of the proposed reading contents estimation method, we calculate the accuracy of a 44 classes (11 pages × 4 regions) classification problem. We also calculate the accuracy of 16 classes (4 pages × 4 regions). The eleven pages contain pages that no subjects read the pages. Additionally, we individually evaluate the subprocesses of the proposed method; the page estimation and the reading region estimation. The evaluation of the page estimation was performed in two settings; calculating the accuracy of the eleven class classification and four class classification problems. The

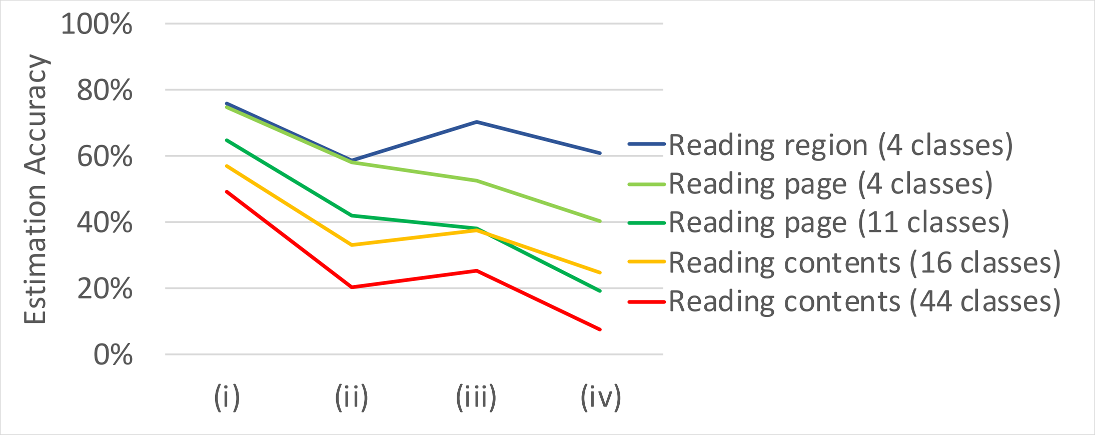


Fig. 7: Estimation accuracy by changing the distance between the camera and a subject. The horizontal axis corresponds to (i), (ii), (iii), and (iv) in Figure 6.

evaluation of reading region estimation was performed by calculating the accuracy of a four classes classification.

For the evaluation, we performed 12-fold cross validation, which consists of the combination of three subjects and four chair positions. We selected a person in a position, trained using the training data which do not contain the person, and evaluate the accuracy of reading region estimation for the selected person. We repeat it for twelve times for each combination of a subject and a chair position, and average the results of twelve cases.

C. Result

First, Table I shows the results for the chair position shown in Figure 6 (i), which is the nearest chair position from the camera. The results are averages of three cases (three subjects) in the twelve cases. The table contains overall and individual evaluations. The overall evaluation stands for the evaluation of reading contents estimation and the individual evaluation stands the evaluations for each procedures, that is, reading page estimation and reading region estimation. The results show the proposed method achieved 49.0% in accuracy.

Table II shows the average of result of all the chair positions. The proposed method achieves 25.6% in 44 class classification accuracy of reading contents estimation. The trend is similar as Table I.

We can guess that the accuracy of reading page estimation and reading region estimation would decrease with increasing distance between the camera and a subject. We performed an analysis on the accuracy by changing the distance. Figure 7 shows the result. The horizontal axis of Figure 7 shows the distances between the camera and a subject, which correspond to (i), (ii), (iii), and (iv) in Figure 6.

Since keypoint detection heavily depends on the resolution of the target, we can confirm that the accuracy of reading page estimation decrease with increasing distance. On the other hand, the accuracy of reading region estimation decreased slightly. It is considered that the training data contained images of other persons sitting on the same distance. Since the degree of poses variation of the subjects was different among the chair positions, it is considered the accuracy in the distance (iii) was better than that of (ii).

V. Conclusion

In this paper, we proposed a method for estimating reading content of booklets using an image captured by an indoor surveillance camera. We divided the reading content estimation into two procedures; the reading page estimation and the reading region estimation. To estimate the reading region of a person, by focusing on the pose when a person is reading a booklet, we introduced a reading region estimation method from the pose of a person. We confirmed that the proposed method achieved 25.6% in accuracy of 44 classes reading content estimation.

Since the proposed reading page estimation method is based on a naïve keypoint matching algorithm, it can be used in a limited situation. As a future work, we need to develop a robust reading page estimation method against heavy occlusion or low resolution. We evaluated the reading page estimation using only one booklet which has eleven pages. To make the method more practical, we need to extend the reading page estimation method which can handle more pages and evaluate it using more booklets.

In the evaluation, we fixed the orientation of subjects, however, for more general situation, we need to evaluate the method over various orientation of subjects. To make the method can handle not only sitting in a chair, but also other pose such as standing or lying would be also a future work.

Acknowledgement

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research.

References

[1] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. of the 7th IEEE International Conference on Computer Vision (ICCV 1999)*, vol. 2, Sep. 1999, pp. 1150–1157.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th European Conference on Computer Vision (ECCV 2006)*, May 2006, pp. 404–417.

[3] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *in Proc. the 6th ACM International Conference on Image and Video Retrieval*, 2007, pp. 494–501.

[4] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. 13th International Conference on Computer Vision (ICCV 2011)*, Nov. 2011, pp. 2548–2555.

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in *Proc. 13th International Conference on Computer Vision (ICCV 2011)*, Nov. 2011, pp. 2564–2571.

[6] H. Jégou, M. Douze, C. Schmid, and P. Prez, "Aggregating local descriptors into a compact image representation," in *Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, Jun. 2010, pp. 3304–3311.

[7] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. 11th European Conference on Computer Vision (ECCV 2010)*, Sep. 2010, pp. 141–154.

[8] F. Perronnin, Y. Liu, J. Snchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, Jun. 2010, pp. 3384–3391.

[9] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *Proc. 7th IAPR International Workshop on Document Analysis Systems (DAS 2006)*, Feb. 2006, pp. 541–552.

[10] A. T. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2007.

[11] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2033–2046, Oct. 2014.

[12] H. Park and D. Kim, "Gaze classification on a mobile device by using deep belief networks," in *Proc. 3rd IAPR Asian Conference on Pattern Recognition (ACPR 2015)*, Nov. 2015, pp. 685–689.

[13] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. 2015 IEEE IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Jun. 2015, pp. 4511–4520.

[14] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 49–56, May 2016.

[15] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Jun. 2016, pp. 2176–2184.

[16] K. Smith, S. O. Ba, J. M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1212–1229, Jul. 2008.

[17] S. Duffner and C. Garcia, "Unsupervised online learning of visual focus of attention," in *Proc. 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2013)*, Aug. 2013, pp. 25–30.

[18] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proc. 24th British Machine Vision Conference (BMVC 2013)*, Sep. 2013.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. 2017 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Jul. 2017, pp. 7291–7299.

[20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.