Segawa *et al. EURASIP Journal on Image and Video Processing* (2018) 2018:33
https://doi.org/10.1186/s13640-018-0272-z

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

First-person reading activity recognition by deep learning with synthetically generated images

Yuta Segawa[1], Kazuhiko Kawamoto[2*] and Kazushi Okamoto[3]

Abstract

We propose a vision-based method for recognizing first-person reading activity with deep learning. For the success of deep learning, it is well known that a large amount of training data plays a vital role. Unlike image classification, there are less publicly available datasets for reading activity recognition, and the collection of book images might cause copyright trouble. In this paper, we develop a synthetic approach for generating positive training images. Our approach synthesizes computer-generated images and real backround images. In experiments, we show that this synthesis is effective in combination with pre-trained deep convolutional neural networks and also our trained neural network outperforms other baselines.

Keywords: First-person activity recognition, Deep learning, Image classification

1 Introduction

With the development of wearable cameras, first-person activity recognition has been a popular topic in recent years [1]. There are many conventional approaches which tackle first-person activity recognition. Some of these approaches employ motion feature such as optical flow and also a classifier, e.g., LogitBoost and SVM (support vector machine)[2, 3]. In recent years, DCNN (deep convolutional neural network), the state-of-the-art model for visual recognition, has been proposed [4] and then applied to several tasks on first-person activity recognition.

Although DCNN models provide remarkable results for image recognition, they require a large amount of labeled training samples. Fine-tuning is a promising method for reducing the amount of required training samples and also reducing training time [5–8]. Unfortunately, there are no large-scale datasets for first-person activity recognition, while datasets for image recognition, such as ImageNet [9] and Places [10], are publically available. For this reason, even if using fine-tuning, DCNN models for first-person activity recognition require collection and annotation of large-scale FPV (first-person vision) videos. Castro et al. [11] actually collected 40,000 images in 26 weeks by recording with a wearable camera and annotated all of the images.

To cope with such time-consuming issues, we propose a synthetic method for generating training images. Our method consists of three steps, as shown in Fig. 1, which are image synthesis, fine-tuning DCNN, and recognizing activities from natural FPV images. In this paper, we focus on reading activity recognition. Reading is a pervasive and intellectual activity in daily life, and its recognition can be useful for building context-aware interfaces [12], life log systems [13, 14] and experience sharing systems [15]. In addition, the image synthesis approach is useful for reading activity recognition, because one must take care of copyright issues when collecting the images of books and magazines for training. Conversely, if one wants to recognize other activities, it is easier to collect the images of the related objects such as displays and keyboards. In this paper, we contribute in:

- Reduction of the collection and annotation costs on deep learning dataset by using simple image generation and synthesis
- Methodology of applying the synthesis approach to recognition of first-person activities

*Correspondence: kawa@faculty.chiba-u.jp
[1]NIFTY Corporation, 2-21-1, Kita-Shinjuku, Shinjuku 169-8333, Tokyo, Japan
Full list of author information is available at the end of the article

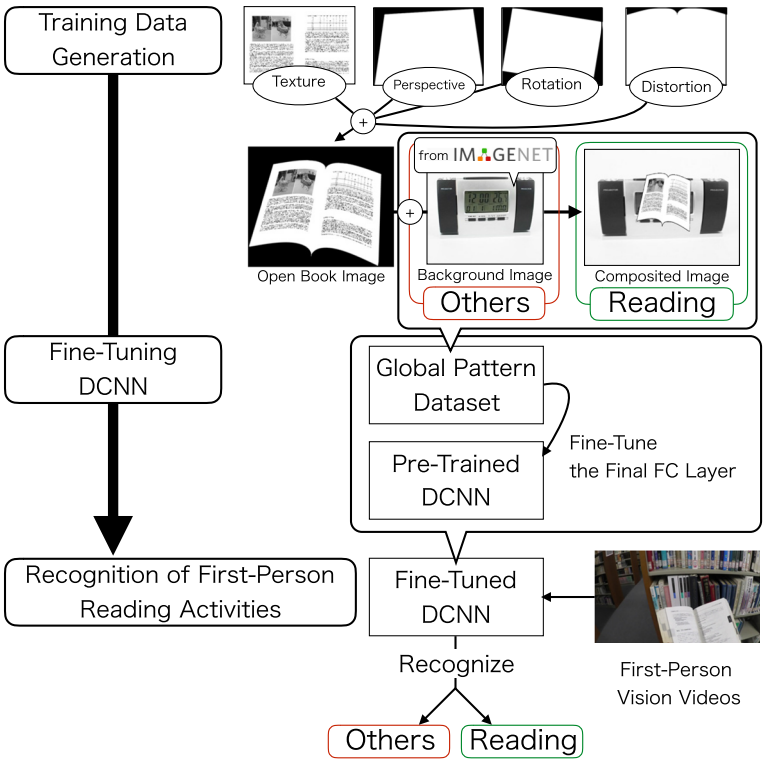

Fig. 1 Overview of our approach

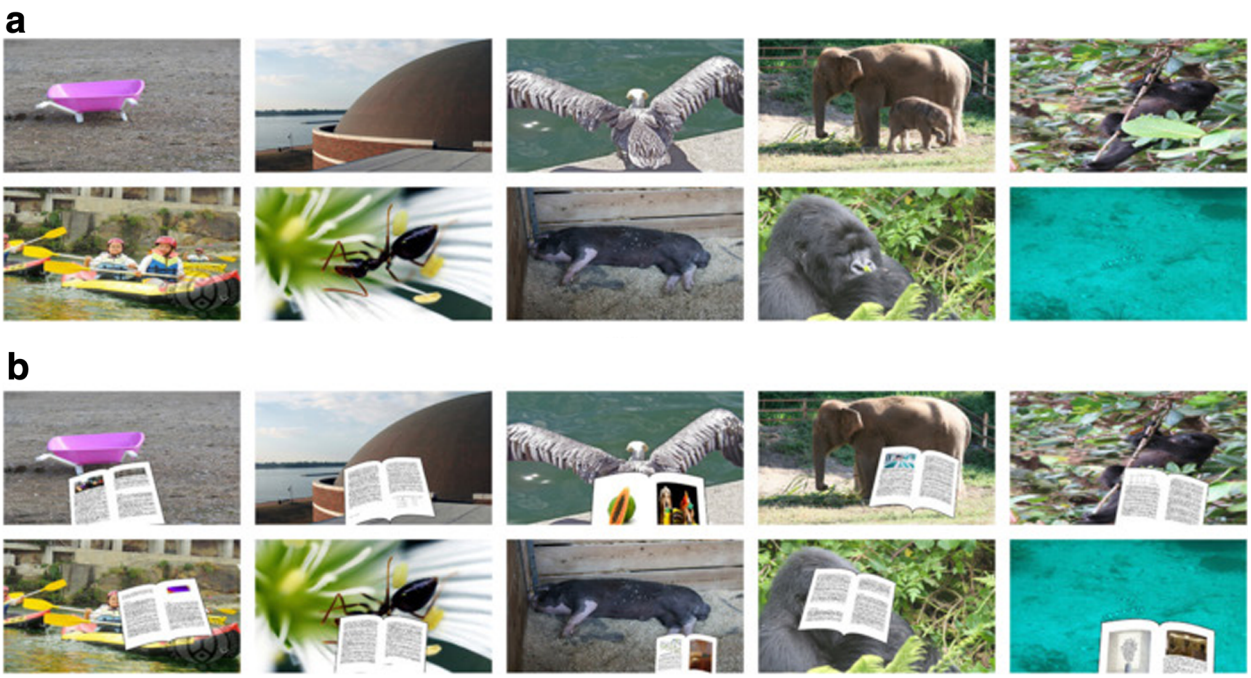

Fig. 2 Synthetic training samples. **a** "Others" class. **b** "Reading" class

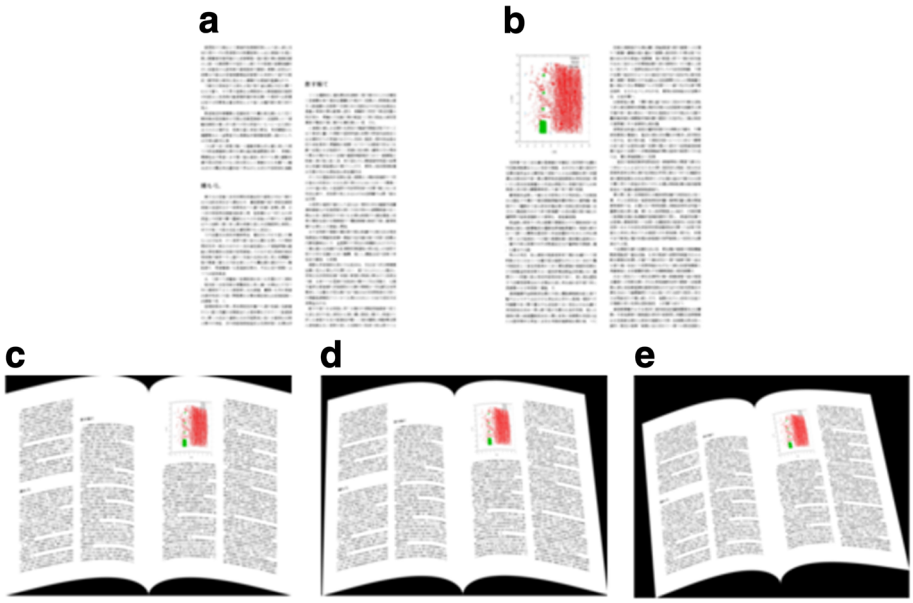

Fig. 3 Computer-generated book image. **a** Texture drawing (left page). **b** Texture drawing (right page). **c** Distortion. **d** Perspective. **e** Rotation

- Interpretations of the synthesis approach using simple visual patterns in terms of representation in deep visual patterns

The rest of this paper is organized as follows. In Section 2, we introduce some related works about our approach: first-person activity recognition, deep convolutional neural networks, and synthetically image generation. In Section 3, we proposes the image synthetic method for generating training images. In Section 4, we show the adaptability of our synthetic images to first-person reading activity recognition with real FPV videos. In addition, we also demonstrate the generalizability of deep features from the synthetic images. In Section 5,

Table 1 Main parameters in the generation processes

Process	Parameters	Values
–	Canvas of each page	$210\sqrt{2} \times 210$ pixel by appropriate scaling
T	Blank space	10% at the top, bottom, left, and right in the canvases
	Columns	1 or 2
	Pages containing figures	80% of the entire generated pages
	Place to put figures	The top or the bottom in a column
	Category of figures	Mathematical figures, tables, and general pictures
	Figure size	Height: 12.5 $\sim$ 50%, or 100% of the column height
		Width: fixed at the same length of the column
	Size of headlines	10% of the page height
	Place to put headlines	Somewhere outside the figure areas
	Place to put texts	Entire areas in the column except the figures and headlines
	Text format	Japanese characters in random order (without any rules)
	Line-breaking	Done with prob. 1% every time putting a character
D	Distortion strength	$\alpha \in [0.1, 0.3]$
P	Homography matrix	Determined as shown in Fig. 4
R	Rotation angles	In $[-10°, 10°]$
–	Resize	The long side shrinks to 256 pixel with keeping the ratio

Fig. 4 Perspective projection by a randomly determined homography matrix

we discuss our method in terms of image synthesis processes through further experiments on several types of the synthetic training dataset.

2 Related works

2.1 First-person activity recognition

Body-mounted devices help to record personal information and to analyze personal activities. A kind of such devices, 3-axis accelerometers, provides user's posture estimations, and many researchers employ them for life-logging [16, 17]. Head-mounted cameras also have become popular with development of camera downsizing and high-efficiency video coding [18–21].

For first-person activity recognition, there are several methods based on image segmentation [22, 23] and object recognition [24, 25]. These approaches include multistage recognition processes, and hence, recognition errors tend to be stacked. To avoid explicit object recognition, many studies use motion feature such as optical flow with a classifier such as LogitBoost and SVM [2, 3, 26–29].

2.2 Deep convolutional neural network

DCNN models such as AlexNet [4], VGGNet [30], GoogLeNet [31], and ResNet [32] have been proposed and demonstrated remarkable performances in image classification. In first-person activity recognition, there are DCNN-based methods in which optical flow [5, 6] and pooled motion features [33] are used as image features. Moreover, LSTM (long short-term memory) model, which is also a kind of deep models for learning data with recursive expressions, has been introduced together with DCNN models aiming at additionally learning temporal correlations of activities [7, 8].

2.3 Synthetic image generation

For data augmentation, image synthesis is a useful approach for reducing the effort of manually annotation. Wong et al. [34] investigate a benefit of data augmentation for MNIST handwritten character dataset. For object detection, Khail et al. [35] propose an image synthetic method in which real object images and real background images are synthesized. For text localization in natural images, Gupta et al.[36] also propose an image synthetic method in which computer-generated texts and natural real images are synthesized. Sun and Saenko [37] and Su et al.[38] employ 3D CAD object models with real background images for image synthesis. Castro et al.[39] propose a method of generating synthetic

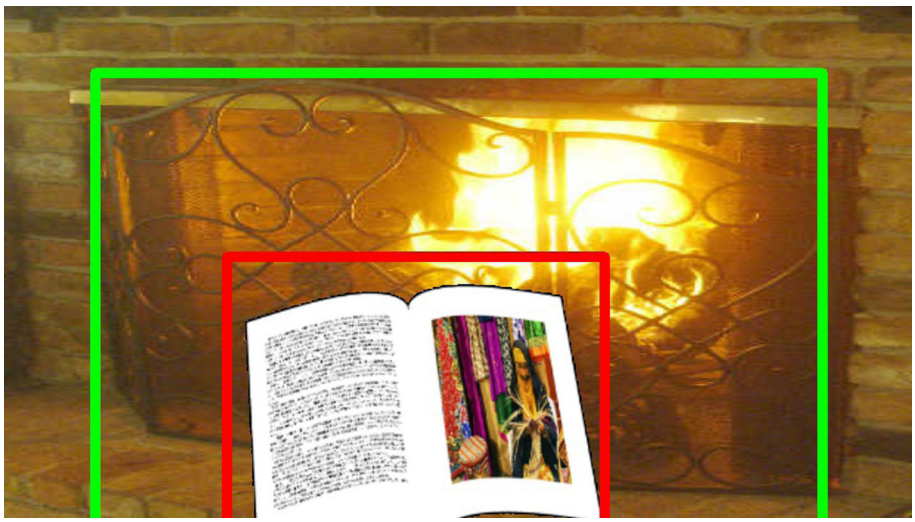

Fig. 5 Bounding region (green rectangle) for book image location

structural magnetic resonance images for learning schizophrenia.

3 Methodology of image synthesis

For recognition of first-person reading activities, we synthesize training samples of "Reading" class. The "Reading" class samples represent visual patterns of open books in FPV images, as shown in Fig. 2b. This section describes a procedure of generating the computer-generated book images and superimposing them on real background images. In addition, we explain how to prepare "Others" class images as the negative class.

3.1 Generation of book images

Generation procedure of the computer-generated book images can be divided into four types of image processing: texture drawing, edge distortion, perspective projection, and rotation. We call these processes T, D, P, and R, respectively. Figure 3 shows examples of results in these processes, and Table 1 shows the parameters for the processes in detail.

The process T aims to reproduce textures like real books on a white canvas so as to be an open book image, as shown in Fig. 3a,b. This process consists of two steps which are the determination of a layout and the drawing of a texture. First, we prepare two white canvases corresponding to the left and the right page of an open book. Next, we determine a layout in each page by randomly placing figures, headlines, and texts on each page.

The process D makes the shape of the book images distorted as shown in Fig. 3c. For the left pages, we set a coordinate system such that the origin is at the left-bottom corner in the canvas, the positive direction of X-axis is right-to-left and of Y-axis is bottom-to-top. For the right pages, we reverse this coordinate horizontally. We distort the image by moving pixel at (x, y) to (x, y') using $y' = y + f(x)$. Here, the distortion function $f(x)$ is defined by $f(x) = \alpha(x\sqrt{1-x^2})$, where α is a parameter for controlling the strength of distortion. In experiments, we set the strength parameters as $\alpha = 0.1$.

The process P is a perspective projection, as shown in Fig. 3d. In order to generate FPV-like appearances, we determine the perspective projection, as shown in Fig. 4, in which each of the original left-top and right-top corners (red points) is moved to a point randomly selected in the rectangles near the corners (blue rectangles). In experiments, we set the width and height of the rectangle as 10% of the image ones.

The process R is a rotation transformation, as shown in Fig. 3e. Here, we only rotates the images with a slight degree of angle to generate open book images caused by reading activity. Note that the black regions around the books are assigned to the transparent color.

3.2 Image synthesis

For image synthesis, we superimpose the computer-generated book images onto real background images, as shown in Fig. 5. As real background images, we use images which are randomly selected from the ImageNet dataset [9]. In Section 5.3, we will demonstrate that the use of the ImageNet images is superior to the use of other domain-specific background images.

We set a bounding region on the background images (the green rectangle in Fig. 5) and randomly put the computer-generated book images inside the bounding region. This region is used for preventing the book images from being put on the periphery of the background images.

3.3 Negative samples against reading activities

The "Others" class is the negative class and hence represents various visual patterns except for the "Reading" class images. For the "Others" class images, we simply use the background images which are used for generating the "Reading" class images. In other words, the "Reading" and the "Others" samples only have a difference whether

Table 2 GoogLeNet(v3) architecture

Index	Module type	Output shape
1	Input	299 × 299 × 3
2	Convolution	149 × 149 × 32
3	Convolution	147 × 147 × 32
4	Convolution	147 × 147 × 64
5	Max pooling	73 × 73 × 64
6	Convolution	73 × 73 × 80
7	Convolution	71 × 71 × 192
8	Max pooling	35 × 35 × 192
9	Inception	35 × 35 × 256
10	Inception	35 × 35 × 256
11	Inception	35 × 35 × 256
12	Inception	17 × 17 × 736
13	Inception	17 × 17 × 768
14	Inception	17 × 17 × 768
15	Inception	17 × 17 × 768
16	Inception	17 × 17 × 768
17	Inception	8 × 8 × 1280
18	Inception	8 × 8 × 2048
19	Inception	8 × 8 × 2048
20	Average pooling	2048
21	Output	2

an open book appears or not, as shown in Fig. 2b and a, respectively.

4 Experimental results with real first-person vision videos

In this section, we report the performance of our synthetic method with real FPV videos. In the experiments, we compare our DCNN model with other baseline models.

4.1 Fine-tuning of DCNN model

We use GoogLeNet(v3) [40] as a pre-trained DCNN model with the ImageNet dataset, as shown in Table 2. In fine-tuning, we retrain only the final layer of the model because the fine-tuning of deeper layers degrades the performance from our preliminary experiment. We optimize the parameters with the cross-entropy loss function using the SGD (stochastic gradient descent) algorithm. In the optimization, we use 25,000 training samples per class and feed mini batches in size of 10.

4.2 Test dataset

For evaluation, we prepare 20 real FPV videos, as shown in Fig. 6. In order to prove the generalization of the proposed method, we record the videos with two types of wearable cameras at four different places. The label "Reading" or "Others" for each image is provided by manual annotation in which "Reading" is provided if an open book is located around the center of the image. Table 3 shows a summary of the our dataset. In addition, we use publicly available two datasets including "Reading"

Fig. 6 Example images of 20 videos for test. **a** "Others" class. **b** "Reading" class

Table 3 Summary of our dataset used for evaluation

Recording devices	HX-A500[2]	Looxcie2[3]
Image size	960 × 540	480 × 320
Number of videos	3 videos	17 videos
Sampling rate	5–6 fps	
Number of images (per video)	100–700 per a video	

Table 4 Summary of public dataset used for evaluation

Database name	MEAD [42]	LENA [41]
Image size	1280 × 720	430×240
Number of videos	10 videos	20 videos
Sampling rate	1 fps	
Number of images (per video)	77	152–156

class: LENA (Life-loggine EgoceNtric Activities) dataset [41] and MEAD (Multimodal Egocentric Activity Dataset) [42], as shown in Fig. 7. Table 4 shows a summary of the two datasets used for our evaluation. Note that Table 4 includes only the "Reading" class data. For "Others" class, we randomly select 3830 images, which is the same as the number of the "Reading" class images, from other activity images such as "Writing" and "Working at PC" in the two datasets.[1] [2]

4.3 Evaluation result

Using our synthetic dataset, we compare the DCNN model with two baselines 1-NN (1-nearest neighbor) and linear SVM in terms of precision, recall, and F-measure. For 1-NN and linear SVM, we use Fisher vectors [43] as image feature. Fisher feature has been often used in the image classification tasks as well as bag-of-visual-words [44].

We show the experimental results in Tables 5 and 6. We find that the DCNN model significantly outperforms the other two baselines on all datasets in terms of the averaged F-measure. Further improvements might be possible if motion features like optical flow are used in DCNN models (e.g., [6]), but we only evaluate the methods without motion features in order to clearly show the effectiveness of using synthetic images.

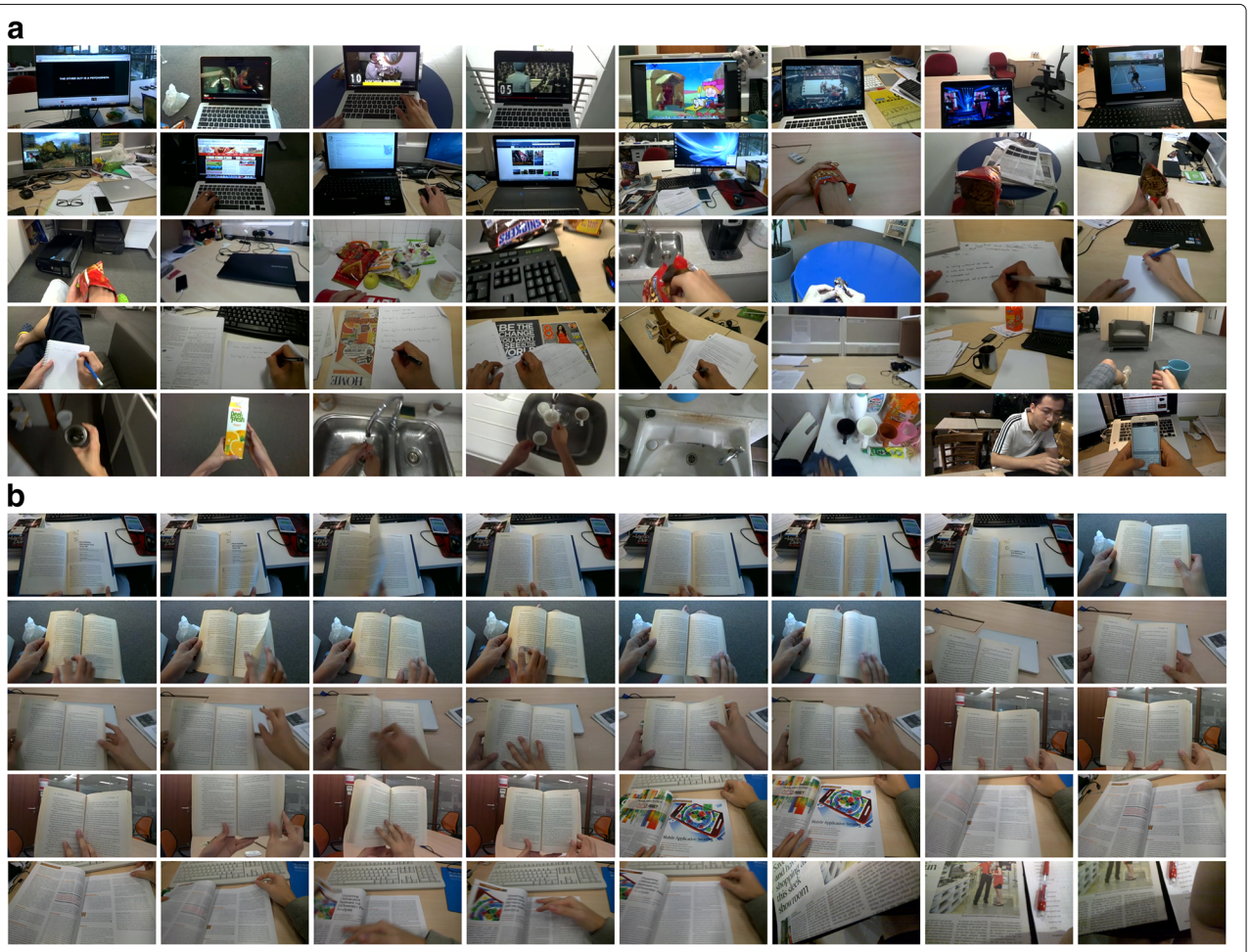

Fig. 7 Example images of two public datasets for test. **a** "Others" class. **b** "Reading" class

Table 5 Comparative result with our dataset in Table 3

	Baselines				Proposed		
	1-NN + FV		SVM + FV		DCNN		
	Others	Reading	Others	Reading	Others	Reading	
Scores	Precision	19.6	84.0	36.6	84.5	*70.8*	*92.9*
for each class	Recall	58.7	51.5	47.4	79.5	*69.9*	*95.0*
	F-measure	24.6	61.0	33.4	80.5	*64.9*	*93.5*
Averages	Precision	51.8		60.5		*81.9*	
over classes	Recall	55.1		63.5		*82.5*	
	F-measure	42.8		56.9		*79.2*	

Best methods for a given measure are specified in italic type

Table 6 Comparative result with the public dataset in Table 4

	Baselines				Proposed		
	1-NN + FV		SVM + FV		DCNN		
	Others	Reading	Others	Reading	Others	Reading	
Scores	Precision	54.2	53.1	*88.5*	50.7	79.0	*92.8*
for each class	Recall	46.1	61.1	3.0	*99.6*	*94.2*	74.9
	F-measure	49.8	55.8	5.8	62.2	*85.9*	*82.9*
Averages	Precision	53.7		69.6		*85.9*	
over classes	Recall	53.6		51.3		*84.6*	
	F-measure	52.8		34.0		*84.4*	

Best methods for a given measure are specified in italic type

Table 7 Effect of changing the number of training samples

	Number of training samples for each class						
	1000		5000		25,000		
	Others	Reading	Others	Reading	Others	Reading	
Scores	Precision	*73.3*	89.2	70.8	92.9	70.1	*96.2*
for each class	Recall	46.0	*98.4*	69.9	95.0	*84.2*	92.0
	F-measure	52.9	93.3	64.9	*93.5*	*72.0*	93.4
Averages	Precision	81.2		81.9		*83.1*	
over classes	Recall	72.2		82.5		*88.1*	
	F-measure	73.1		79.2		*82.7*	

Best methods for a given measure are specified in italic type

Fig. 8 Example images synthesized by a combination of the four generation processes: *R*, *P*, *D*, and *T*

Table 8 Results of the comparison in the generation processes

	None	*P*	*D*	*T*	PD	DT	PT	PDT	Avg.
Without R	56.7	41.4	62.0	71.3	60.5	79.5	71.6	81.2	65.5
With R	50.1	43.4	50.6	73.5	64.9	79.2	66.2	79.2	63.4
Differences	−6.6	2.0	−11.4	2.2	4.4	−0.3	−5.4	−2.0	−2.1
	None	*R*	*D*	*T*	RD	DT	RT	RDT	Avg.
Without P	56.7	50.1	62.0	71.3	50.6	79.5	73.5	79.2	65.4
With P	41.4	43.4	60.5	71.6	64.9	81.2	66.2	79.2	63.6
Differences	−15.3	−6.7	−1.5	0.3	14.3	1.7	−7.3	0.0	−1.8
	None	*R*	*P*	*T*	RP	PT	RT	RPT	Avg.
Without D	56.7	50.1	41.4	71.3	43.4	71.6	73.5	66.2	59.3
With D	62.0	50.6	60.5	79.5	64.9	81.2	79.2	79.2	69.6
Differences	5.3	0.5	19.1	8.2	21.5	9.6	5.7	13.0	10.4
	None	*R*	*P*	*D*	RP	PD	RD	RPD	Avg.
Without T	56.7	50.1	41.4	62.0	43.4	60.5	50.6	64.9	53.7
With T	71.3	73.5	71.6	79.5	66.2	81.2	79.2	79.2	75.2
Differences	14.6	23.4	30.2	17.5	22.8	20.7	28.6	14.3	21.5

Differences in each comparison about a process *X* show the values by substraction of scores with *X* from ones without *X*

5 Discussion on the image synthesis in deep learning

In this section, we verify our synthetic approach in more detail and discuss it.

5.1 Effect of changing the number of training samples

First, we verify the effect of changing the number of training samples. Table 7 shows the results with 1000, 5000, and 25,000 training samples. For the "Reading" class, we observe that the F-measure roughly keeps constant over the three cases. On the other hand, for the "Others" class, we observe that the F-measure improves with increasing the number of training samples. Since images in the "Others" class are diverse, the increase is especially effective for recognizing other activities.

5.2 Effect of changing combinations of the generation processes

Our synthesis approach consists of the four image processings, *T*, *D*, *P*, and *R* as mentioned in Section 3.1. Here, we verify which process is effective in improving the recognition performance. To do this, we generate sets of images as shown in Fig. 8 in which the most left column indicates the processes used for image synthesis. For examples, the images at the "None" row are generated by no image processing and the images at "PDT" are generated by the combination of the three image processings: *P*, *D*, and *T*.

From Table 8, we observe that the process *T* always brings the best improvement of the F-measure (21.5% in average) and the process *D* the second one (10.4% in average). This result means that the two processes *T* and *D* produce discriminative features for recognition while the other two processes *R* and *P* provide less discriminative power. In fact, if book regions are overexposed like Fig. 9, the proposed method fails to recognize such images. We further verify which combination contributes in improving the performance. In Fig. 10, we summarize each contribution (average F-measure difference) of the possible combinations of the processes. For example,

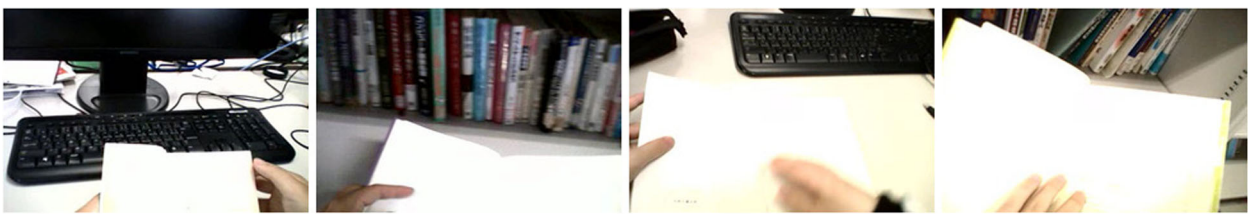

Fig. 9 Recognition failure examples

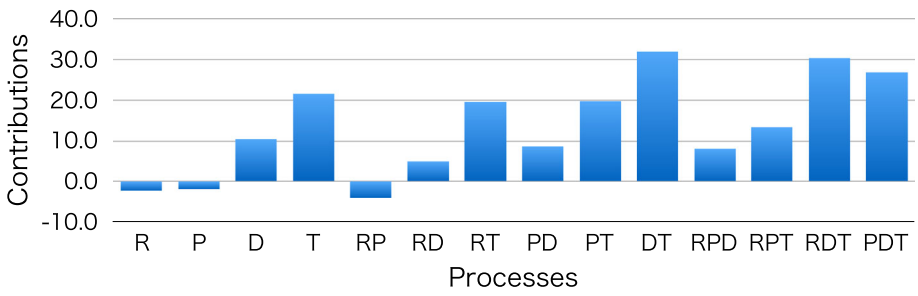

Fig. 10 Contributions of each process combination

the bar chart at "DT" indicates the F-measure difference averaged between DT and the other possible combinations, None, *R*, *P*, and RP. We observe that the combination of *D* and *T* is the most effective in improving the F-measure, and based on the above, we conclude that the two processes *D* and *T* are required in the image synthesis for producing discriminative features and the other processes *R* and *P* should be used in combination with DT.

5.3 Effect of using domain-specific backgrounds

In the abovementioned experiments, we use images in the ImageNet dataset as background images, as shown in Fig. 2. In order to verify the effect of using the background images, we evaluate the recognition performance with domain-specific backgrounds, as shown in Fig. 11. We recorded the background images at the same places where we do our dataset in Table 3.

We show the experimental results in Table 9. We find that the ImageNet dataset provides better performance than the domain-specific one in terms of F-measure. In particular, we observe the increase of F-measure for the "Others" class. Since we use the pre-trained DCNN model with ImageNet, the use of ImageNet enables the efficient learning.

6 Conclusions

We propose a method of synthetically generating training samples for deep learning. The proposed method synthesizes book images from simple computer-generated patterns and real background images. The synthetic approach is particularly useful for recognizing reading activity because of the copy right issues, i.e., capturing books with a digital camera and their use often causes trouble.

From the comparison with the two baselines, we find that our synthetic dataset is effective in combination with

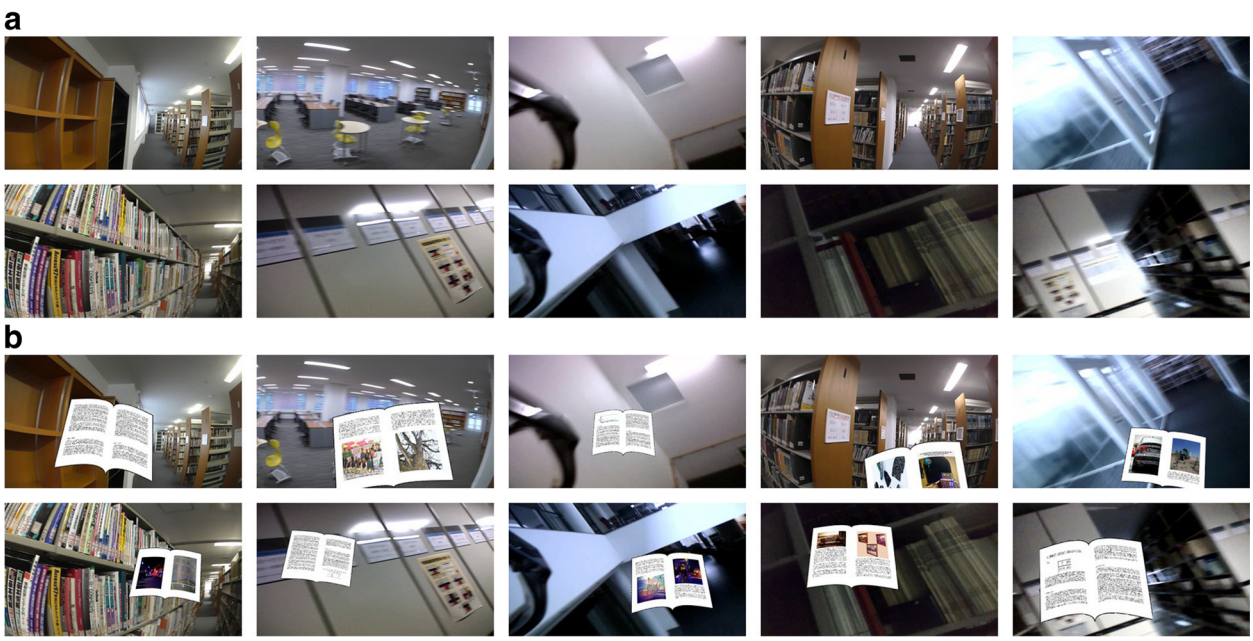

Fig. 11 Synthetic training samples with domain-specific background. **a** "Others" class. **b** "Reading" class

Table 9 Results of the comparison with the dataset synthesized by domain-specific background images

	Synthesized background images				
	Domain-specific		ImageNet		
	Others	Reading	Others	Reading	
Scores	Precision	68.0	92.2	70.8	*92.9*
for each class	Recall	57.3	*95.3*	69.9	95.0
	F-measure	57.1	*93.5*	*64.9*	*93.5*
Averages	Precision	81.2		*81.9*	
over classes	Recall	72.2		*82.5*	
	F-measure	73.1		*79.2*	

Best methods for a given measure are specified in italic type

the DCNN model. In addition, we find that the use of ImageNet images as background brings an improvement in recognizing the activities in the "Others" class. These results are promising for deep learning-based recognition because we are able to easily prepare a large number of training images.

Endnotes

[1] http://panasonic.jp/wearable/a500/

[2] http://www.looxcie.com/

Acknowledgements
The authors would like to thank I. Nakamura for the technical assistance with the experiments.

Funding
This work was supported by JSPS KAKENHI grant number 16K00231.

Availability of data and materials
You can see an example video of the experimental results at https://www.kawa-lab.org/en/research/fpv_ar.

Authors' contributions
The first author mainly developed and evaluated the system and wrote this manuscript. The second author managed this research project by continually advising the first author. The third author supported the system development and experimental evaluation. All authors read and approved the final manuscript.

Authors' information
Yota Segawa received the B.E. and M.E. degrees from Chiba University, Japan, in 2015 and 2017, respectively. He is currently with NIFTY Corporation, Japan. Kazuhiko Kawamoto received the B.E., M.E., and Ph.D. degrees from Chiba University, Japan, in 1997, 1999, and 2002, respectively. From 2002 to 2005, he was an assistant professor in Tokyo Institute of Technology, Japan. From 2005 to 2009, he was an associate professor in Kyushu Institute of Technology, Japan. He is currently a professor in Chiba University, Japan. His research interests include computer vision, pattern recognition, and statistical signal processing. Kazushi Okamoto received the B.E. and M.E. degrees from Kochi University of Technology, Japan, in 2006 and 2008, respectively. He received the Ph.D. degree from Tokyo Institute of Technology, Japan, in 2011. From 2011 to 2015, he was an assistant professor in Chiba University, Japan. He is currently an assistant professor in the University of Electro-Communications, Japan. His research interests include machine learning, data mining, and recommender system.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1] NIFTY Corporation, 2-21-1, Kita-Shinjuku, Shinjuku 169-8333, Tokyo, Japan. [2] Graduate School of Engineering, Chiba University, 1-33, Yayoicho, Inage 263-8522, Chiba, Japan. [3] Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu 182-8585, Japan.

Received: 20 June 2017 Accepted: 24 April 2018
Published online: 14 May 2018

References

1. A Betancourt, P Morerio, CS Regazzoni, M Rauterberg, The evolution of first person vision methods: a survey. IEEE Trans. Circ. Syst. Video Technol. **25**(5), 744–760 (2015)
2. Z Lu, K Grauman, in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference On*. Story-driven summarization for egocentric video, (2013), pp. 2714–2721
3. MS Ryoo, L Matthies, in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference On*. First-person activity recognition: what are they doing to me?, (2013), pp. 2730–2737
4. A Krizhevsky, I Sutskever, GE Hinton, in *Advances in Neural Information Processing Systems 25*, ed. by F Pereira, CJC Burges, L Bottou, and KQ Weinberger. Imagenet classification with deep convolutional neural networks, (2012), pp. 1097–1105
5. A Takamine, Y Iwashita, R Kurazume, in *2015 IEEE/SICE International Symposium on System Integration (SII)*. First-person activity recognition with c3d features from optical flow images, (2015), pp. 619–622
6. M Ma, H Fan, KM Kitani, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Going deeper into first-person activity recognition, (2016), pp. 1894–1903
7. S Song, V Chandrasekhar, B Mandal, L Li, J-H Lim, G Sateesh Babu, P Phyo San, N-M Cheung, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Multimodal multi-stream deep learning for egocentric activity recognition, (2016), pp. 378–385
8. D Crandall, C Fan, in *European Conference on Computer Vision International Workshop on Egocentric Perception, Interaction, and Computing (EPIC)*. Deepdiary: automatically captioning lifelogging image streams (Springer International Publishing, Cham, 2016), pp. 459–473
9. O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, A Karpathy, A Khosla, M Bernstein, AC Berg, L Fei-Fei, ImageNet large scale visual recognition challenge. Int. J. Comput. Vision (IJCV). **115**(3), 211–252 (2015)
10. B Zhou, A Khosla, À Lapedriza, A Torralba, A Oliva, Places: an image database for deep scene understanding. CoRR. **abs/1610.02055** (2016)
11. D Castro, S Hickson, V Bettadapura, E Thomaz, G Abowd, H Christensen, I Essa, in *Proceedings of the 2015 ACM International Symposium on Wearable Computers. ISWC '15*. Predicting daily activities from egocentric images using deep learning, (2015), pp. 75–82

12. A Bulling, JA Ward, H Gellersen, G Tröster, *Robust recognition of reading activity in transit using wearable electrooculography*. (J Indulska, DJ Patterson, T Rodden, M Ott, eds.) (Springer, Berlin, Heidelberg, 2008), pp. 19–37
13. T Kimura, R Huang, S Uchida, M Iwamura, S Omachi, K Kise, in *2013 12th International Conference on Document Analysis and Recognition*. The reading-life log—technologies to recognize texts that we read, (2013), pp. 91–95
14. K Kunze, Y Shiga, S Ishimaru, K Kise, in *2013 12th International Conference on Document Analysis and Recognition*. Reading activity recognition using an off-the-shelf EEG—detecting reading activities and distinguishing genres of documents, (2013), pp. 96–100
15. K Kise, O Augereau, Y Utsumi, M Iwamura, K Kunze, S Ishimaru, A Dengel, in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. UbiComp '17*. Quantified reading and learning for sharing experiences, (2017), pp. 724–731
16. AM Khan, YK Lee, SY Lee, TS Kim, A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. IEEE Trans. Inf. Technol. Biomed. **14**(5), 1166–1172 (2010)
17. M Zhang, AA Sawchuk, Human daily activity recognition with sparse representation using wearable sensors. IEEE J. Biomed. Health Inf. **17**(3), 553–560 (2013)
18. C Yan, Y Zhang, F Dai, J Zhang, L Li, Q Dai, Efficient parallel hevc intra-prediction on many-core processor. Electron. Lett. **50**(11), 805–806 (2014)
19. C Yan, Y Zhang, F Dai, X Wang, L Li, Q Dai, Parallel deblocking filter for hevc on many-core processor. Electron. Lett. **50**(5), 367–368 (2014)
20. C Yan, Y Zhang, J Xu, F Dai, J Zhang, Q Dai, F Wu, Efficient parallel framework for HEVC motion estimation on many-core processors. IEEE Trans. Circ. Syst. Video Technol. **24**(12), 2077–2089 (2014)
21. C Yan, Y Zhang, J Xu, F Dai, L Li, Q Dai, F Wu, A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. IEEE Signal Proc. Lett. **21**(5), 573–576 (2014)
22. EH Spriggs, FDL Torre, M Hebert, in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Temporal segmentation and activity classification from first-person sensing, (2009), pp. 17–24
23. KM Kitani, T Okabe, Y Sato, A Sugimoto, in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*. Fast unsupervised ego-action learning for first-person sports videos, (2011), pp. 3241–3248
24. A Fathi, A Farhadi, JM Rehg, in *Proceedings of the 2011 International Conference on Computer Vision. ICCV '11*. Understanding egocentric activities, (2011), pp. 407–414
25. H Pirsiavash, D Ramanan, in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On*. Detecting activities of daily living in first-person camera views, (2012), pp. 2847–2854
26. K Zhan, F Ramos, S Faux, in *Control Automation Robotics Vision (ICARCV), 2012 12th International Conference On*. Activity recognition from a wearable camera, (2012), pp. 365–370
27. K Zhan, V Guizilini, F Ramos, in *Control Automation Robotics Vision (ICARCV), 2014 13th International Conference On*. Dense motion segmentation for first-person activity recognition, (2014), pp. 123–128
28. Y Yan, E Ricci, G Liu, N Sebe, Egocentric daily activity recognition via multitask clustering. IEEE Trans. Image Process. **24**(10), 2984–2995 (2015)
29. L Xia, I Gori, JK Aggarwal, MS Ryoo, in *2015 IEEE Winter Conference on Applications of Computer Vision*. Robot-centric activity recognition from first-person RGB-D videos, (2015), pp. 357–364
30. K Simonyan, A Zisserman, Very deep convolutional networks for large-scale image recognition. CoRR. **abs/1409.1556** (2014)
31. C Szegedy, W Liu, Y Jia, P Sermanet, SE Reed, D Anguelov, D Erhan, V Vanhoucke, A Rabinovich, Going deeper with convolutions. CoRR. **abs/1409.4842** (2014)
32. K He, X Zhang, S Ren, J Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Deep residual learning for image recognition, (2016), pp. 770–778
33. MS Ryoo, B Rothrock, L Matthies, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Pooled motion features for first-person videos, (2015), pp. 896–904
34. SC Wong, A Gatt, V Stamatescu, MD McDonnell, Understanding data augmentation for classification: when to warp?. CoRR. **abs/1609.08764** (2016)
35. O Khalil, ME Fathy, DKE Kholy, ME Saban, P Kohli, J Shotton, Y Badr, in *2013 IEEE International Conference on Image Processing*. Synthetic training in object detection, (2013), pp. 3113–3117
36. A Gupta, A Vedaldi, A Zisserman, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Synthetic data for text localisation in natural images, (2016), pp. 2315–2324
37. B Sun, K Saenko, in *Proceedings of the British Machine Vision Conference*. From virtual to reality: fast adaptation of virtual object detectors to real domains, (2014)
38. H Su, CR Qi, Y Li, LJ Guibas, in *2015 IEEE International Conference on Computer Vision (ICCV)*. Render for CNN: viewpoint estimation in images using CNNC trained with rendered 3d model views, (2015), pp. 2686–2694
39. E Castro, A Ulloa, SM Plis, JA Turner, VD Calhoun, in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. Generation of synthetic structural magnetic resonance images for deep learning pre-training, (2015), pp. 1057–1060
40. C Szegedy, V Vanhoucke, S Ioffe, J Shlens, Z Wojna, Rethinking the inception architecture for computer vision. CoRR. **abs/1512.00567** (2015)
41. S Song, V Chandrasekhar, N-M Cheung, S Narayan, L Li, J-H Lim, in *ACCV Workshops (3)*. Activity recognition in egocentric life-logging videos (Springer International Publishing, Cham, 2014), pp. 445–458
42. S Song, NM Cheung, V Chandrasekhar, B Mandal, J Liri, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Egocentric activity recognition with multimodal Fisher vector, (2016), pp. 2717–2721
43. F Perronnin, J Sánchez, T Mensink, in *Proceedings of the 11th European Conference on Computer Vision: Part IV. ECCV'10*. Improving the Fisher kernel for large-scale image classification (Springer, Berlin, 2010), pp. 143–156
44. G Csurka, CR Dance, L Fan, J Willamowski, C Bray, in *In Workshop on Statistical Learning in Computer Vision, ECCV*. Visual categorization with bags of keypoints, (2004), pp. 1–22