Recognizing Context for Privacy Preserving of First Person Vision Image Sequences

Sebastiano Battiato, Giovanni Maria Farinella(✉), Christian Napoli,
Gabriele Nicotra, and Salvatore Riccobene

Department of Mathematics and Computer Science, University of Catania,
Viale A. Doria 6, 95125 Catania, Italy
{battiato,gfarinella,napoli,riccobene}@dmi.unict.it

Abstract. The constant increasing evolution of life-logging wearable devices, as well as the fast grow of their market, has introduced relevant changes in the acquisition, storage and automatic understanding of images and videos. Along with the novel users' opportunities, this technology is introducing a large amount of privacy-related concerns, mainly regarding the unaware or unwilling contexts subject that could get recorded by a life-logging device. In this work, we devise an approach to help life-logging wearable devices enforcing restrictions for context-related users' privacy preservation. The proposed approach joins different technological innovations, from computer vision techniques to bluetooth beacon technology and network security.

1 Introduction

Since late 1980 s the world of wearable devices has encountered a tremendous evolution while components miniaturization enabled us to freely interact with a wide range of mobile systems and implement them in many aspects of everyday life [1]. Such devices have begun to overwhelmingly interact with personal information domain. Moreover an increasing amount of such devices is equipped with built-in cameras and has introduced relevant changes in the acquisition, storage and automatic understanding of images and videos. Therefore, while the novel availability of images and videos has encouraged personal creativity, it has also raised a certain amount of privacy concerns, mainly regarding the unaware or unwilling contexts subject, that could get caught on such multimedia contents. Moreover relevant legal implication should be taken into account [2], especially regarding the large amount of data continuously produced by the so called life-logging devices [3–6]. Nowadays, these devices allow their users to continuously record and share online many different kinds of data, as videos, audio, pictures, personal data, as well as collective information or individual activities. On the other hand, while traditional devices as cameras or audio recording devices were only used sporadically and deliberately, modern life-logging devices can record and share their data continuously, therefore tampering with bystanders' expectations about privacy and discretion [7]. For these reasons privacy and discretion aspects have gained great importance; as a matter of facts the typical user of

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 580–590, 2017.
https://doi.org/10.1007/978-3-319-68548-9_53

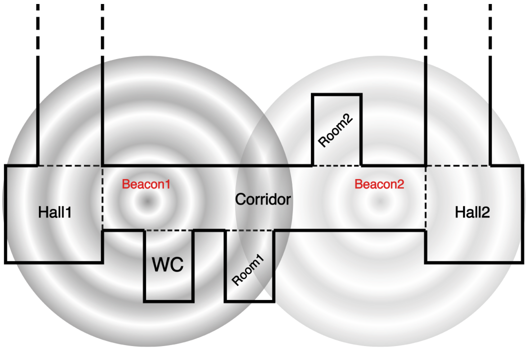


Fig. 1. Contexts recorded into video recorded at the Department of Mathematics and Computer Science of Catania's University Campus.

such life-logging devices may prefer to enforce privacy through location based control of image collection, in order to avoid later burdensome review of all collected media. Finally, automatic face recognition software performances are now almost as good as human abilities [8], therefore on one hand they offer a useful service, on the other hand they can put at even greater risk personal privacy e.g. taking into account the treat represented by malwares which could seize private multimedial contents surreptitiously [9,10].

2 Proposed System

In this work, we present an overall architecture for context related privacy preservation. The system has been designed to work in places affected by an high level of similarity among different contexts. Therefore the presented approach enforces privacy constraints by applying computer vision methods as well as low energy bluetooth technology for context recognition.

We tested the proposed method in nontrivial use cases, therefore, we decided to use a low-end commercial wearable to record portions of our University campus facilities with an high degree of similarity (e.g. offices). Specifically the raw video data were collected wearing the Recon Jet $^{\text{TM}}$ smart glasses and recording while walking trough several rooms, lounges and hallways (see Fig. 1).

2.1 Scenario and Communication Protocol

In order to grant users' privacy and enforce all the required security measures, the proposed system has been provided with an ad hoc communication protocol (Fig. 2). The protocol is enforced with the following steps:

1. Environment identification;
2. Generation of session encryption key for row file transmission;
3. Cloud Service for Policies handling;
4. Handled file retrieving.

In our protocol we assume the presence of trustworthy users and untampered device. This restriction is based on the fact that no-one can prevent the recording of image or sound by uncooperative or nasty user, with hidden camera. In these cases, defining privacy policy or restriction is totally useless.

Instead we want to focus our attention on a scenario where a user with wearable device wants to respect rules relative to the environment where it is, obtaining from the environment itself the privacy policy defined by others. In this sense, all the encryption operations are finalized to prevent any image acquisition by unauthorized user before privacy rules application.

Finally, we assume that the "owner" (or at least the bystanders) of a specific location, have uploaded to the Cloud System a set of preferences or rules in order to determine whether or not enforce any privacy-related restriction when the context of interest is recognized.

In the following formalism we will define three agents: a generic wearable device W^n that grabs the environment images, a generic beacon B^m that identifies the particular portion of the environment, and a Cloud Service C that handles the recorded images.

The first phase (*environment identification*) involves both the wearable device and the nearest beacon. The beacon broadcasts continuously its identity, providing its $\mathtt{ID}_{\mathtt{b}}^{m}$ and a cloud-related public key $\mathtt{K}_{\mathtt{Pub}}^{C}$, so that

$$B^m \rightarrow W^n : \mathtt{ID}_{\mathtt{b}}^{m}, \mathtt{K}_{\mathtt{Pub}}^{C}$$

The Cloud Service couple of public and private keys $\mathtt{K}_{\mathtt{Pub}}^{C}\mathtt{K}_{\mathtt{Priv}}^{C}$ is generated by an independent Certification Authority; this step provides the properties of authentication and confidentiality for the Cloud Service. The public key can be obviously retrieved also in other ways.

After the detection of the beacon's presence, the wearable device generates a session key $\mathtt{K}_{\mathtt{S}}^{n,i}$ which is used to encrypt the video recorded, in the follow called $V^{n,i}$, and a timestamp $T^{n,i}$, which is used to identify univocally the video.

In the follow, the encrypted information is represented with the common bracket formalism $\{\cdot\}_k$, where k is the encryption key.

The encrypted video is stored locally until network connection availability or a user interaction. When the connection is available or the device owner decides, the stored cyphered data is upload to the Cloud Service:

$$W^n \rightarrow C : \mathtt{ID}_{\mathtt{b}}^{m}, \left\{V^{n,i}\right\}_{\mathtt{K}_{\mathtt{S}}^{n,i}}, \left\{T^{n,i}, \mathtt{K}_{\mathtt{S}}^{n,i}, \mathtt{K}_{\mathtt{R}}^{n,i}, [\mathtt{ID}_{\mathtt{b}}^{l}]_{l \neq m}\right\}_{\mathtt{K}_{\mathtt{Pub}}^{C}}$$

In this transmission the wearable device sends:

- the $\mathtt{ID}_{\mathtt{b}}^{m}$ in clear text;
- the recorded video $\{\tilde{V}^{n,i}\}$ encrypted with the Cloud public key;
- a tuple with a transmission timestamp $T^{n,i}$, the session encryption key, a response encryption key $\mathtt{K}_{\mathtt{R}}^{n,i}$, the list of beacon listened by the device, except $\mathtt{ID}_{\mathtt{b}}^{m}$.

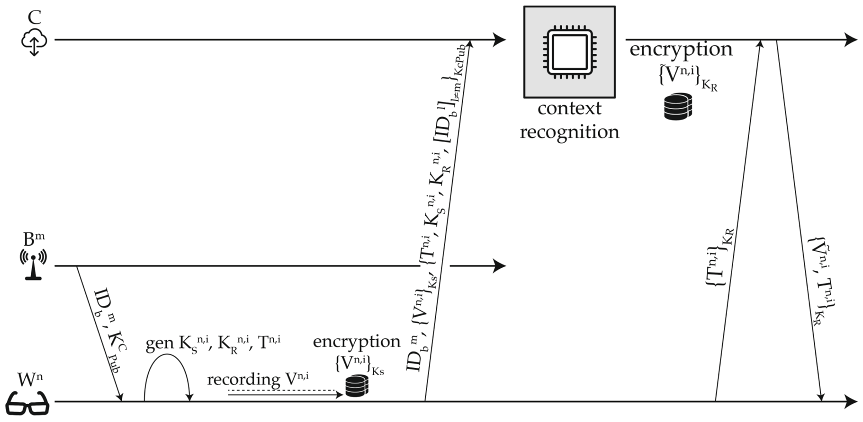


Fig. 2. Network protocol

The Cloud Service, that owns its $\mathtt{K}^{C}_{\mathtt{Priv}}$ key, is the only one able to decrypt the last part of the received communication. It retrieves the session key $\mathtt{K}_{\mathtt{S}}^{n,i}$, so it cat decode the video. By means of the list of beacon ID, it can retrieve the privacy policies previously defined, applying them to the video.

After this described communication, and the related message decoding, the Cloud resident application is able to recognize the context of each image through computer vision algorithms, and apply the required privacy enforcement rules. Only after this process, and the blurring of privacy concerned images, the resulting video can be transmitted back to the wearable device's owner.

Before its transmission from the cloud service to the users client, the processed video is re-encrypted using the response key $\mathtt{K}_{\mathtt{R}}^{n,i}$, to avoid unauthorized accesses.

The wearable device finally requests to the Cloud Service repository the transmission of the handled video:

$$W^n \rightarrow C : \left\{T^{n,i}\right\}_{\mathtt{K}_{\mathtt{R}}^{n,i}}$$

$$C \rightarrow W^n : \left\{T^{n,i}, \tilde{V}^{n,i}\right\}_{\mathtt{K}_{\mathtt{R}}^{n,i}}$$

Note that the response key $\mathtt{K}_{\mathtt{R}}^{n,i}$ and the timestamp $T^{n,i}$ can be provided by the wearable device to every authorized user, to realized independently steps 3 and 4. For this reason $\mathtt{K}_{\mathtt{R}}^{n,i}$ must be different from $\mathtt{K}_{\mathtt{S}}^{n,i}$.

3 Visual Context Recognition

In Sect. 2.1 we described the designed communication protocol between the wearable device and a cloud-resident application. In this section we will describe how the cloud application proceeds with the required context recognition. This latter goal is achieved by using dedicated computer vision algorithms as well as machine learning solutions. Once the context is identified, then, the cloud is responsible for applying the required privacy preserving policies. Such policies will be applied by blurring the images regarding contexts for which the users have required a privacy enforcement rule. In the following we will compare two different implementations for the proposed approach. The first uses Bag of Words for feature extraction and k-Nearest Neighbors algorithm for context recognition (see Sect. 3.1). The second approach uses AlexNet for feature extraction and Support Vector Machine for context recognition (see Sect. 3.2). Finally these two implementation are compared on the base of their results and performances (see Sect. 4).

3.1 Bag-of-Words and k-Nearest Neighbors Algorithm

The Bag-of-Words (BoW) [11] method was born for information retrieval in text document analysis. For image processing purposes it is possible to apply the same model by creating a vocabulary of *visual words* constructed as a catalog of visual features. BoW model relies a distance based features clustering. The features are extracted from local regions after keypoint detection. It is possible to apply the BoW model for image classification by the following steps:

1. extract local regions from Points of Interests;
2. compute and extract local descriptors on these local regions;
3. compute a visual vocabulary through the clustering of the local descriptor;
4. represent an image as distribution of its visual word with respect to the computed visual vocabulary.

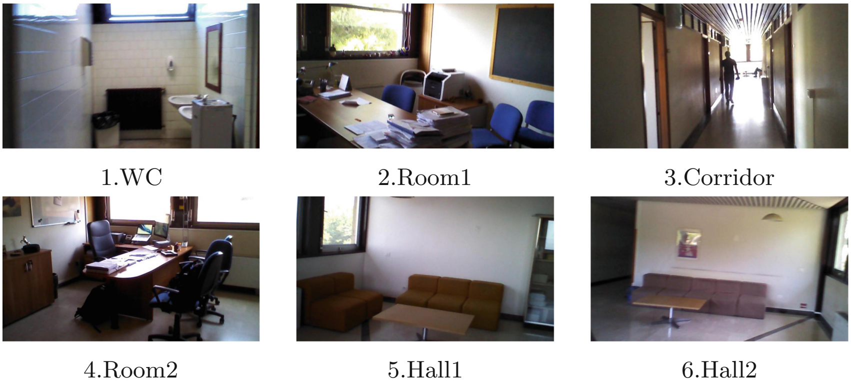
In this work, the BoW model has been used with Dense-SURF as features. The algorithm has been instructed to use an 8 by 8 pixel grid. The visual vocabulary obtained with k-means clustering is constituted by 1024 visual words.

In our solution we created different classifiers, one for each beacon, in order to assist context recognition. We used the following set-ups. We split the dataset in three parts and we used one or two parts for training and only one for testing.

The k-Nearest Neighbors algorithm [12] (k-NN) is the algorithm we used for classification when BoW is employed as representation. This algorithm is based on the prediction of the class of an image, considering k training data neighbors. In our study we used 1-NN algorithm implementation.

3.2 AlexNet and SVM Algorithm

AlexNet [13] is a convolution neural network (CNN) for objects recognition. AlexNet is composed by 650000 neurones triggered by 60 millions input parameters. The AlexNet model has been trained on a subset of ImageNet dataset composed by 1.2 million images of 1000 categories. We used AlexNet as alternative of BoW for image representation purpose.


1.WC

2.Room1

3.Corridor

4.Room2

5.Hall1

6.Hall2

Fig. 3. Example of images into dataset

We have coupled AlexNet representation with an SVM [12] classifier. We used SVM in multiclass procedure. This algorithm is based on the construction of detach clusters. In this study we have six detach cluster one for each class.

4 Experimental Settings and Results

The experiments were designed to test the proposed system employing beacon by comparing it's efficacy while used to improve two well known classification methods based on Bag-of-Words [11] and AlexNet [13]. The classic classification methods, therefore, will be taken as reference baseline for the result presented in the next sections. The Bag-of-Words model has been used jointly with the k-Nearest Neighbors [12] algorithm for classification. Similarly, AlexNet has been used and then feed a SVM [12] algorithm for classification purpose. Our dataset is composed by six classes each of them related to a different context (see Fig. 1).

4.1 Dataset

The dataset is composed of video frames (Fig. 3). The frames have been collected from a set of recorded videos. Such videos have been captured with a ReconJet by an operator walking through one wing of the building (Fig. 1). We performed many simulations regarding each one of the presented methods. In order to collect sufficient data and minimize statistical interferences we used several configurations for both the training and the testing set. The dataset has been split in three equally sized partitions (T1, T2, and T3), moreover three mixed partitions have been created: T12 combining T1 and T2, T13 combining T1 and T3, and T23 combining T2 and T3. Each of these partitions have been used independently for training in each simulation while paired with a complementary test set partition (see Table 1).

Table 1. Correlation between beacon, images and classes. N_B is the number of images for the classifier training step in the based method and our solution. N_C is the number of images per class. The parameter K is 1 for single part of the dataset (e.g. `T1`) and 2 for combined parts (e.g. `T12`).

Beacon #	N_B	Class	N_C	Policy
-	$K * 4170$	Hall1	$K * 695$	Yes
		Corridor		Yes
		WC		No
		Room1		No
		Room2		Yes
		Hall2		No
Beacon #	N_B	Class	N_C	Policy
1	$K * 2780$	Hall1	$K * 695$	Yes
		Corridor		Yes
		WC		No
		Room1		No
2	$K * 2085$	Corridor	$K * 695$	Yes
		Room2		Yes
		Hall2		No

4.2 BoW Model

In order to compare our beacon based solution with the foremost standards, initially we used a BoW model and k-NN algorithm to obtain a reference baseline. Table 2 shows the results of such a model for each possible combination of training and testing sets (see Table 2). In this phase the classifier has been trained for context recognition among all possible classes with no restriction. Therefore the BoW/k-NN model has been applied to the context of each frame with respect to six different contexts (see Fig. 1). While each test has produced consistent results, on the other hand, it should be noticed that, when we used `T3` as test set, the accuracy of this classifier produces less accurate results. Similarly, also when `T3` is used for training, the resulting classifier obtains a very low accuracy. Effectively the `T3` dataset is affected by a relevant noise (e.g. blurred or overexposed frames, too dark or too bright scenes, etc.). On the other hand we also noticed that using `T3` combined with another set among `T1` or `T2` for training highly increases the classification capabilities of the classifier. We suspect that `T3` contributes to train the classifier for context recognition even with noisy data.

4.3 Beacon-Enhanced BoW Model

Table 2 also shows the performances obtained by an improved version of the BoW model which makes use of beacon-driven context recognition (see Sect. 2). In this

Table 2. Performance of the different proposed setups based on BoW/kNN and AlexNet/SVM representation: α^{BoW} is the baseline accuracy of the standard BoW model, α^{AN} is the baseline accuracy of the standard AN model, α_*^{BoW} is the accuracy of our improved BoW-based model with beacon driven context classification, α_*^{AN} is the accuracy of our improved AN-based model.

Training set	Test set	α^{BoW}	α^{AN}	Beacon	α_*^{BoW}	α_*^{AN}
T1	T2	78.56%	74.75%	1	83.17%	84.14%
				2	93.33%	92.04%
T1	T3	67.41%	67.99%	1	80.68%	79.10%
				2	84.56%	82.97%
T2	T1	80.79%	75.97%	1	79.96%	84.10%
				2	96.45%	95.78%
T2	T3	69.40%	69.86%	1	78.13%	86.73%
				2	81.58%	81.97%
T3	T1	73.43%	70.29%	1	82.73%	80.72%
				2	83.31%	87.43%
T3	T2	68.80%	71.70%	1	82.09%	85.42%
				2	77.79%	84.70%
T12	T3	74.68%	75.01%	1	86.33%	84.03%
				2	85.13%	83.88%
T13	T2	83.29%	80.10%	1	88.13%	87.05%
				2	92.71%	89.74%
T23	T1	83.55%	78.03%	1	85.43%	85.58%
				2	95.20%	94.20%

phase we used two classifiers, one for each beacon involved in our experiments. The first classifier has been used to detect the classes associated with context related to the first beacon: *Hall1*, *WC*, *Room1* and *Corridor*. The second classifier has been used to recognize the remaining classes related to the second beacon: *Corridor*, *Room2* and *Hall2*. The data provided to the classifiers where similar to the data used for the standard BoW modes (Sect. 4.2). Moreover, for this second experiment, the device also stored a tag for each frame with a ID list of the beacons in range at recording time. This setup permitted us to obtain an higher system's accuracy (compare columns 3 and 4 of Table 2 w.r.t. the columns 6 and 7). Finally, as in the previous experiment, also in this second scenario the T3 showed the same noise-related issues.

4.4 AlexNet Model

In order to prove the efficacy of the proposed beacon-driven context recognition with respect to the standard image recognition based models, we tested and

Table 3. Improvements with respect to BoW model: α^{BoW} is the baseline accuracy of the standard BoW model, α_1^{BoW} and α_2^{BoW} are the accuracies related respectively to the classes belonging to the first or second beacon in our improved BoW model, $\alpha_{1,2}^{BoW}$ is the average accuracy of our improved BoW model

Training set	α^{BoW}	α_1^{BoW}	α_2^{BoW}	$\alpha_{1,2}^{BoW}$	Improvement
T1	72.98%	81.92%	88.94%	84.92%	∼12%
T2	75.09%	79.04%	89.01%	83.31%	∼12%
T3	71.11%	82.41%	80.55%	81.61%	∼10%
T12	74.68%	86.33%	85.13%	85.82%	∼11%
T13	83.29%	88.13%	92.71%	90.09%	∼7%
T23	83.55%	85.43%	95.20%	89.61%	∼6%

Table 4. Improvements with respect to the AlexNet (AN) model: α^{AN} is the baseline accuracy of the standard AN model, α_1^{AN} and α_2^{AN} are the accuracies related respectively to the classes belonging to the first or second beacon in our AN-based model, $\alpha_{1,2}^{AN}$ is the average accuracy of our AN-based model

Training set	α^{AN}	α_1^{AN}	α_2^{AN}	$\alpha_{1,2}^{AN}$	Improvement
T1	71.37%	81.98%	88.15%	84.62%	∼13%
T2	72.91%	85.66%	89.86%	87.46%	∼15%
T3	70.99%	82.73%	86.26%	84.24%	∼13%
T12	75.01%	84.03%	83.88%	83.96%	∼9%
T13	80.10%	87.05%	89.74%	88.20%	∼8%
T23	78.03%	85.58%	94.20%	89.27%	∼11%

compared a hybrid approach. In this setup we preprocessed the video frames by using AlexNet [13] obtaining for each frame a feature vector. Then we used such a feature vector as input for an SVM classification algorithm. As done previously (see Sects. 4.2 and 4.3), also for this hybrid method we compare the results of an unconstrained test, that we used as comparison baseline, with our improved beacon-driven approach. Similarly to the previous experiments, also this time the noisy dataset T3 affected the classification accuracy of our implemented models. Moreover, despite AlexNet architecture should be robust with respect to such kind of noise, in our experiment we noticed that a strongly noise video recording could tamper it. On the other hand, if T3 is used in conjunction with a low-noise dataset, it seems to improve the accuracy of the classifier (see Sect. 4.5).

4.5 Discussion

The results of the experiments are reported in Tables 2, 3, and 4. In Table 2 we report the performance of the standard Bag-of-Words (α^{BoW}) and AlexNet (α^{AN}) approaches, as well as the performances of our improved models (α_*^{BoW}

and α_*^{AN}). These two latter also make use of beacon driven context classification to improve their accuracy. The same results are reported in columns third and fourth of Table 2. In Table 3 the performance of the implemented BoW-based models are analyzed with respect to the different set of classes (whether if related to beacon 1 or beacon 2): α^{BoW} is the baseline accuracy of the standard BoW model, α_1^{BoW} and α_2^{BoW} are the accuracies related respectively to the classes belonging to the first or second beacon in our improved BoW model, $\alpha_{1,2}^{BoW}$ is the average accuracy of our improved BoW-based model. Table 4 shows the improvement introduced by our modifications to the AlexNet model: α^{AN} is the baseline accuracy of the standard AlexNet model, α_1^{AN} and α_2^{AN} are the accuracies related respectively to the classes belonging to the first or second beacon in our improved AN model, $\alpha_{1,2}^{AN}$ is the average accuracy of our improved AN model (see Tables 3 and 4). Finally, Fig. 4 shows an overview of the implemented methods and the related improvements introduced with the proposed beacon-driven recognition techniques.

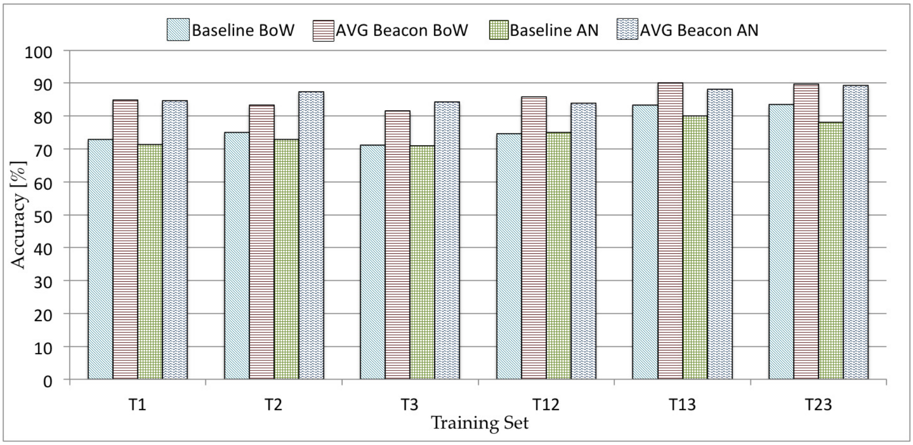


Fig. 4. Comparison of Bag-of-Words and AlexNet representation with and without the exploitation of beacon

5 Conclusions

In this work, we presented a hybrid approach to help life-logging wearable devices enforcing restrictions for context-related users' privacy preservation. The introduction of bluetooth beacon technology have been proven useful to improve the context recognition accuracy of some known image classification solutions based on Bag-of-Words and AlexNet representation. The results showed that the proposed solution is both robust to noise affected datasets as well as efficient for environments that presents an high degree of similarity between different contexts. Moreover, the developed system is highly customizable to enforce the privacy choices of the context owners or bystanders. Finally, the cloud oriented support make it suitable for a wide range of different devices and applications.

References

1. Mann, S.: Wearable computing: a first step toward personal imaging. Computer **30**(2), 25–32 (1997)
2. Cheng, W.C., Golubchik, L., Kay, D.G.: Total recall: are privacy changes inevitable? In: Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, pp. 86–92. ACM (2004)
3. Allen, A.L.: Dredging up the past: lifelogging, memory, and surveillance. Univ. Chicago Law Rev. **75**(1), 47–74 (2008)
4. Chen, Y., Jones, G.J.: Augmenting human memory using personal lifelogs. In: Proceedings of the 1st Augmented Human International Conference, p. 24. ACM (2010)
5. Ortis, A., Farinella, G.M., D'Amico, V., Addesso, L., Torrisi, G., Battiato, S.: Organizing egocentric videos for daily living monitoring. In: Proceedings of the First Workshop on Lifelogging Tools and Applications, pp. 45–54. ACM (2016)
6. Furnari, A., Farinella, G.M., Battiato, S.: Temporal segmentation of egocentric videos to highlight personal locations of interest. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9913, pp. 474–489. Springer, Cham (2016). doi:10.1007/978-3-319-46604-0_34
7. Teraoka, T.: Organization and exploration of heterogeneous personal data collected in daily life. Hum.-Centric Comput. Inf. Sci. **2**(1), 1 (2012)
8. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
9. Templeman, R., Rahman, Z., Crandall, D., Kapadia, A.: Placeraider: virtual theft in physical spaces with smartphones. arXiv preprint arXiv:1209.5982 (2012)
10. Ryoo, M.S., Rothrock, B., Fleming, C.: Privacy-preserving egocentric activity recognition from extreme low resolution. arXiv preprint arXiv:1604.03196 (2016)
11. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, Heidelberg (2011). doi:10.1007/978-1-84882-935-0
12. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks, pp. 1097–1105 (2012)