Straightforward Recognition of Daily Objects in Smart Environments from Wearable Vision Sensor

Javier Medina Quero
Department of Computer Science
University of Jaén
Jaén, Spain
Campus Las Lagunillas, Jaén 23071

Federico Cruciani
School of Computing
Ulster University
Newtownabbey, United Kingdom
Co. Antrim, Northern Ireland BT15 1ED, UK

Lorenzo Seidenari
Department of Information Engineering
University of Florence
Florence, Italy
Via Santa Marta 3, Firenze 50139

Macarena Espinilla
Department of Computer Science
University of Jaén
Jaén, Spain
Campus Las Lagunillas, Jaén. 23071

Chris Nugent
School of Computing
Ulster University
Newtownabbey, United Kingdom
Co. Antrim, Northern Ireland BT15 1ED, UK

***Abstract*—In this work, we propose a method to create and synthesize a new set of virtual images of daily objects within a smart environment partially automating the labeling process. Proposed methods enable the generation of a large dataset from a set of few images using an ad hoc data augmentation, which increases the original dataset size, generating new items through partial modification of available images. The proposed method for data augmentation is accomplished through the following steps: (i) object tracking is proposed to identify and label *static* objects; and (ii) background subtraction is used to select the masked foreground object of *moving* objects, which are virtually projected with geometry transformation over room images used as background. Furthermore, a case study is carried out, where an inhabitant wears a wearable vision sensor in a daily scene. Eight objects are learned using the proposed methodology. Finally, obtained results and successful recognition rates are discussed.**

***Index Terms*—object recognition, data augmentation, smart environments, wearable vision sensors, deep learning, activity recognition.**

I. Introduction

Activity Recognition (AR) defines models able to detect human actions and their goals in smart environments in order to provide assistance. Such methods have increasingly been adopted in smart homes and health-care applications aiming both at improving the quality of care services and providing assistance for instance in emergency situations [1]. AR is an open field of research where approaches based on different types of sensors have been proposed. In the firs stages, binary sensors were proposed as suitable devices for describing daily human activities within a smart environment setting [2]. More recently, wearable devices have been used to analyze activities and gestures in AR [3].

Vision based sensors have also been used as a rich data source for the recognition of human activities. Within indoor-based works, the description of activities through the detection of human joints with a vision sensor has been considered [4]. Within this context, the success of these approaches highlights a potential major role that egocentric view could play in indoor environments [5] [6]. Moreover, the use of wearable vision sensors with first-person point-of-view has been reported as viable in detecting daily object interaction [7].

On the other hand, deep learning approaches have emerged as a powerful method to recognize objects from vision images by means of Convolutional Neural Networks (CNNs) [8], which have generated a new trend of vision models [9]. However, the weak point of deep learning methods resides in the fact that the training in CNN requires a large amount of image data [10].

To palliate this lack, data augmentation provides a straightforward solution to enlarge the number of learning cases from a limited set [11] and therefore, reducing the risk of over-fitting [8]. A similar approach has been proposed in recent works [12], [13], where the selection of images from objects in a small number of human-annotated examples is next projected in the environmental background to provide new synthetic examples. Similar data transformations and operations improve the recognition of objects [14], which have also been related to improve the learning with new synthetic images from a real world in a similar way to a *dream process*. In the context of AR, a large quantity of labeled data is required [15] and videos have often been used to produce the annotation of data portions and specific annotation tools have been developed [16], [17]. Similar tools can facilitate the annotation process alleviating the problem, yet labeling remains a time consuming task. In this paper we propose a straightforward method that allows to easily recognize objects involved with ADLs with minimal user interaction.

In addition, transfer learning has resulted a suitable method to efficiently transfer patterns when recognizing objects from a new scope [18]. Moreover using transfer learning from previous pre-trained networks produces a boost in learning which is suitable to problems with limited training data [18], [19].

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

Taking into account these references and the most recent progress in this topic, in this work, we present a methodology for straightforward recognition of daily objects in smart environments from wearable vision sensors focused on the following points:

- Recognition of object instances. We aim to identify personal and daily objects, such as, *my cupboard* as distinct from other cups or objects in the environments.
- Straightforward data generation and bounding-box labeling [20] has been developed and adapted to static and moving objects in the environment. We propose: i) an automatic object tracking [21] in case of static objects, and ii), a virtual projections of objects for which foreground have previously been extracted [22] in case of moving objects.
- A neural network for object detection is integrated via fine-tuning to increase the results and reduce the required learning time.
- Description of human activities by egocentric first-person point-of-view [6] using a wearable vision sensor [23]. Such sensors have been successfully described as a employed to identify interaction of objects in human activity recognition [7].

In Figure 1, the components of the proposed methodology are illustrated. This methodology aims to facilitate the integration of vision recognition from wearable vision sensor in smart environments through a short-time data collection facilitating labeling of daily objects. The reason behind is evaluating the capabilities of a quick collection of visual data to describe daily scenes while an inhabitant interacts with objects in order to promote the object recognition as information source in other models of AR.

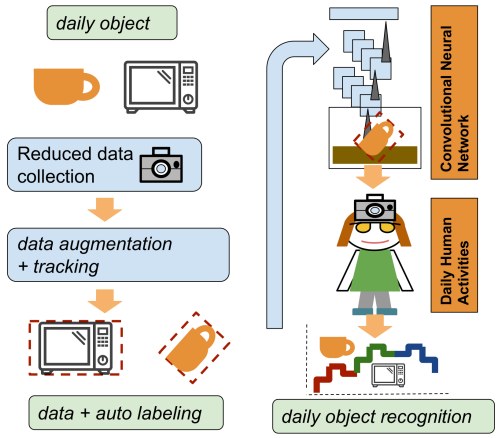

Fig. 1. Components of the methodology. First, a limited number of poses from objects are collected. Second, several techniques of data augmentation and tracking are applied to ease the creation and labeling of dataset. Third, a Convolutional Neural Networks (CNN) learns the images and generated labeling. Fourth, the evaluation of the model is developed by an inhabitant which wears a wearable vision sensor while developing daily activities.

The remainder of the paper is structured as follows: in Section II, the new methodology to easily annotate daily objects in smart environments is proposed; in Section III, a case study of daily living activities in an intelligent environment is presented. Finally, in Section IV, conclusions and future works are pointed out.

II. Methodology

In this work, a methodology to easily recognize daily objects in smart environments from wearable vision sensor is proposed. Specifically, we focus on recognition of objects' instances in an indoor context using a first-person point-of-view.

The following section presents two different approaches for data collection and bounding-box labeling; respectively with reference to *static* and *moving* objects within a smart environment .

A. Collecting and labelling data for static objects

Some objects can be assumed to be static and fixed in a location which does not change over time. Static objects' appearance may only change because of camera pose and lighting variation, also induced by changes of the observer position. This assumption provides a major advantage for labeling purposes. Hence, for static objects we propose the collection of short videos of the object within fixed background from different perspective and distances.

Subsequently, an auto tracking is applied to the instance of the static objects in order to auto detect the bounding box where the object is located in the video. To this end an object tracking approach [25] has been used, where just an initial bounding box selection is required by the user in order to track the object in the complete video sequence. After visual inspection of the results, we concluded that the approach of [25] provided better results than Tracking-Learning-Detection [26] or median flow tracker [27].

This data collection approach provides easy data recording without requiring long time collection, as well as, an automatic labeling of the bounding box using an automatic visual tracker. We note that learning in this static context is not robust to changes in the background or location of the instance object.

Figure 2, depicts the process of collecting and labelling data for static objects.

B. Collecting and labelling data for moving objects

Other objects are usually handled by the inhabitant, thus their location changes in the smart environment. Moreover, the point-of-view of the person while looking at and interacting with them is not fixed and their location in the smart environment changes.

In this case, the proposed method aims to generate thousand of images from just few raw images of the target object. This approach inherently presents further complexity with respect to the case of *static instances*, due to the variability introduced by a dynamic context.

First, few images of the object are taken from different points of view, rotations and orientations. Then, a foreground extraction using iterated graph cuts is applied [22] to obtain the

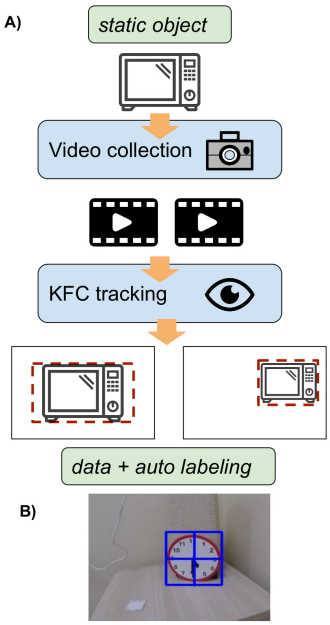

Fig. 2. A) Collection of short-videos from static objects. The KFC tracking provides auto-labeling of the bounding box in the images. B) Example of the image of clock with the bounding box auto-labeling from KFC tracker.

pixels of the objects without the background, which configure a *mask* for each frame with a transparent background (masked foreground objects). This process is straightforward from the user perspective who is simply required to select the bounding box where the object is located. Finally, for each pose an image of the object is obtained depicting only the target object with no background. In Figure 3, example of this process is shown.

Subsequently, a short-time video for each room of the smart environment is collected from a vision sensor, whose frames are used as background. Finally, the masked foreground objects (obtained in the previous step) is superposed over the new background obtained from the video of the room) thus generating a new synthetic set of virtual images. This annotation strategy for moving objects provides also the bounding box labeling which determines both the location and size of the object; this is beneficial in training detectors that work in this specific environments.

In order to locate masked foreground objects in the background of the rooms, a significant data augmentation from original images is developed by means of random transformations: translation, rotation and scale, which have been successfully proposed in visual recognition within smart environments from limited set [28].

- *Translation*. The masked foreground objects and bounding box labels are relocated within a maximal window size $[t_x, t_y]^+$ using a random process which generates a random translation transformation $[t_x, t_y], t_x \in [0, t_x^+], t_y \in [0, t_y^+]$.
- *Rotation*. The rotations of the masked foreground objects and bounding box labels are provided in two steps. First, the translated image is flipped horizontally and vertically, using a random process that applies the transformation to a percentage of cases, defined by wH, wR respectively. Second, a rotation transformation is defined by a maximal rotation angle α^+ which generates a random rotation with an angle $\alpha \in [0, \alpha^+]$.
- *Scale*. A random scale within a maximal angle $s \in [0, s^+]$ is applied to the mask and bounding box labeling.
- *Flipping*. The image is flipped horizontally and vertically, using a random process that applies the transformation in a given percentage of cases, defined by wH, wR respectively,

The final result is a new synthetic set of images, where the moving objects are virtually located, scaled and rotated in the background. An example is shown in detail in Figure 3. Although these images look like *surrealistic*, we note they are really close to the point-of-view of the person when interacting with the objects and these images are used in the learning process.

III. Case Study

The case study was carried out in the smart lab of the CEATIC (Center for Advanced Studies in Information Technology and Communication) of University of Jaen (Spain) [24] [1]. This smart lab integrates a living space with bedroom, kitchen, living-room and toilet. Eight objects within this smart lab were selected to be recognized: 1) the microwave of the kitchen, 2) the bedroom clock, 3) exit door, 4) the toothbrush, 5) the preferred book, 6) tetra brick of milk, 7) the cup, and 8) the mobile device of the inhabitant.

The selection criterion was integrating moving and static objects which were relevant in most popular target ADLs in AR approaches, such as waking up, taking some breakfast and resting in the smart environment. For each static object (room clock, microwave and exit door), 3 short videos (with a duration between 10 and 15 seconds) were collected. For each moving object (toothbrush, cup, book, mobile and tetra), 20 images from different orientations were collected. For each room involved (kitchen, living room and toilet), 3 short videos (with a duration between 10 and 15 seconds) were collected. A wearable vision sensor (GoPro Hero 5) was used to collect the images and videos. In Table I, an image of static and moving objects and the room involved in the case study are illustrated.

Second, for each static object, we applied the adaptive color attributes approach for auto tracking [25] to auto label the bounding box where the object is located in the video, as we described in Section II-A.

Third, for each moving object, we applied a foreground extraction to obtain the masked foreground object using iterated graph cuts [22]. Then, we augmented the data projecting the masked foreground objects in the background of the rooms. 200 images for each masked foreground object were generated with next parameter of random transformations: *translation*) maximal window size $[t_x, t_y]^+ = [50, 50]$, *rotation*) maximal

[1] https://ceatic.ujaen.es/en/smartlabweb

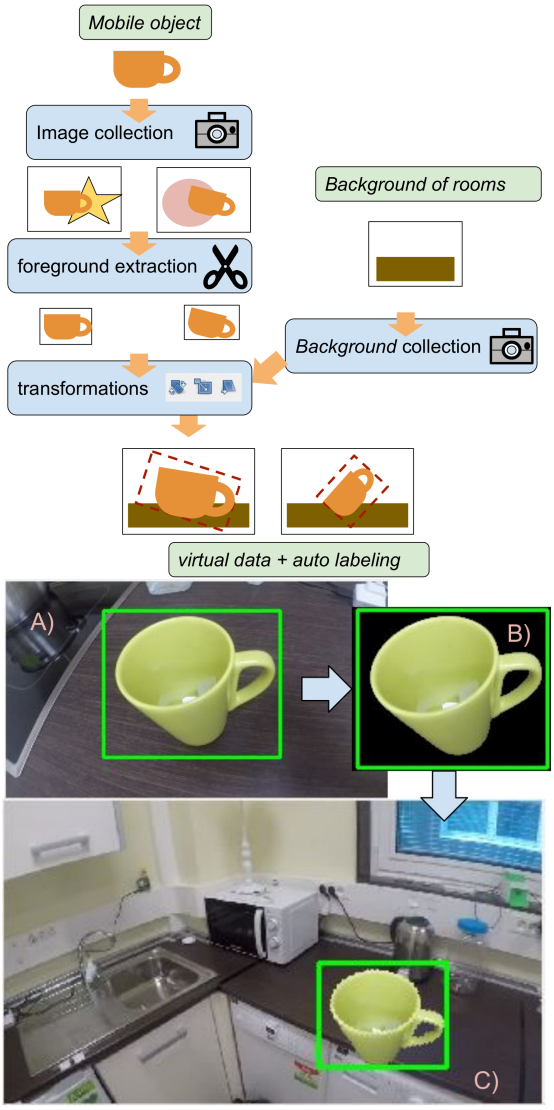

Fig. 3. On the Top) Steps in the collection and data labelling process for moving objects: i) taking of few images of the object, ii) extracting background to generate a masked foreground object, iii) collection of background, and iv) random transformation of the masked foreground object and bounding box of the labeling when projecting over the background. On the bottom) An example of a cup in the smart environment. A) cup with original background, B) masked foreground object as result from foreground extraction, and C) virtual random transformation of the masked foreground object and projection over a background.

rotation angle $\alpha^+ = \pi/4$, *scale*) maximal scale $s^+ = 0.5$ and *flipping*) in the percentage of cases $wH = wR = 0.5$.

Fourth, we have learned the data to detect the objects from an efficient and robust Convolutional Neural Network [29]: a CNN model with inception configuration within layers [30] with a faster learning approach [31]. We have used the Object Detection API of Tensorflow [32] and the CNN of Faster with an Inception backbone pre-trained on MSCOCO. While learning with Tensorflow, the next data augmentation with default parameters were also included: random change of pixel values, contrast and saturation, and random displacement of

TABLE I
OBJECTS AND ROOMS INVOLVED IN THE CASE STUDY. FROM TOP TO BOTTOM AND FROM LEFT TO RIGHT: MICROWAVE, CLOCK, DOOR, TOOTHBRUSH, BOOK, TETRA, CUP, MOBILE DEVICE, KITCHEN, LIVING ROOM AND TOILET.

boxes. The 9 objects where learned for 24 hours in i5 CPU 2.3 GHz without GPU.

Fifth, for the evaluation process, a scene where an inhabitant wears a wearable vision sensor (Go Pro 5) were recorded. In this scene, the inhabitant wakes up from bed and go to toilet to toothbrushing. Then goes to the kitchen to pick up a cup, which is filled by tetra brick of milk. Afterwards, the cup of milk is heated in the microwave, and then, the inhabitant drinks it and goes to sit in the sofa to read a book and review the mobile devices. Finally, inhabitant goes to the main door to exit. In the evaluation process, two similar scenes were collected two times: scene 0 with a duration of 160 seconds, and scene 1 with a duration of 183 seconds. The original images of objects, the augmented data with the auto-labeling generated by the methodology and the images from the wearable vision sensor are available in the next URL[2].

A. Results

In this section, the results of detecting the daily object from images collected by a wearable vision sensor using an egocentric point-of-view are presented.

In Figure 4, the score of the recognition for each object in the time-line for the two scenes is presented. First, we identify that the range of detection in the scores exhibits significant variations between objects. Consequently, some objects present a great difference of scores between them: some with low score and few number of recognition, against others exhibiting high scores and frequent recognition. This is mainly due to the similarity of an object with the environment and or with other objects. For example, microwave and book contain black and white regions which are very common in the environment, so only from scores higher to 0.8 we recognize properly the object. To identify the correct recognitions from the scores, a threshold for each object has been applied. As

[2]http://serezade.ujaen.es:8054/smart-lab-wearable-vision/

two similar samples of the scene were collected, scene 0 was used to identify the thresholds, which are shown in Table II, which minimize the false positives rates of objects. Scene 1 was used to evaluate object recognition with these thresholds and the ground truth labeled by a human observer.

Fig. 4. Top) Raw scores of the object recognition in the time-line of the scene 0 (top) and scene 1 (bottom).

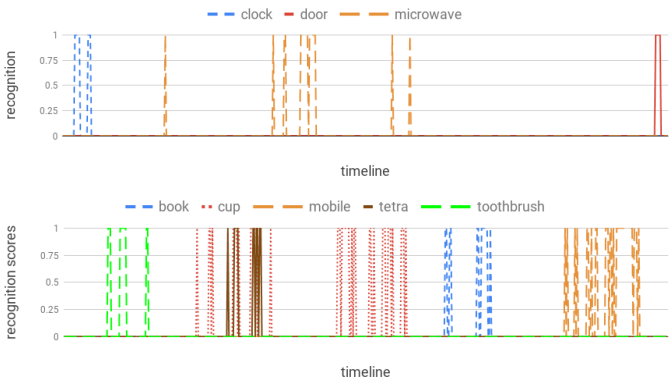

Fig. 5. Recognition of static objects (top) and moving objects (bottom) after applying threshold.

TABLE II
THRESHOLD OF SCORES FOR EACH OBJECT

Microwave(0.7)	Clock(0.9)	Door (0.8)
Toothbrush(0.6)	Book(0.4)	Tetra brick(0.9)
Cup(0.15)	Mobile (0.95)	

In Figure 5, we illustrate the object recognition in the time-line of the scene 1 for each moving and static objects. From this data of object recognition, we have evaluated the capabilities of the recognition on a real scene. First, we present precision and recall of the object detection for each frame and detection. Precision represents the percentage of corrected detection in the frames where the object appears. Recall represents the percentage of real appearance of the object in the frames where the object is identified.

TABLE III
PRECISION AND RECALL OF THE DETECTION FOR EACH OBJECT.

Object	**Precision**	**Recall**
book	1.0	0.11
clock	1.0	0.89
cup	0.86	0.42
door	1.0	0.5
microwave	0.89	0.25
mobile	0.97	0.48
tetra brick	1.0	0.28
toothbrush	1.0	0.37

As shown in the table III, precision is high but recall is notably lower. We note extremely high metrics are not usually presented in object detection from first-person point-of-view. This is caused by the blurry images collected by a wearable vision sensor while movements and the occlusion of objects in real interaction of daily activities. However, it does not represent a disadvantage in detecting and identifying the user interaction with objects. Indeed, once an object is detected, it can be straightforwardly linked with an interaction of the inhabitant and the object. To evaluate it in Table IV, the number of objects detected in each *action*: waking up, breakfast and resting is shown.

TABLE IV
INVOLVED AND RECOGNIZED OBJECTS FOR EACH SCENE. FOR EACH SCENE, FIRST COLUMN + REPRESENTS IF THE OBJECTS WAS INVOLVED, SECOND COLUMN $|N|$ REPRESENTS THE NUMBER OF DETECTION FOR EACH OBJECT.

	waking up		breakfast		resting	
Obj	+	$\|N\|$	+	$\|N\|$	+	$\|N\|$
book		0		0	+	13
clock	+	8		0		0
cup		0	+	37		0
door		0		0	+	5
microwave		0	+	17		0
mobile		0		0	+	33
tetra brick		0	+	11		0
toothbrush	+	11	0			

IV. CONCLUSIONS AND ON GOING WORKS

In this work, a methodology to easily recognize daily objects in smart environments from a wearable vision sensor using a first-person point-of-view has been proposed. The proposed methodology is focused on minimizing the time of data collection and labeling using virtual images from moving objects and automatic tracking from static objects. Main functionalities and advantages introduced with our method have been described in a case study, where eight objects have been identified in a smart lab, while an inhabitant developed the most popular human daily activities in a home. The augmented data with and auto-labeling of bounding boxes generated by the methodology for the CNN have provided an encouraging detection of the objects while the actions of the inhabitants in the scene. However, the recall is notably lower due to the complexity of recognizing from frames collected by a wearable vision sensor, which include blur and fuzzy frames due to movement. It enable a suitable daily object recognition within a sequence of frames, but it is not appropriate for single

image detection. In future works, instance based tracking can be integrated to improve recall to avoid losing object detection because of blur or other rapid changes in imagery [33], [34].

Finally, we note that visual recognition of daily objects provides a straightforward representation of inhabitant's interaction with objects, which can be integrated in multi-sensor activity recognition system in future works.

ACKNOWLEDGMENT

This research has received funding under the REMIND project Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020, under Grant Agreement No. 734355. This contribution has been also supported by the PI-0203-2016 project from the Council of Health for the Andalusian Health Service (Spain) and the postdoctoral research grant Action-6 of the University of Jaen.

REFERENCES

[1] Chen, L., Nugent, C. D., Biswas, J., & Hoey, J. (Eds.). (2011). Activity recognition in pervasive intelligent environments (Vol. 4). Springer Science & Business Media.

[2] Medina-Quero, J., Zhang, S., Nugent, C., & Espinilla, M. (2018). Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition. Expert Systems with Applications, 114, 441-453.

[3] Ordoñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors, 16(1), 115.

[4] Rege, A., Mehra, S., Vann, A., & Luo, Z (2017). Vision-Based Approach to Senior Healthcare: Depth-Based Activity Recognition with Convolutional Neural Networks.

[5] Bettadapura, V., Essa, I., & Pantofaru, C. (2015, January). Egocentric field-of-view localization using first-person point-of-view devices. In Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on (pp. 626-633). IEEE.

[6] Betancourt, A., Morerio, P., Regazzoni, C. S., & Rauterberg, M. (2015). The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 25(5), 744-760.

[7] Shewell, C., MedinaQuero, J., Espinilla, M., Nugent, C., Donnelly, M., & Wang, H. (2017). Comparison of fiducial marker detection and object interaction in activities of daily living utilising a wearable vision sensor. International Journal of Communication Systems, 30(5), e3223.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[9] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[10] Yamashita, T.; Watasue, T.; Yamauchi, Y.; Fujiyoshi, H. Improving Quality of Training Samples Through Exhaustless Generation and Effective Selection for Deep Convolutional Neural Networks. *VISAPP* **2015**, *2*, 228–235.

[11] Ciresan, D.C.; Meier, U.; Masci, J.; Gambardella, L.M.; Schmidhuber, J. High-Performance Neural Networks for Visual Object Classification. *arXiv* **2011**, arXiv:1102.0183.

[12] Georgakis, G., Mousavian, A., Berg, A. C., & Kosecka, J. (2017). Synthesizing training data for object detection in indoor scenes. arXiv preprint arXiv:1702.07836.

[13] Dwibedi, D., Misra, I., & Hebert, M. (2017, October). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In The IEEE international conference on computer vision (ICCV).

[14] Khoreva, A., Benenson, R., Ilg, E., Brox, T., & Schiele, B. (2017). Lucid Data Dreaming for Multiple Object Tracking. arXiv preprint arXiv:1703.09554.

[15] Tonkin E, Burrows A, Woznowski P, Laskowski P, Yordanova K, Twomey N, Craddock I. Talk, Text, Tag? Understanding Self-Annotation of Smart Home Data from a Users Perspective. Sensors. 2018 Jul 20;18(7):2365.

[16] Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S., & Noldus, L. P. (2009). The Observer XT: A tool for the integration and synchronization of multimodal signals. Behavior research methods, 41(3), 731-735.

[17] Cruciani F, Donnelly MP, Nugent CD, Parente G, Paggetti C, Burns W. DANTE: a video based annotation tool for smart environments. InInternational Conference on Sensor Systems and Software 2010 Dec 13 (pp. 179-188). Springer, Berlin, Heidelberg.

[18] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In Advances in neural information processing systems (pp. 3320-3328).

[19] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1717-1724).

[20] Uricchio, T., Ballan, L., Seidenari, L., & Del Bimbo, A. (2017). Automatic image annotation via label transfer in the semantic space. Pattern Recognition, 71, 144-157.

[21] Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2012, October). Exploiting the circulant structure of tracking-by-detection with kernels. In European conference on computer vision (pp. 702-715). Springer, Berlin, Heidelberg.

[22] Rother, C., Kolmogorov, V., & Blake, A. (2004, August). Grabcut: Interactive foreground extraction using iterated graph cuts. In ACM transactions on graphics (TOG) (Vol. 23, No. 3, pp. 309-314). ACM.

[23] Sumi, K., Sugimoto, A., Matsuyama, T., Toda, M., & Tsukizawa, S. (2004, March). Active wearable vision sensor: recognition of human activities and environments. In Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004. International Conference on (pp. 15-22). IEEE.

[24] M. Espinilla, L. Martinez, J. Medina, C. Nugent. (2018). The Experience of Developing the UJAmI Smart Lab. IEEE Access, vol. 6. pp. 34631-34642.

[25] Danelljan, M., Shahbaz Khan, F., Felsberg, M., & Van de Weijer, J. (2014). Adaptive color attributes for real-time visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1090-1097).

[26] Kalal, Z., Mikolajczyk, K., & Matas, J. (2012). Tracking-learning-detection. IEEE transactions on pattern analysis and machine intelligence, 34(7), 1409.

[27] Kalal, Z., Mikolajczyk, K., & Matas, J. (2010, August). Forward-backward error: Automatic detection of tracking failures. In Pattern recognition (ICPR), 2010 20th international conference on (pp. 2756-2759). IEEE.

[28] Medina Quero, J., Burns, M., Razzaq, M., Nugent, C., & Espinilla, M. (2018, October). Detection of Falls from Non-Invasive Thermal Vision Sensors Using Convolutional Neural Networks. In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 2, No. 19, p. 1236).

[29] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[30] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI (Vol. 4, p. 12).

[31] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[32] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016, November). Tensorflow: a system for large-scale machine learning. In OSDI (Vol. 16, pp. 265-283).

[33] Seidenari, L., Baecchi, C., Uricchio, T., Ferracani, A., Bertini, M., & Del Bimbo, A. (2018, January). Object Recognition and Tracking for Smart Audio Guides. In Italian Research Conference on Digital Libraries (pp. 163-168). Springer, Cham.

[34] Taverriti, G., Lombini, S., Seidenari, L., Bertini, M., & Del Bimbo, A. (2016, October). Real-time Wearable Computer Vision System for Improved Museum Experience. In Proceedings of the 2016 ACM on Multimedia Conference (pp. 703-704). ACM.