Convolutional Neural Networks for Detecting and Mapping Crowds in First Person Vision Applications

Juan Sebastian Olier[1,2(✉)], Carlo Regazzoni[1], Lucio Marcenaro[1], and Matthias Rauterberg[2]

[1] Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture (DITEN), University of Genoa, Genoa, Italy
sebastian.olier@ginevra.dibe.unige.it
[2] Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands

Abstract. There has been an increasing interest on the analysis of First Person Videos in the last few years due to the spread of low-cost wearable devices. Nevertheless, the understanding of the environment surrounding the wearer is a difficult task with many elements involved. In this work, a method for detecting and mapping the presence of people and crowds around the wearer is presented. Features extracted at the crowd level are used for building a robust representation that can handle the variations and occlusion of people's visual characteristics inside a crowd. To this aim, convolutional neural networks have been exploited. Results demonstrate that this approach achieves a high accuracy on the recognition of crowds, as well as the possibility of a general interpretation of the context trough the classification of characteristics of the segmented background.

Keywords: Convolutional neural networks · Crowds detection · First-person vision · Egocentric videos

1 Introduction

The increasing development of wearable devices, and in particular of wearable cameras at a low price, gives rise to unexplored scenarios where many new applications can be developed, along with which many new challenges are posed. In particular, first person vision (FPV) greatly enhances the possibilities of understanding the surroundings of the wearer, but equally implies many problems to be faced in image processing.

When it comes to the interaction with and of people around the wearer of an FPV device (e.g. smart glasses), the perspective from which the video is taken offers a different understanding of the behaviors, since wearers can move and interacts more naturally, and thus produce data from this more fluent and versatile point of view. Therefore, such capability can be used to develop a context

I. Rojas et al. (Eds.): IWANN 2015, Part I, LNCS 9094, pp. 475–485, 2015.
DOI: 10.1007/978-3-319-19258-1_39

awareness and understanding processing tool, capable of extracting meaningful analytics with respect to changing environments (in particular crowds) and their dynamic evolution. Such tools could be applied in fields like group and social dynamics understanding, affective computing, surveillance, or assistive technologies. Particularly, in crowded environments this would allow innovative ways of understanding activities in specific areas of interest, as a source of new measures for crowd monitoring and as complement to standard surveillance cameras.

In general, any application based on the ideas mentioned implies the segmentation of crowds and background in the FPV video. Thus, in this paper, a method for the detection of crowds from FPV data is proposed. The estimation of people's positions is addressed as an image classification task with video features extracted directly at the crowd level. The aim of this approach is to provide a processing method to be used for modeling interactions inside the crowds, as well as between them and the wearer of the FPV device.

However, the detection of crowds from this kind of videos poses new ambitious challenges due to the high amount of ego-motion, as well as to the variations in the background and environment that might happen in short periods [5]. Moreover, the irregular features of crowds represent yet another challenge, since many possible occlusions can occur and no particular part of the body, nor any specific visual characteristic, can be expected to be prominent or always present when estimating the presence of a person or a multitude.

Thus, a requirement of the method to be used is that it should be capable of extracting features directly from sample data, and to robustly handle the kind of variations described for the problem at hand. These requirements make Convolutional Neural Networks (CNN) an interesting option, for their capability of learning relevant features from data, and the outstanding performance that has been achieved with them in recent years for image classification tasks as shown in [13] and [6].

Furthermore, the method presented here represents a contribution in itself as, to the best of our knowledge, the problem of mapping crowds around a wearer, from a first person vision perspective, has not been yet addressed.

The remainder of the article is structured as follows; section 2 briefly reviews the state of the art on FPV research. In section 3 the statement of the problem, the architecture and training data are described. In section 4 the results found are described and finally in section 5 conclusions and future work are discussed.

2 Previous Work

The research interest in FPV has been increasing in the last few years along with the development of new technologies and the search for applications as described in [1]. One of the main areas investigated so far is activity recognition [9], where the focus is the assessment of interactions between the wearer and the environment, paying particular attention on the manipulation of objects and hands movements as in [3] and [14].

Similarly, the understanding of situations can be related to the interpretation of interactions with people and the way these evolve dynamically, which in turn

can be used to segment video sequences into relevant periods. However, most of the contributions on scene understanding and segmentation of videos are based on similarities on the actions, situations or environments, with interesting results like the ones found in [8] [12] and [11]. Non has directly focused on the presence and interactions of people as the basic segmentation criteria, being a task which may require a tool for extracting and mapping people like the one proposed in the present work.

Nonetheless, in relation to social interactions, some research has been carried out in works like [4], where the gaze direction is used to infer the attention of the people surrounding the wearer in order to recognize communications and social relations. To this aim face detection is used to detect people, what is done because of the particular goals, but limits the problem to particular kinds of interactions and to the presence of faces which generally represents a small subset of all the possible cases in which a person can appear in a scene.

Equally, in pedestrian's detection, a relevant topic deeply explored in the last decades, most of the contributions have focused on fixed cameras, normally above the crowd, and with the aim of analyzing its dynamics. Likewise, relevant works have recently presented important increases in detection rates by using deep architectures, outperforming most of or all previous methods as in [2] and [10], and even from FPV in [15].

Nevertheless, very limited efforts have been devoted towards the estimation of position of people surrounding the wearer in complex environments and from any perspective, being that all of these approaches emphasize on the detection of persons, searching for the complete shape of the bodies and not directly on the crowds and variable visible parts. That implies a training data focused on people and specific bodily characteristics, parts or poses. On the contrary, the present work focuses on the detection of people and crowds in general without assumptions on particular visible characteristics.

3 Approach

This section describes the general approach through a statement of the general problem, then the selection of the specific method to be used, the data used to train the system, and finally the architecture used.

3.1 General Description of the Problem

Scene understanding from FPV videos is the main goal, with a particular focus on describing the environment and the states of people around a wearer as the main actors in crowded environments. The aim is to provide context awareness and understanding processing tool, with which it would be possible to extract meaningful analytics of changing crowded environments and their dynamic evolution. Additionally, given that the interpretation of the context is not only dependent on the crowd but is also related to the characteristics of the environment, in the extracted background it is possible to find some information about

situations in which the wearer is immersed, which could be, at least partially, analyzed through the classification of certain features of the background.

This way, the main task is to segment the image into crowd and background. However, due to the complex visual appearance of natural scenes and the abrupt variations of background caused by the ego-motion, the detection of people from first person videos represents new challenges, many of which are not addressable through approaches normally adopted for pedestrian recognition or crowd analysis.

In particular, as the situation is highly variable, the position, orientation and movements of the people, as well as the crowd density is not predictable, and then it is not possible to make an assumption about detecting particular parts of the body, or any specific visual characteristic. For instance, there is no guarantee to find body parts in all situations, simply because big parts of the body can be out of the field of view when the person is close to the camera, and probably occluded by other people or objects when far. Consequently, the approach of the present work focuses on the detection of crowds and the extraction of features directly from them.

Finally, the desired result is expected to contain a spatial representation of the crowd's distribution, which might allow mapping and analyzing the situation surrounding the wearer as immersed in a particular scenario. Thus, the method developed cannot only be focused on the detection of the crowds, but must allow the approximation of their position in a 3D-like way, implying the estimation of a plausible distance from the camera.

3.2 Detecting Crowds

The task of estimating the presence of crowds in a particular part of the image, as well as to obtain features that allow to perform classification, are the main goals to have in mind when selecting a method to face the problem at hand. However, many methods that could be suitable for this task may assume some previous knowledge of the features, the presence of some characteristics in all the images, or in some cases the capability of extracting the background, which as described above is not feasible for FPV videos.

Then, the crowd's classification problem implies features that can be extracted only from images containing people and crowds, and on features related to the environment in the case of the background. This statement has led to the selection of Convolutional Neural Networks (CNN) as the main method to address the proposed goal; mainly due to the outstanding performance that has been shown for the CNNs in object recognition tasks and feature extraction from sample data [13][7][6]. Particularly in [13], CNNs are compared to state of the art methods for image recognition, showing how deep architectures outperforms them in most of the datasets used. Equally, in [7] the detection of pedestrians using an architecture based on deep CNNs, in general outperforms or performs comparably to the state of the art methods in the most challenging pedestrian detection data sets.

Additionally, the separation between the crowd and the background resulting of the segmentation, leaves the opportunity to further analyze the information

in the image in order to infer more characteristics of the situation and the environment. Thus, the possibility of extracting such kinds of information from the background are taken into account when designing the network and the dataset to train it.

Design of the Training Data. In order to train a CNN diverse approaches can be explored, from whole frames containing different amount of crowds, aiming to classify them by the density found in them, to the estimation of crowds inside small patches from the image. Obviously, the first option turns out to be impractical in many cases, and as the main features to learn are the ones specific to the crowds, using small patches of the images containing them becomes the best option. Extracting patches is beneficial as well when building the data set to train the network, for it makes it possible to easily include variability as a single image containing a crowd may be sampled in such a way that many patches with different variations can be created.

Patches from crowds and background are used to train a CNN, nonetheless, the variability of the features contained in the background is much higher than the ones in people and multitudes, and in some cases even similar to the ones in crowds. Thus the number of background samples must be higher than the one of crowds in order to achieve a good performance.

The set of crowd images has been split into three classes based on the distance of the crowd to the camera. Hence, the classes "Crowds", "Mid-distance crowds" and "Far crowds" are created. This approach can reduce the variability in the features, but also allows estimating the position of the crowd in three dimensions.

The definition of these three classes is done by assuming a 4 to 1 ratio as a regular proportion of the height of a person and the width of the torso. Assuming this ratio, and considering the closeness to the people that a wearer can have when immersed in a crowd; the first distance is assumed to include at most the forth part of the height of a person, and therefore the whole width of the torso. Subsequently, the next distance doubles the previous one, meaning half of the height. Then the third, doubling again, includes the whole height. In this way, the distance to the camera follows a linear function of the portion of height included, where each one is the double of the previous, and the basis depends on the angular view of the camera and the size of the patches.

Having a higher number of samples for background also allows to create different classes for it. Such classes can be selected in different ways, but finding the best division is beyond the scope of this paper; the dataset for background has been split into six different classes (Plants, Interior, Sky, Buildings, Exterior and Floor). With such division, it is possible to estimate other kind of information present in natural scenes representing basic cues of the context and the environment, and could be used even for a better estimation of possible activities.

Consequently, a set of images to train a CNN with these classes has been created based on FPV videos dataset presented in [4] and other images from different publicly available sources. The set is balanced in the number of images per class and contains a total of 57.600, out of which a 15% is used for validation.

Fig. 1. Samples of the three classes of crowds used to train the CNN. a) "Crowds", b) "Mid-distance crowds", and c) "Far crowds"

Fig. 2. Samples of the six classes of background used to train the CNN. Plants, Interior, Sky, Buildings, Exterior and Floor

Architecture of the CNN. The network used has 3 convolutional layers and 2 fully connected ones, and inputs used are 50 by 50 pixels RGB images (the mentioned patches). All convolutional layers are followed by Rectified Linear Units ('ReLU'), and the first two are followed by a 2 by 2 max pooling. The first convolutional layer has 27 filter kernels of size 8x8x3; the second has 90 kernels of size 7x7x27 and the last has 180, 5x5x90 filter kernels. After the third convolutional layer and the first fully connected one a dropout is performed during training with a 50% rate. The first fully connected layer has 420 units, and the last 9, corresponding to the number of classes. Finally, the training for classification is performed with a 9-way softmax.

The architecture and used regularizations for training has been selected through experimentation but also following the results for improving performance of CNNs described in [16].

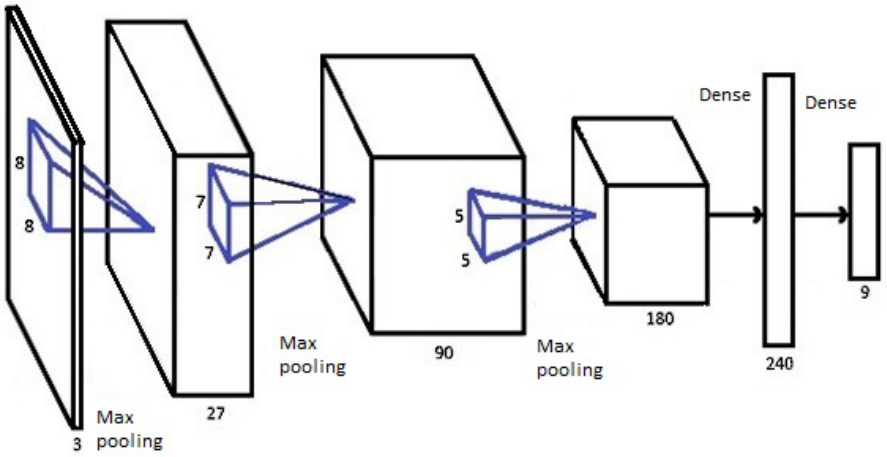

Fig. 3. Illustration of the architecture used for the CNN

Table 1. Confusion matrix of the network for the nine classes tested on the validation set

	Build-ings	Sky	Plants	Interior	Exterior	Floor	Far C.	Mid-dist.	Crowd
Buildings	**0,70**	0,01	0,02	0,14	0,07	0,02	0,00	0,02	0,03
Sky	0,00	**0,96**	0,01	0,01	0,00	0,01	0,00	0,00	0,01
Plants	0,02	0,02	**0,72**	0,00	0,08	0,08	0,02	0,02	0,04
Interior	0,03	0,02	0,00	**0,91**	0,00	0,02	0,00	0,01	0,00
Exterior	0,20	0,01	0,04	0,04	**0,49**	0,06	0,05	0,02	0,09
Floor	0,03	0,05	0,02	0,04	0,02	**0,83**	0,00	0,00	0,01
Far C.	0,00	0,00	0,01	0,00	0,01	0,00	**0,85**	0,12	0,01
Mid-dist.	0,01	0,00	0,00	0,01	0,01	0,00	0,12	**0,77**	0,09
Crowd	0,02	0,00	0,00	0,00	0,00	0,00	0,01	0,18	**0,79**

The network has been trained by using the MatConvNet implementation [17].

In this way, for classification a sliding window extracts patches from the frame and uses the classification delivered by the network to determine the presence of one of the three class of crowd in different areas of the image.

4 Results

With the selected architecture and training on the dataset for 200 iterations, the results indicate that the problem of segmenting crowds in FPV videos can definitely be addressed as an image classification task with features learnt at the crowd level. The classification of the three possible crowd's classes with different distances from the camera shows a good performance and allows the

Fig. 4. Examples of crowd detection, the three classes are represented with different colors, from Red for crowds close to the camera, "Crowd" class, to "Far crowd" in blue. In a, the original frames and in b the respective visualization of the crowd's detection

Table 2. Confusion matrix for the classification of crowds and background in general

	Background	Crowd
Background	0.95	0.05
Crowd	0.03	0.97

mapping of crowds around the wearer just as intended. In Fig. 4 are visualized the classifications of two particular situations form a video in the data set presented in [4]. In these examples the classification of the three possible distances from the camera are illustrated.

Table 1 shows the confusion matrix of the classification performed by the selected CNN. It can be seen that the three classes for crowds have a classification rate of at least 77%, yet most part of the error in these classes are misclassifications between themselves. Thus, if the accuracy is measured as the classification of crowds and background in general, it goes up to 97% as shown in Table 2.

Even if the classification rate is quite high, the misclassification errors can generate noise, which would accumulate when data are used for tracking or mapping of the surroundings. This problem is beyond the scope of this work but could certainly be addressed by the correlation of data in a temporal framework, taking into account the information of consecutive frames in the video, and probably additional information form the ego-motion.

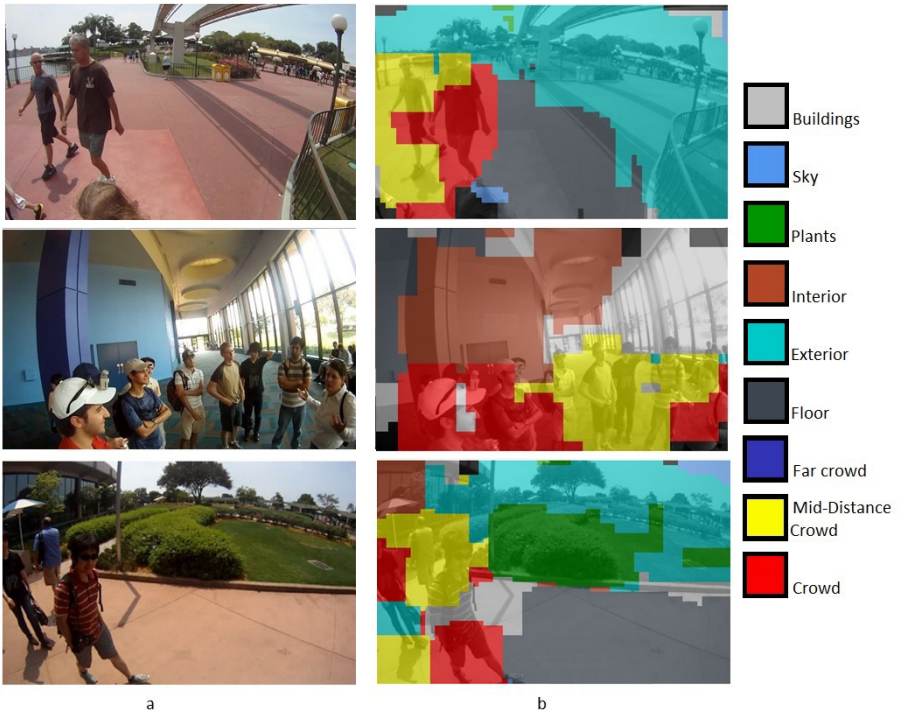

Fig. 5. Examples of segmentation of images based on the classification performed for background and crowds

Additionally to the detection of crowds, it is shown in Fig. 5 that as the background has been grouped in separated classes, though not with the same accuracy than for crowds in all of them, such division can be used to segment images as a basis for understanding context. It can be appreciated in the illustrations of Fig. 5 that, for example, it might be possible to infer the difference between a scenario inside a building where the classes "Interior" and "Buildings" are predominant on an area of the image, while in the scenes outdoors the class "Exterior" is predominant, and for example some areas where plants are present appear clearly segmented giving more hints about the context.

5 Discussion and Future Work

This work discussed how detection of people and crowds from first-person view can be achieved with good accuracy by learning features at the crowd level for classification through Convolutional neural networks.

It has been also shown that this kind of approach could be extended to wider segmentation of the image in order to generate relevant information for context awareness or other applications. This approach could be the basis for

more abstract processing that deals with interaction understanding or dynamic adjustment of the system depending on the environment.

Finally, once the detection of the crowds has been performed by using the three defined classes, it could be possible to map the crowds by projecting the angle and depth, assuming the angle as a linear function of the horizontal position of the crowd and the depth depending on the class it belongs to. This can be the next step to bring the temporal information of a video to the stabilization of crowd's detection and the extraction of information about interactions of the wearer with the surrounding people.

References

1. Betancourt, A., Morerio, P., Regazzoni, C., Rauterberg, M.: The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology **PP**(99), 1–1 (2015)
2. Bourdev, L., Yang, F., Fergus, R.: Deep poselets for human detection (2014). arXiv preprint arXiv:1407.0717
3. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 407–414. IEEE (2011)
4. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: a first-person perspective. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1226–1233. IEEE (2012)
5. Kanade, T., Hebert, M.: First-person vision. Proceedings of the IEEE **100**(8), 2442–2453 (2012)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
7. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 253–256, May 2010
8. Lee, Y.J., Grauman, K.: Predicting important objects for egocentric video summarization. International Journal of Computer Vision, 1–18 (2014)
9. Narayan, S., Kankanhalli, M.S., Ramakrishnan, K.R.: Action and interaction recognition in first-person videos. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 526–532. IEEE (2014)
10. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2056–2063. IEEE (2013)
11. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2847–2854. IEEE (2012)
12. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2537–2544. IEEE (2014)

13. Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 512–519, June 2014
14. Ryoo, M.S., Matthies, L.: First-person activity recognition: what are they doing to me? In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2730–2737. IEEE (2013)
15. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3626–3633. IEEE (2013)
16. Smirnov, E.A., Timoshenko, D.M., Andrianov, S.N.: Comparison of regularization methods for imagenet classification with deep convolutional neural networks. AASRI Procedia **6**, 89–94 (2014)
17. Vedaldi, A., Lenc, K.: Matconvnet-convolutional neural networks for matlab (2014). arXiv preprint arXiv:1412.4564