Visitors Localization in Natural Sites Exploiting EgoVision and GPS

Filippo L. M. Milotta[1,*], Antonino Furnari[1,*], Sebastiano Battiato[1],
Maria De Salvo[2], Giovanni Signorello[2] and Giovanni M. Farinella[1,2]
[1] *University of Catania, Department of Mathematics and Computer Science, Via Santa Sofia - 64, Catania 95125, Italy*
[2] *University of Catania, CUTGANA, Viale A. Doria 6, Catania 95123, Italy*

Keywords: Egocentric (First Person) Vision, Localization, GPS, Multimodal Data Fusion.

Abstract: Localization in outdoor contexts such as parks and natural reserves can be used to augment the visitors' experience and to provide the site manager with valid analytics to improve the fruition of the site. In this paper, we address the problem of visitors localization in natural sites by exploiting both egocentric vision and GPS data. To this aim, we gathered a dataset of first person videos in the Botanical Garden of the University of Catania. Along with the videos, we also acquired GPS coordinates. The data have been acquired by 12 different users, each walking all around the garden for an average of 30 minutes (i.e., a total of about 6 hours of recording). Using the collected dataset, we show that localizing visitors based solely on GPS data is not sufficient to understand the location of the visitors in a natural site. We hence investigate how to exploit visual data to perform localization by casting the problem as the one of classifying images among the different contexts of the natural site. Our investigation highlights that visual information can be leveraged to achieve better localization and that Egocentric Vision and GPS can be exploited jointly to improve accuracy.

1 INTRODUCTION

Localizing visitors in natural sites can be useful in many ways. The information about the position of a visitor can be used to augment the tour experience (e.g., describing plants that can be observed at a specific location). Also, the collected information can provide the site manager with valid data useful to understand the visitors' behaviour and to improve the services of the site. Last, but not least, localization may be useful also for safety reasons, being a valuable technology to retrieve the position of the users in wide natural outdoor environments, where it may be easy for a visitor to get lost.

Despite the fact that GPS is a popular technology to perform localization outdoor, we found that GPS information is not suitable for supporting the localization of the visitors of a natural site in a reliable way. This is due to many factors which limit the localization accuracy, such as trees covering the sky and the occasional presence of indoor spaces, like, for instance, green-houses. An alternative technology to localize the visitors of a natural site is provided by image based localization. In particular, egocentric vision offers a convenient setting to collect visual information of the visits which can be used to aid localization and, more in general, to understand the behavior of the visitors to answer questions like "What are the most common paths chosen by visitors?", "What are the most viewed points of interests?", or "Are there important points of interest that, for some reason, are not viewed by the visitors?".

In this paper, we investigate the use of egocentric vision as an aid for localization purposes. Visitors are supposed to wear smart glasses able to acquire and process videos from their point of view. The collected visual information is employed to perform automatic localization of the users in the natural site.

Our study compares solutions based on vision and GPS to localize the visitors of a natural site and explores different modalities that can be combined to improve the performance of a localization system. As proposed in previous work (Starner et al., 1998; Weyand et al., 2016; Ragusa et al., 2019), we address localization as a classification task. In this setting, the area of the natural site is divided into cells representing meaningful environments (e.g., "main entrance", "sicilian garden", etc.). A classifier is then trained to recognize the correct class given measurements form

[*] These authors contributed equally to this work.

Milotta, F., Furnari, A., Battiato, S., De Salvo, M., Signorello, G. and Farinella, G.
Visitors Localization in Natural Sites Exploiting EgoVision and GPS.
DOI: 10.5220/0007484205560563
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 556-563
ISBN: 978-989-758-354-4

GPS and/or visual signals. Our goal is to design a system for the localization of visitors in a natural site which can run in embedded settings such as on wearable and mobile devices. To this purpose, we compare the considered methods taking into account localization accuracy, computational time, as well as the amount of memory required by the system to perform localization. To support the experiments, we acquired a dataset of egocentric videos in the Botanical Garden of the University of Catania[1]. The garden covers an area of about 170×130 m^2. The first person videos of the dataset have been acquired using a Pupil 3D Eye Tracker[2]. During the acquisition, a smartphone has been employed to record the GPS location of the visitor. Video and GPS measurements have been synced in order to attach each video frame a specific set of GPS coordinates. The area of the considered natural site has been divided into 9 contexts relevant for the visitors, in accordance with the experts and the manager of the site. Each frame of the dataset has been labeled to indicate the context in which the frame was captured.

Experiments show that performing localization based solely on GPS information is not sufficient to achieve reasonable performance on the considered dataset, whereas localization based on visual data achieves better accuracy, albeit at a higher computational cost. We also show that an improvement can be obtained by exploiting vision and GPS jointly at a slightly higher computational cost. Specifically, the performed analysis points out that a vision based approach requiring just 0.1MB allows to improve the accuracy of localization systems based on GPS by 7.60% at a negligible computational time (i.e., 4.71E–3 seconds for image on CPU).

The remainder of the paper is structured as follows. An overview of the related works is given in Section 2. The collected dataset is described in Section 3. The proposed method and the experimental results are detailed in Sections 4 and 5 respectively. Section 6 concludes the paper with final remarks and insights for future research.

2 RELATED WORK

In this section we briefly review previous work related to four lines of research relevant to this paper: (i) computer vision in natural environments, (ii) localization based on wireless and BLE devices, (iii) image-based localization, and (iv) localization based on both images and GPS.

Computer Vision in Natural Environments. In Kumar et al. (Kumar et al., 2012), a computer vision system named *Leafsnap* has been proposed to recognize leaves and to identify the species of 184 trees of the North-eastern United States. This system is integrated in a mobile application which allows users to take pictures of leaves placed on a white sheet in order to segment them and remove stems. The silhouettes of the leaves are represented through histograms of curvature over different scales. Leaves are then identified from their representation with a Nearest Neighbors approach considering an intersection distance measure.

Wagner et al. (Wegner et al., 2016) designed a framework to recognize trees in a urban context. The framework has been specifically developed to match aerial images of trees from Google maps with respect to street view images. In this way, trees are assigned positions on public street sides. The authors also released the “Pasadena Urban Trees” dataset containing more than 100,000 images related to 80,000 trees tagged with species and locations.

Van et al. (Van Horn et al., 2017) gathered a dataset called “iNat2017” by employing iNaturalist expert network[3], which allows naturalists to map and share photographic observations of biodiversity across the world. The dataset contains images of more than 8,000 species acquired in natural places. The species are characterized by high visual variability, high similarity among species and a large number of imbalanced and fine-grained categories. A challenge on this dataset has been proposed by the authors to encourage research on the field.

Joly et al. (Joly et al., 2017) collected the “LifeCLEF” dataset and proposed a challenge on natural species classification. The dataset is proposed together with four challenges: audio based bird identification, image-based plant identification, visual-based sea-related organisms monitoring and location-based species recommendation.

The studies discussed above have proposed datasets and algorithms based on computer vision to address specific issues in the domain of natural sites. While most of these works addressed tasks related to the classification of plants, in this work we consider the problem of localizing the visitors of a natural site using GPS and visual data.

[1] Botanical Garden of the University of Catania: http://www.ortobotanicoitalia.it/sicilia/catania/ (4-Nov-2018)

[2] Pupil 3D Eye Tracker Website: https://pupil-labs.com/pupil/ (4-Nov-2018)

[3] iNaturalist Website: https://www.inaturalist.org/ (4-Nov-2018)

Localization based on Wireless and BLE Devices. Localization can be performed employing several devices and signals, such as antennas, RGB cameras, mobile wireless devices (Alahi et al., 2015), and bluetooth low energy (BLE) (Ishihara et al., 2017b; Ishihara et al., 2017a).

Alahi et al. (Alahi et al., 2015) developed a method to improve human localization based on GPS employing a set of fixed antennas coupled with fixed RGB cameras and mobile wireless devices (i.e., smartphones/beacons). The authors used a multimodal approach in which visual information (RGB) is considered jointly with wireless signals (W) obtaining the so called RGB-W data. Signal trilateration and propagation models are at the core of this wireless-based approach. These signals are used jointly with tracking methods in the RGB domain to localize users in indoor environments.

Ishihara et al. (Ishihara et al., 2017b) have shown how the user localization can be performed through a beacon-guided approach, instrumenting the environment with bluetooth low energy (BLE) signals emitters. The authors designed a method in which radio-wave-based localization is combined with Structure from Motion (SfM) starting from visual input. However, as stated by the authors, SfM is still a challenging task in a real world context (particularly in natural outdoor scenarios) as it does not perform well in environments with little to no distinctive visual features or when there is a large amount of repetitive features, like in a natural site (i.e., a garden with a lot of plants as in our case). An improvement of the approach as been proposed in (Ishihara et al., 2017a) where inference machines have been trained on previously collected pedestrian paths to perform user localization. In this way, the authors managed to reduce localization and orientation error with respect to their previous method.

While the exploitation of Wireless and BLE devices is convenient in indoor settings, this is not generally the case in large outdoor and natural environments. The main problems due to the lack of existing infrastructures (e.g., WiFi) and due to the difficulties arising from the installation of specific hardware in such settings. Therefore, in this paper, we consider the exploitation of visual and GPS signal, which do not require the installation of specific hardware in the site.

Image based Localization. In this paper we address localization of the visitors as a classification problem, where each class represents a context of a large outdoor natural place. This approach has been already considered by other authors, as briefly summarized in the following.

Furnari et al. (Furnari et al., 2017) considered the problem of recognizing personal locations specified by the user from egocentric videos. The segmentation problem is addressed as an “open-set” classification problem where the classes specified by the user have to be identified and the other environments, which are unseen during training, need to be rejected.

Body-mounted video cameras have been employed by Starner et al. (Starner et al., 1998) to localize users from first person images. Localization is in this case considered at the room level in a “close-set” scenario in which the users can move in a limited set of environments.

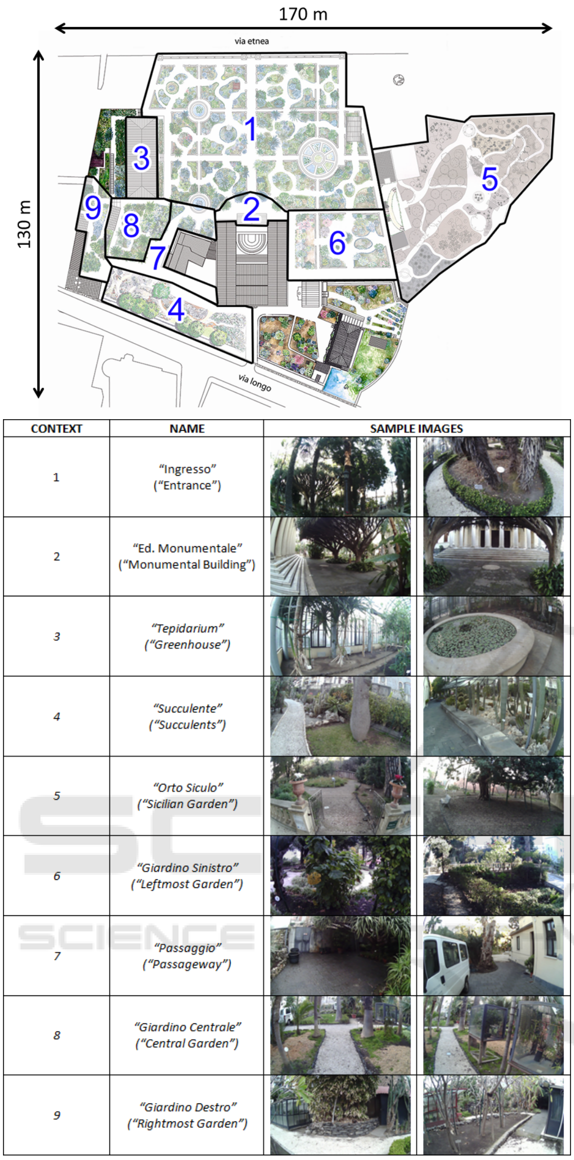
Santarcangelo et al. (Santarcangelo et al., 2016) have investigated the use of multimodal signals collected from shopping carts to localize customers in a retail store. The inferred location information is then exploited fo infer the behavior of customers for marketing purposes in a “Visual Market Basket Analysis” scenario.

Ragusa et al. (Ragusa et al., 2018; Ragusa et al., 2019) considered the problem of localizing the visitors of a cultural site from egocentric video. In the considered settings, museum curators and site managers could take advantage from the inferred information to improve the arrangement of their collections and increase the interest of their audience. The system has been extended to automatically produce summaries of the tours to be sent to the visitors of the cultural site as digital memory.

Classification based localization has been studied by Weyand et al. (Weyand et al., 2016). Specifically, the authors presented *PlaNet*, a deep network able to localize images of places through different cues such as landmarks, weather patterns, vegetation, road markings, or architectural details.

Similarly to the aforementioned works, we tackle localization as a classification problem, dividing the space of interest into areas. Differently from the above approaches, we explore the combination of GPS and visual input to achieve better accuracy at a low computational cost.

Joint Exploitation of Images and GPS for Localization. Previous works investigated the combination of GPS and vision to localize users in an environment. Capi et al. (Capi et al., 2014) presented an assistive robotic system to guide visually impaired people in urban environments. Electronic navigation aid is achieved through a multimodal approach: data from GPS, compass, laser range finders, and visual information are merged together and used for training neural networks. The assistive robotic system



CONTEXT	NAME	SAMPLE IMAGES
1	"Ingresso" ("Entrance")	
2	"Ed. Monumentale" ("Monumental Building")	
3	"Tepidarium" ("Greenhouse")	
4	"Succulente" ("Succulents")	
5	"Orto Siculo" ("Sicilian Garden")	
6	"Giardino Sinistro" ("Leftmost Garden")	
7	"Passaggio" ("Passageway")	
8	"Giardino Centrale" ("Central Garden")	
9	"Giardino Destro" ("Rightmost Garden")	

Figure 1: Topology of the natural site considered in this work. The table in the bottom part of the figure reports details on the 9 contexts highlighted in the map reported at the top of the figure.

has been validated in a controlled environment, but authors shown it is also capable to adapt to changes in the environment (i.e., obstacles, slopes, shelves, etc.). Other works jointly exploiting images and GPS have been proposed in literature. As example, *NavCog* (Ahmetovic et al., 2016) is a smartphone-based system that performs an accurate real-time localization over large spaces.

Similarly to the works discussed above, we investigate methods to combine GPS and vision. However, differently from previous works, our study focuses on the context of outdoor natural environments.

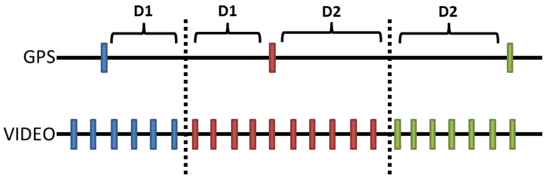


Figure 2: Illustration of the process to align GPS measurements to the frames of the videos. Each video frame is associated to the closest GPS measurement in time. This implicitly defines variable-length time-slot, i.e., the boundaries marked with black dotted lines in the figure. All the frames falling in the same time-slot are associated to the same GPS measurement.

3 DATASET

The dataset used in this work has been collected asking 12 volunteers to visit the Botanical Garden of the University of Catania. The garden has a length of about 170m and a width of about 130m. In accordance with experts, we defined 9 contexts of interest which are relevant to collect behavioral information from the visitors of the site (Fig. 1). The volunteers have been instructed to visit all 9 contexts without any specific constraint, allowing them to spend how much time they wished in each context.

We asked each volunteer to explore the natural site wearing a camera and a smartphone during their visit. The wearable camera has been used to collect egocentric videos of the visits, while the smartphone has been used to collect GPS locations. GPS data has been later synced with the collected video frames. As a wearable camera, we have used a Pupil 3D Eye Tracker headset. Videos have been acquired at a resolution of 1280×720 pixels and a framerate of 60 fps. GPS locations have been recorded using a Honor 9 smartphone. Due to the limitations of using GPS devices when the sky is not clear or because of the presence of trees, GPS locations have been acquired at a slower rate as compared to videos. Specifically, a new GPS signal has been recoreded about every 14 seconds, depending on the capability of the device to communicate with the GPS satellites. Leveraging video and GPS timestamps stored during the data acquisition, each frame is associated to the closest GPS measurement in time. This leads to the replication of a given GPS location over time as it is illustrated in Fig. 2. Each frame of the egocentric videos has been labelled to specify the context in which the volunteer was actually located during the visit (Fig. 1). The labeling has been performed by experts using the ELAN annotation tool[4].

[4] ELAN Website: https://www.mpi.nl/corpus/html/elan/index.html (4-Nov-2018)

Table 1: Number and percentages of training and test frames, as well as the total number of frames contained in the dataset.

	Training	**Test**	**Total**
Number of Frames	$40,436$	$23,145$	$63,581$
Percentage w.r.t Total	63.59%	36.41%	100%

Using the described protocol, we collected and labeled almost 6 hours of recording, from which we sampled a total of $63,581$ frames for our experiments. The selected frames have been resized to 128×128 pixels to decrease the computational load. This image size is coherent with previous work which highlights that a resolution of 128×128 pixels is enough for context recognition (Torralba, 2009).

The dataset has been partitioned by considering about 64% of the data for training and the remaining data for test. This is achieved using frames extracted from 12 videos acquired by 4 volunteers for training, and frames belonging to the remaining 21 videos acquired by 8 volunteers for test. The videos have been selected in order to obtain similar class distributions among the two sets. Table 1 summarizes the number of frames belonging to training and test sets as well as the total number of frames contained in the dataset.

4 METHODS

We compared different approaches to localize the visitors of a natural site from GPS data and egocentric images. We tackle localization as a classification problem, hence we aim at building classifiers able to identify the area in which the visitor is currently located from the considered data. To localize visitors from GPS data only, we train a Decision Classification Tree (DCT). During training, we optimize the “maximum height” parameter of the DCT performing a grid search in the range $[1,100]$. Since a video frame can correspond to multiple identical GPS positions (see Section 3), the training set will contain duplicate samples with the same label, which we experienced degrading the performance of the DCT. Hence, in our experiments, we removed duplicates from the training set. It is worth to note that duplicates are not removed when test set is used for evaluation purpose.

Localization using vision only is performed finetuning a SqueezeNet architecture (Iandola et al., 2016) pre-trained on the ImageNet dataset to classify video frames according to the 9 considered contexts. We choose the SqueezeNet architecture for its compactness and the low computational cost required at test time. The network is trained using Stochastic Gradient Descent with a learning rate of 0.001 for 300 epochs and a batch size of 256. The model scoring the best test accuracy among the different epochs is hence retained to perform evaluation.

We also investigated 4 streamlined CNNs derived from SqueezeNet by considering subsets of its layers. The employment of streamlined CNNs is taken into account to investigate the design of systems which can efficiently combine GPS and vision when the computational budget is low. This is motivated by the need to deploy the localization system in embedded settings as a complementary service. In such cases, a low computational cost is required to save battery and computational resources needed for other services (e.g., to recognize plants). The streamlined models have been obtained considering subsets of the layers of the SqueezeNet architecture. Specifically, we considered the first 6, 9 and 11 layers of the SqueezeNet architecture[5]. Fig. 3 summarizes the architecture of SqueezeNet and highlights the subsets of layers considered to define the three streamlined CNNs. Each of the networks is complemented with a classification module which takes over the activations of the final layer. The classification module is composed by two layers: a convolutional layer generating 9 maps and a global pooling layer computing the average of each of the map and returning 9 class scores. The latter two layers are initialized randomly, while all other layers have been pre-trained on ImageNet. We refer to the considered architectures as SqueezeNet-6, SqueezeNet-9, and SqueezeNet-11, respectively. The networks are trained to perform classification from images only using the same settings used for the full SqueezeNet model, except for the batch size, which is set to 512.

We finally explore how visual information and GPS data can be combined to improve localization. This is obtained by performing late fusion on the probabilities computed by the DCT considering GPS data and the ones predicted by the CNNs. Specifically, late fusion is performed as a linear combination of the probabilities predicted by the two models:

$$p_f(c|x) = w_i \cdot p_i(c|x) + w_g \cdot p_g(c|x) \qquad (1)$$

where $p_f(c|x)$ is the final probability obtained by late fusion for class c when observing the sample x, $p_i(c|x)$ and $p_g(c|x)$ are the probabilities predicted using respectively images and GPS data, w_i and w_g are weights regulating the contribution of each modality to the final prediction. We set $w_i = 2$ and $w_g = 1$ in our experiments, as they lead to best results.

[5] We considered the SqueezeNet v1.1 model, as implemented in Torchvision - https://github.com/pytorch/vision/blob/master/torchvision/models/squeezenet.py (4-Nov-2018)

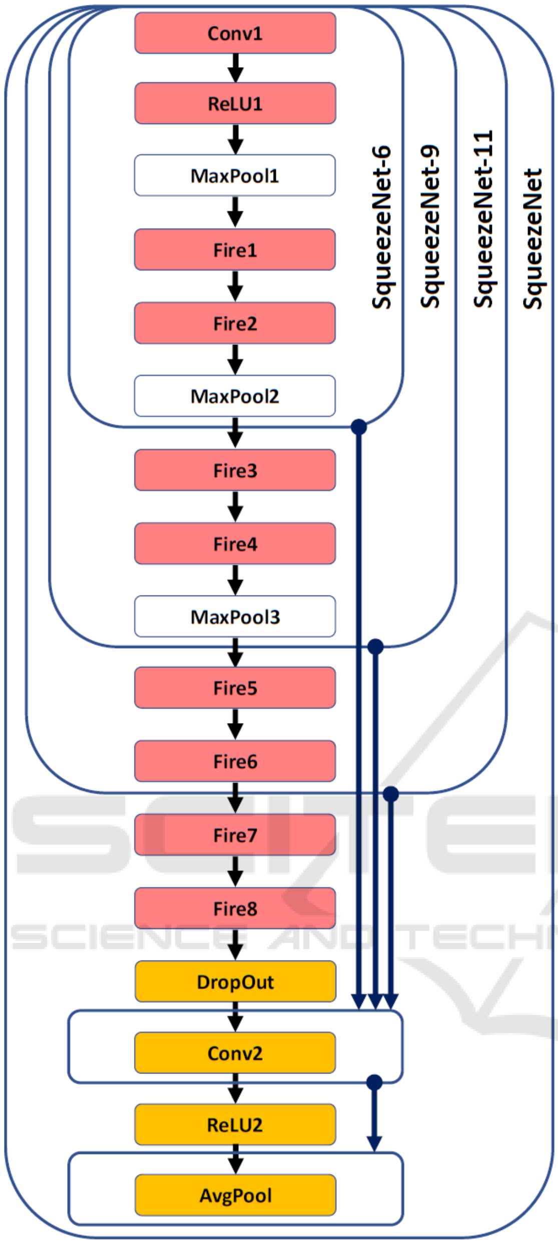


Figure 3: The SqueezeNet architecture and the considered subsets of layers used to build streamlined CNNs for image-based localization.

5 EXPERIMENTAL RESULTS

Table 2 compares the two considered methods to perform localization using only GPS data (DCT) and only visual data (SqueezeNet) in terms of accuracy, memory and computational time[6]. Best results are reported in bold column-wise. Experiments pointed out that, using only vision allows to greatly outperform the classification method based only on GPS in the considered experimental settings. This highlights the noisy nature of GPS measurements in the considered context and the potential of computer vision to address localization. Nevertheless, the approach based on GPS requires far less memory and runs many times faster than the approach based on Squeezenet, which may make it more suited to be deployed in embedded devices despite the reduced performance.

Table 2: Comparison between DCT, using only GPS to perform localization and SqueezeNet, using only vision to perform localization. For each model, we report the amount of memory required, the computational time needed to process a single sample and the accuracy on the test set.

Method	Memory	Time (s)	Accuracy
DCT (GPS)	**0.03 MB**	**$1.30E-07$**	78.76%
SqueezeNet (Vision)	2.80 MB	$2.29E-02$	**91.24%**

Table 3 compares the performances of the methods which combine through late fusion the results of the DCT classifier based on GPS with SqueezeNet and the three streamlined CNNs SqueezeNet-6, SqueezeNet-9 and SqueezeNet-11 based on visual data. For each method, we report the amount of required memory, the computational time needed to process one sample, the accuracy of the CNN, the late fusion accuracy, the improvement of the fused classifier with respect to DCT, and the improvement with respect to the CNN. Best results and second-best results are reported in bold and underlined numbers respectively. As can be seen from the table, the fusion between SqueezeNet-6 and DCT allows to obtain an improvement of $+7.6\%$ with respect to DCT alone and $+4.58\%$ with respect to the SqueezeNet-6 alone, scoring a final accuracy of 86.36%, while requiring only 0.01 *MB* and $4.71E-03$ seconds to process a sample. While the accuracy of the CNNs improve with the increasing number of layers, the improvements over DCT and CNNs obtained by late fusion tend to drop down. This indicates that the shallower models are more kin to learn representations of the data which are complementary with the rough localization already provided GPS. Moreover, it should be noted that deeper CNNs require more memory and computational time. SqueezeNet allows to obtain a boost of about 5.63% with respect to the best of the proposed architecture scoring an accuracy of 91.99%. However, it should be noted that the contribution of fusion with DCT is rather modest (about $+0.75\%$), which suggests that SqueezeNet does not learn representations which are complementary to the available GPS

[6] All time measurements have been performed on CPU using a four-cores Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz. Please note that higher computational times should be expected in embedded settings.

Table 3: Performances of the considered methods which perform classification using both images and GPS data. For each method, we report the needed memory, the computation time required to process one sample, the accuracy of the CNN model, the accuracy of the whole model after late fusion, the improvement with respect to the DCT alone, and the improvement with respect to the CNN alone.

Model	Memory	Time (s)	CNN Acc.	Fusion Acc.	Imp. wrt DCT	Imp. wrt CNN
SqueezeNet-6 + DCT	**0.1 MB**	**4.71E − 03**	81.78%	86.36%	+7.60%	+**4.58**%
SqueezeNet-9 + DCT	0.5 *MB*	6.09*E* − 03	82.52%	86.32%	+7.56%	+3.80%
SqueezeNet-11 + DCT	1.4 *MB*	6.60*E* − 03	85.78%	86.36%	+7.02%	+2.54%
SqueezeNet + DCT	2.8 *MB*	2.29*E* − 02	**91**.**24**%	**91**.**99**%	+**13**.**23**%	+0.75%

data. Moreover, it should be noted that this is achieved with more expensive computation. Indeed, the proposed SqueezeNet-6+DCT method requires 28 times less memory and is 4.9 times faster than the combination of SqueezeNet and DCT.

As a final remark, the results show how shallow and computationally inexpensive CNN models can be leveraged to greatly improve the performance of classifiers based on GPS by naturally learning complementary representations.

6 CONCLUSION

We investigated the use of GPS data and vision to localize people in natural outdoor contexts. To carry out the study, we collected a dataset of egocentric videos and GPS measurements in the Botanical Garden of the University of Catania, Italy. The area of the considered natural site has been divided into meaningful contexts and each frame of the dataset has been labeled according to the context in which the visitor was actually located. We tackle localization as a classification problem and compare different methods aiming at performing localization using only GPS data, visual data and the combination of both modalities through late fusion. Our investigation shows that: 1) localization based on vision is more accurate than localization based on GPS in the considered context, but requires more computational resources, which may hinder its use in embedded settings; 2) the performance of localization methods based on GPS can be greatly improved by fusion with inexpensive shallow CNNs derived from pre-trained networks.

Future works will be devoted to perform a more thorough benchmark study of methods based on GPS and CNNs on the proposed dataset. Moreover, in accordance with the experts, we will evaluate new possible labeling schemes to indicate the presence of contexts and subcontexts in the collected data. Other efforts could also be dedicated to investigating more sophisticated ways to fuse GPS and visual information.

ACKNOWLEDGEMENTS

This research is supported by PON MISE - Horizon 2020, Project VEDI - Vision Exploitation for Data Interpretation, Prog. n. F/050457/02/X32 - CUP: B68I17000800008 - COR: 128032, and Piano della Ricerca 2016-2018 linea di Intervento 2 of DMI of the University of Catania. The authors would like to thank Costantino Laureanti for his invaluable help.

REFERENCES

Ahmetovic, D., Gleason, C., Ruan, C., Kitani, K., Takagi, H., and Asakawa, C. (2016). Navcog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 90–99. ACM.

Alahi, A., Haque, A., and Fei-Fei, L. (2015). Rgb-w: When vision meets wireless. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3289–3297.

Capi, G., Kitani, M., and Ueki, K. (2014). Guide robot intelligent navigation in urban environments. *Advanced Robotics*, 28(15):1043–1053.

Furnari, A., Farinella, G. M., and Battiato, S. (2017). Recognizing personal locations from egocentric videos. *IEEE Transactions on Human-Machine Systems*, 47(1):6–18.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Ishihara, T., Kitani, K. M., Asakawa, C., and Hirose, M. (2017a). Inference machines for supervised bluetooth localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5950–5954. IEEE.

Ishihara, T., Vongkulbhisal, J., Kitani, K. M., and Asakawa, C. (2017b). Beacon-guided structure from motion for smartphone-based navigation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 769–777. IEEE.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.-P., Lombardo, J.-C., Planque, R., Palazzo, S., and Müller, H. (2017). Lifeclef 2017 lab overview: multimedia species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 255–274. Springer.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In *European Conference on Computer Vision*, pages 502–516. Springer.

Ragusa, F., Furnari, A., Battiato, S., Signorello, G., and Farinella, G. M. (2019). Egocentric visitors localization in cultural sites. *Journal on Computing and Cultural Heritage*.

Ragusa, F., Guarnera, L., Furnari, A., Battiato, S., Signorello, G., and Farinella, G. M. (2018). Localization of visitors for cultural sites management. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 1: ICETE,*, pages 407–413. INSTICC, SciTePress.

Santarcangelo, V., Farinella, G. M., and Battiato, S. (2016). Egocentric vision for visual market basket analysis. In *European Conference on Computer Vision Workshop (EPIC)*, pages 518–531. Springer.

Starner, T., Schiele, B., and Pentland, A. S. (1998). Visual contextual awareness in wearable computing. In *Proceedings of the International Symposium on Wearable Computing*, pages 50–57.

Torralba, A. (2009). How many pixels make an image? *Visual neuroscience*, 26(1):123–131.

Van Horn, G., Mac Aodha, O., Song, Y., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2017). The iNaturalist challenge 2017 dataset. *arXiv preprint arXiv:1707.06642*.

Wegner, J. D., Branson, S., Hall, D., Schindler, K., and Perona, P. (2016). Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023.

Weyand, T., Kostrikov, I., and Philbin, J. (2016). Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer.