Localizing Visitors in Natural Sites Exploiting Modality Attention on Egocentric Images and GPS Data

Giovanni Pasqualino[1,*], Stefano Scafiti[1,*], Antonino Furnari[1] and Giovanni Maria Farinella[1,2]

[1]*Department of Mathematics and Computer Science, University of Catania, Catania, Italy*

[2]*Cognitive Robotics and Social Sensing Laboratory, ICAR-CNR, Palermo, Italy*

Keywords: Egocentric (First Person) Vision, Localization, GPS, Multi-modal Data Fusion.

Abstract: Localizing the visitors of an outdoor natural site can be advantageous to study their behavior as well as to provide them information on where they are and what to visit in the site. Despite GPS can generally be used to perform outdoor localization, we show that this kind of signal is not always accurate enough in real-case scenarios. On the contrary, localization based on egocentric images can be more accurate but it generally results in more expensive computation. In this paper, we investigate how fusing image- and GPS-based predictions can allow to achieve efficient and accurate localization of the visitors of a natural site. Specifically, we compare different fusion techniques, including a modality attention approach which is shown to provide the best performances. Results point out that the proposed technique achieve promising results, allowing to obtain the performances of very deep models (e.g., DenseNet) with a less expensive architecture (e.g., SqueezeNet) which employ a memory footprint of about $3MB$ and an inference speed of about $25ms$.

1 INTRODUCTION

Smart wearable and mobile devices equipped with a camera and a display offer a convenient platform to improve the fruition of indoor cultural sites such as art galleries and museums (Cucchiara and Del Bimbo, 2014; Seidenari et al., 2017), outdoor urban places such as cities (Alkhafaji et al., 2019; Alletto et al., 2016), as well as outdoor natural environments such as parks and gardens (Milotta et al., 2019a; Milotta et al., 2019b). Notably, smart glasses allow to explore the cultural site and receive additional information and services through augmented reality in a "hands-free" fashion. At the same time, such devices allow to effortlessly collect visual information about the behavior of the user which, if properly analyzed, can provide value to the site manager (Farinella et al., 2019). In particular, the continuous localization of the camera wearers allows to provide the visitors with a "where am I" service, which can be useful to help them navigate the site and provide additional information on the current area or alternative routes to visit. At the same time, a "where are they" service can be provided to the site manager to help locate all visitors to understand which areas are the most visited and where people tend to spend more time.

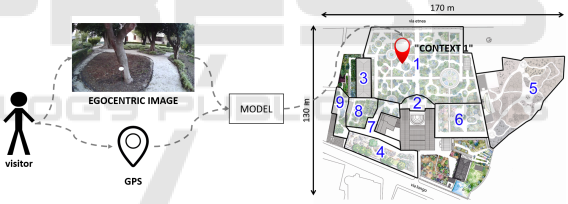

Figure 1: In the considered scenario, users are localized in a natural site using both egocentric images and GPS data. The natural site is divided into non-overlapping areas and localization is addressed as a classification problem.

Most previous works focused on recognizing scene context (Battiato et al., 2008) and on localizing the visitors in indoor scenarios (Giuliano et al., 2014; Ragusa et al., 2019; Farinella et al., 2019), whereas the localization of users in outdoor natural sites has been relatively less investigated (Milotta et al., 2019a; Milotta et al., 2019b). While GPS can be generally exploited for localization in outdoor contexts, previous works (Milotta et al., 2019a; Milotta et al., 2019b) have shown that this is not always the case due to the limited accuracy of consumer GPS receivers, especially in natural scenarios in which the sky might be covered by trees. In particular, in (Milotta et al., 2019a) it has been shown that localizing the visitors from egocentric images is much more accurate than using GPS. The work of (Milotta et al., 2019b) further

[*]These authors are co-first authors.

Pasqualino, G., Scafiti, S., Furnari, A. and Farinella, G.
Localizing Visitors in Natural Sites Exploiting Modality Attention on Egocentric Images and GPS Data.
DOI: 10.5220/0009103906090617
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 609-617
ISBN: 978-989-758-402-2; ISSN: 2184-4321

investigated how combining a shallow Convolutional Neural Network (CNN) for image-based localization with a simple decision tree to process GPS data can allow to obtain a good localization accuracy at a low computational cost, which is a key factor when the system has to be deployed in the embedded settings imposed by wearable devices.

In this paper, we consider the problem of localizing the visitors of natural sites from egocentric images and GPS. As proposed in previous investigations (Ragusa et al., 2019; Milotta et al., 2019a; Milotta et al., 2019b), in the considered scenario, the natural site is divided into non-overlapping areas and localization is addressed as a classification problem (see Figure 1). To carry out the investigation, we rely on the dataset proposed in (Milotta et al., 2019a). Similarly to (Milotta et al., 2019a; Milotta et al., 2019b) we consider a multi-modal architecture which processes images and GPS data using specialized branches. While (Milotta et al., 2019a; Milotta et al., 2019b) rely on late fusion to combine the predictions of the different branches, we show that better results can be achieved introducing a "modality attention" (Furnari and Farinella, 2019) module, which uses intermediate representations from the two branches to compute fusion weights. This allows the overall architecture to "vote" for which modality is more to be trusted on a sample-by-sample basis. We validate the proposed approach by comparing it with the results reported in (Milotta et al., 2019a; Milotta et al., 2019b), as well as with several multi-modal fusion approaches based on weighted late fusion, and "learned" late fusion.

The reminder of the paper is organized as following. The related works are discussed in Section 2. The compared localization approaches are detailed in Section 3. Experimental settings and results are given in Section 4. Section 5 concludes the paper.

2 RELATED WORKS

Previous works have investigated the use of computer vision in natural environments. Kumar et al. (Kumar et al., 2012) proposed a system which allows to identify 184 species of trees from pictures acquired using a mobile application. Wegner et al. (Wegner et al., 2016) introduced a framework for the recognition of trees which leverages data collected from Google street view. Horn et al. (Horn et al., 2017) proposed the iNat2007 dataset, which contains images of about $8,000$ different species of plants collected in natural environments. Joly et al. (Joly et al., 2017) introduced the LifeCLEF dataset to address the investigation of different tasks, including identification of birds based on audio, recognition of plants from images, vision-based monitoring of organisms living in seas, and recommendation of species based on location. Milotta et al. (Milotta et al., 2019a; Milotta et al., 2019b) introduced a dataset of egocentric videos paired with GPS measurements collected by different visitors in a natural site. A subset of video frames has been labeled to specify the area of the natural site in which the visitor was located during the acquisition. In particular, labels are provided at different granularity levels, including high-level contexts and fine-grained subcontexts. Localization has been hence addressed as a classification problem. The authors also investigated localization from images and GPS using standard late fusion techniques, showing how accurate localization can be obtained at a low computational budget using a shallow CNN processing images and a decision tree processing GPS.

Our work builds on the investigation previously proposed by Milotta et al. (Milotta et al., 2019a; Milotta et al., 2019b). In particular, we show that the use of a modality attention mechanism (Furnari and Farinella, 2019) can improve localization accuracy based on images and GPS. The proposed approach is compared both computationally and in terms of accuracy with respect to the baselines proposed in (Milotta et al., 2019a; Milotta et al., 2019b), which are outperformed.

3 METHODS

In this Section, we present the methods to address the localization problem which have been compared in this study. All considered approaches are classifiers over m possible classes $C = 1,\dots,m$ which take as input either a pair of GPS coordinates $\mathbf{x} = (x,y)$ collected using a GPS receiver, an egocentric image I, or a pair $(\mathbf{x},I)$ consisting in a sensed GPS location $\mathbf{x}$ and an egocentric image I collected at the same time. It is worth noting that, as it will be clear from the experiments, the sensed GPS location tend to be noisy. Hence, the task of learning a classifier $f(\mathbf{x})$ from GPS data is not trivial. Indeed, we will show that different methods based on GPS have different performance. We will hence consider two main classes of methods: 1) single-modal approaches, which perform localization by inferring the current class from either the GPS location ($f(\mathbf{x}) \to \hat{y} \in C$) or the egocentric image ($f(I) \to \hat{y} \in C$), 2) multi-modal approaches, which infer the area in which the visitor is currently located by processing both GPS locations and images ($f(\mathbf{x},I) \to \hat{y} \in C$). The following sections discuss the details of the methods considered for the two families

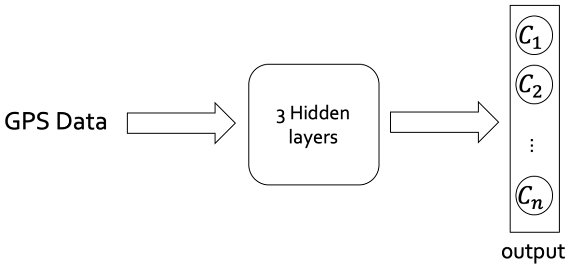

Figure 2: Multilayer Perceptron architecture used to localize the visitors based on GPS data.

of approaches.

3.1 Single-modal Localization based on GPS Data or Egocentric Images

GPS: We used a multilayer perceptron which takes a pair of GPS coordinates as input to infer in which area the user is located. Specifically, the multilayer perceptron is composed of five fully connected layers: one input layer, three hidden layers and one output layer. All layers have the same activation function (Relu) and each layer doubles the dimensionality of the received input. The network has been trained for 30 epochs using the cross-entropy loss fuction. Batch size has been set to 8, momentum to 0.9 and the learning rate to 0.001. The architecture is sketched in Figure 2.

Egocentric Images: We considered different CNN architectures: ResNet (He et al., 2016), DenseNet (Huang et al., 2017) and SqueezeNet (Iandola et al., 2016). We also considered a shallow variant of SqueezeNet obtained by truncating the network to the first 6 layer as proposed in (Milotta et al., 2019b). We refer to this architecture as SqueezeNet-6. In all cases, we consider the models pre-trained on ImageNet (Deng et al., 2009). Each network is fine-tuned to predict the area in which the visitor is currently located from an input egocentric image. The hyperparameters were set in the same way as the multilayer discussed in previous paragraph, only the batch size was changed to 3. To assess the effect of fusing two networks operating on the same modality, as compared to multi-modal fusion, we also consider an architecture which combines the two approaches as depicted in Figure 3. The two vectors have a size of 32 and represent the probability distribution over the classes of the two networks. This two vectors are concatenated and given as input of a multilayer perceptron.

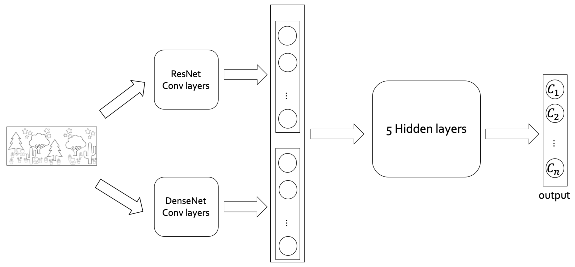

Figure 3: Combined CNNs Architecture.

3.2 Multi-modal Localization Exploiting Images and GPS

We propose to perform multi-modal localization by processing both images and GPS using a "modality attention" approach similar to the one proposed in (Furnari and Farinella, 2019). This approach is compared with respect to two baseline fusion methods: "weighted late fusion" and "learned late fusion". All these approaches assume that two modality-specific branches are available. The "image branch" f_{im} takes as input an egocentric image I and outputs the conditional probability distribution $p(c|I) = f_{im}(I)$. The "GPS branch" f_{GPS} takes as input the GPS coordinates $\mathbf{x}$ and outputs the conditional probability distribution $p(c|\mathbf{x}) = f_{GPS}(\mathbf{x})$. In all our experiments, f_{im} is implemented as a CNN and f_{GPS} is implemented as a multilayer perceptron. The different fusion approaches considered for comparisons seek ways to fuse the two probability distributions, to obtain a multi-modal prediction $p(c|I,\mathbf{x})$.

Weighted Late Fusion: This approach computes the multi-modal prediction $p(c|I,\mathbf{x})$ by computing a weighted average between the modality specific predictions:

$$p(c|I,\mathbf{x}) = \lambda f_{im}(I) + (1-\lambda) f_{GPS}(\mathbf{x}) \qquad (1)$$

where $\lambda \in [0,1]$ is a fusion parameter which regulates the trade-off between image-based predictions and GPS-based predictions. In our experiments, the two branches are trained independently (we have used the model described in previous section) on their respective modalities and the λ parameter is later tuned with a grid search on the training set. Figure 4 illustrates the weighted late fusion architecture. It should be noted that this fusion technique is a very standard approach, as investigated in (Milotta et al., 2019a; Milotta et al., 2019b).

Learned Late Fusion: The weighted late fusion approach might be limited by the fact that modality-specifc predictions are aggregated with a simple linear combination. To allow the model to learn better ways to combine the predictions, we explore "learned

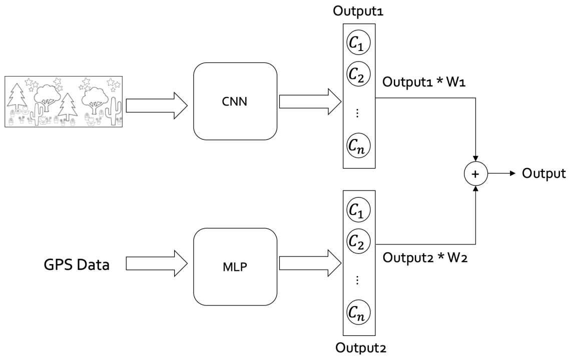

Figure 4: Weighted Late Fusion Architecture.

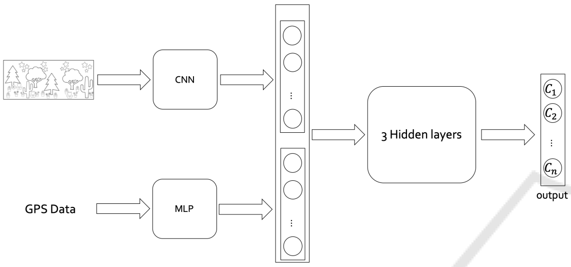

Figure 5: Learned Late Fusion Architecture.

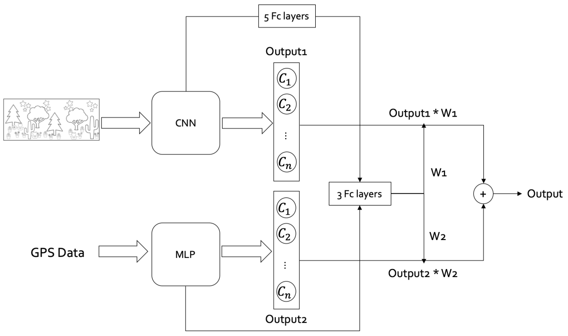

Figure 6: Modality Attention Architecture.

late fusion” approach in which the outputs of the two branches are concatenated and fed to a fusion network f_{fusion} which directly outputs the multi-modal prediction: $p(c|I,\mathbf{x}) = f_{fusion}(f_{img}(I), f_{GPS}(\mathbf{x}))$. We implement the fusion network as a multilayer perceptron with three fully connected layers. Figure 5 illustrates the learned late fusion architecture.

Modality Attention: While the learned late fusion approach can in principle learn how to best combine the modality-specific predictions, we observe that it is more exposed to over-fitting, due to the highly increased number of parameters. On the contrary, the simple weighted late fusion approach keeps a very simple late fusion scheme but imposes the fusion weights to be fixed at inference time. Inspired by recent work (Furnari and Farinella, 2019), we consider a modality attention mechanism which computes the late fusion weights on a sample-by-sample basis. This allows for more flexibility as the model can learn to assign different weights to the different modalities depending on the input samples. To obtain the final prediction, we first consider some intermediate representations computed by the modality-specific branches f_{img} and f_{GPS}. We will refer to those as $\phi_{img}(I)$ and $\phi_{GPS}(\mathbf{x})$. These representations are concatenated and fed to a Modality Attention (MATT) network f_{MATT}, which outputs two scores s_1 and s_2: $s_1, s_2 = f_{MATT}(\phi_{img}(I), \phi_{GPS}(\mathbf{x}))$. The scores are hence normalized using the Softmax function to obtain suitable late fusion weights which sum to 1:

$$w_1 = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}, \quad w_2 = \frac{e^{s_2}}{e^{s_1} + e^{s_2}}. \qquad (2)$$

The computed weights are hence used to perform late fusion as follows:

$$p(c|I,\mathbf{x}) = w_1 \cdot f_{img}(I) + w_2 \cdot f_{GPS}(\mathbf{x}) \qquad (3)$$

In our experiments, we extract the internal representations of the two branches just before the final classification layers to obtain $\phi_{img}(I)$ and $\phi_{GPS}(\mathbf{x})$. The modality attention network f_{MATT} is implemented as a multilayer perceptron with three fully connected layers which each of them halves its input. To maximize performance, we first train each modality-specific branch independently. Then, we initialize f_{MATT} randomly and fine-tune the whole architecture by applying the cross-entropy loss to the final output $p(c|I,\mathbf{x})$. It should be noted that this loss is differentiable with respect to the parameters of both f_{img}, f_{GPS} and f_{MATT}. The network has been trained for 15 epochs, the batch size has been set to 3, momentum to 0.9 and the learning rate to 0.001.

3.3 State of the Art Approaches

We also consider the approaches investigated in (Milotta et al., 2019a; Milotta et al., 2019b) for comparison. These consist in the fusion of different CNN architectures, including SqueezeNet (Iandola et al., 2016), AlexNet (Krizhevsky et al., 2012) and VGG (Simonyan and Zisserman, 2015), with different classifiers based on GPS, including a K-Nearest Neighbor (KNN), a Decision Classification Tree (DCT) and a Support Vector Machine (SVM). All these approaches perform fusion using a standard late fusion approach similar in spirit to the “weighted late fusion” method considered in this paper. We compare with respect to these approaches by reporting the accuracy values declared by the authors in (Milotta et al., 2019a; Milotta et al., 2019b).

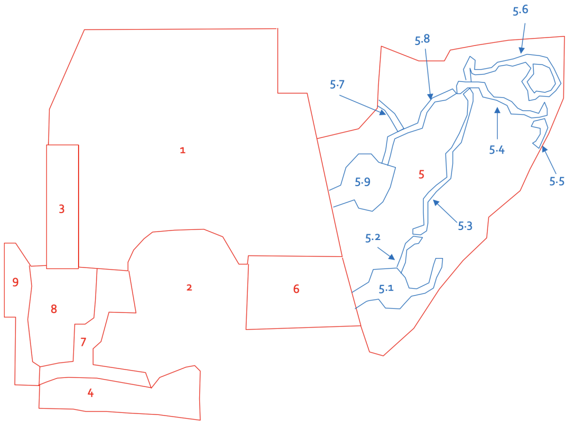

Figure 7: Topology of the natural site. The space is composed by 9 context (coloured in red). For the area number 5 the botanic expert have defined 9 subcontext (coloured in blue). The 17 context area are obtained substituting the 9 subcontext to the context numbers.

4 EXPERIMENTAL RESULTS

4.1 Dataset

For our experiments, we considered the dataset introduced in (Milotta et al., 2019a). This dataset comprises about 6 hours of egocentric videos collected by different visitors while exploring the natural site of the Botanical Garden of the University of Catania. Specifically, the authors released $63,581$ images sampled from the videos. To allow evaluation by cross-validation, the set of data has been partitioned into three folds. Each image is associated to a GPS position retrieved with a smartphone and labeled according to the area in which the visitor was located at the time of acquisition. The labels are provided according to three different levels of localization granularity:

- 9 Contexts: the area of the site has been divided into 9 different contexts, relevant for the behavioral analysis of the visitors to understand for instance which plants they have seen and which areas they have spent more time into;
- 9 Subcontexts: the area of context 5, one of the 9 original contexts, has been further divided into 9 subcontexts. These provide a more fine-grained understanding of where the user has spent more time in context 5, which is the most characteristic part of the garden. Note that this classification task requires a much more accurate localization;
- 17 Contexts: this is a mixed set of classes in which the 8 contexts (excluding context 5) are merged with the 9 subcontexts of context 5. To address this classification task, the methods needs to be able to infer location at both a coarse (the 8 contexts) and a fine (the 9 subcontexts) level.

Table 1: Accuracy of different single-modal methods based on GPS in the 9 Contexts scenario. The performance scores of the methods marked with the ”*” sign are reported from (Milotta et al., 2019a).

Model	Fold 0	Fold 1	Fold 2	AVG
DCT*	78.76 %	46.83 %	76.53%	67.37%
SVM*	78.89 %	52.59 %	78.50%	69.99%
KNN*	80.38 %	54.34 %	81.05%	71.92%
MLP	**82.44%**	**63.13%**	**82.57%**	**76.04%**

Figure 7 illustrates the topology of the natural site including the subdivision of the space into contexts and subcontexts.

All experiments have been performed using an NVIDIA GTX 1050 GPU with 2GB of RAM. We compared all architectures on the 9 contexts settings to provide ablation with respect to the different fusion approaches. The methods achieving best performance in these settings are further evaluated on the 9 subcontexts and 17 contexts scenarios.

4.2 Results on the 9 Contexts Settings

4.2.1 Methods based Only on GPS

Table 1 compares the performance of the considered Multilayer Perceptron (MLP) with respect to the Decision Classification Tree (DCT), the Support Vector Machine (SVM) and the K-Nearest Neighbor (KNN) classifiers investigated in (Milotta et al., 2019a). The table reports the results obtained on each fold, as well as the average performance across folds. Best results per-column are highlighted in bold. In average, the accuracy of MLP increases by 9% compared to DCT, 6% compared to SVM and 4% compared to KNN. These observed performance gaps suggest that GPS coordinates are too noisy to allow for accurate localization on the 9 contexts. Indeed, if they were accurate enough, geometrical approaches such as decision trees, SVM and KNN should be able to easily divide the space into regions to perform classification. Instead, the non-linearity of a properly tuned MLP allows to improve performance by significant margins.

4.2.2 Methods based Only on Images

Table 2 compares the performance of different CNNs for image-based localization. The best performances per-column are reported in bold numbers. As it can be expected, the shallow SqueezeNet-6 obtains lower performance as compared to the full SqueezeNet,

Table 2: Accuracy of different CNNs for image-based localization in the 9 Contexts scenario. The performance scores of the methods marked with the "*" symbol are reported from (Milotta et al., 2019a).

Model	Fold 0	Fold 1	Fold 2	AVG
SqueezeNet-6	82.01%	79.12%	82.70%	81.27%
AlexNet*	90.99%	90.72%	89.63%	90.45%
SqueezeNet*	91.24 %	93.23%	91.40%	91.91%
ResNet18	94.01%	94.89%	93.78%	94.25%
VGG16*	94.26%	**95.59%**	94.08%	94.64%
Combined ResNet18 + DenseNet121	94.43%	95.00%	94.99%	94.80%
DenseNet121	**94.76%**	95.29%	**95.18%**	**95.07%**

Table 3: Accuracy of the multi-modal methods processing both images and GPS on the 9 Contexts scenario. For reference, the top part of the table also reports the performances of the single-modal branches. The performance scores of the methods marked with the "*" symbol are reported from (Milotta et al., 2019a). Per-fold results have not been made available by the authors of (Milotta et al., 2019a) for these methods.

Model	Fold 0	Fold 1	Fold 2	AVG
DCT*	78.76 %	46.83 %	76.53%	67.37%
MLP	82.44%	63.13%	82.57%	76.04%
SqueezeNet-6	82.01%	79.12%	82.70%	81.27%
AlexNet*	90.99%	90.72%	89.63%	90.45%
SqueezeNet*	91.24%	93.23%	91.40%	91.91%
VGG16*	94.26%	95.59%	94.08%	94.64%
DenseNet121	94.76%	95.29%	95.18%	95.07%
MLP + SqueezeNet-6 (learned late fusion)	92.65%	79.84%	91.97%	88.15%
MLP + SqueezeNet-6 (weighted late fusion)	92.13%	81.56%	91.52%	88.40%
MLP + SqueezeNet-6 (modality attention)	93.44%	79.99%	93.58%	89.00%
AlexNet + DCT (weighted late fusion)*	-	-	-	91.33%
SqueezeNet + DCT (weighted late fusion)*	-	-	-	92.47%
VGG16 + DCT (weighted late fusion)*	-	-	-	94.86%
MLP + SqueezeNet (weighted late fusion)	94.15%	94.02%	94.11%	94.09%
MLP + SqueezeNet (learned late fusion)	94.49%	93.35%	94.56%	94.13%
MLP + SqueezeNet (modality attention)	95.57%	93.53%	96.06%	95.05%
MLP + DenseNet121 (modality attention)	**96.15%**	**95.60%**	**96.09%**	**95.94%**

which is probably due to the lower number of parameters and to the limited flexibility due to the reduced depth. In general, deeper models tend to perform better than shallow ones (e.g., VGG16 vs AlexNet and DenseNet121 vs VGG16 and ResNet18). To assess whether fusing two CNNs can be beneficial, we used the "Combined CNNs" fusion architecture illustrated in Figure 3 to combine ResNet18 and DenseNet121. As can be noted, this results in a performance drop, probably due to overfitting induced by the increased capacity of the overall architecture and to the redundancy of the visual features extracted by the two networks. It is worth noting that even a very shallow CNN such as SqueezeNet-6 already outperforms GPS-based localization obtaining a 81.27% accuracy, versus the accuracy of the MLP for GPS-based classification which is equal to 76.04%. This highlights how vision-based classification can be more accurate than classification obtained from noisy GPS measurements.

4.2.3 Methods based on Images and GPS

Table 3 reports the accuracy achieved by multi-modal approaches on the 9 Contexts scenario. To assess the improvements due to fusion, the top part of the table also reports the performance of the single-branch models. The table also includes the results reported in (Milotta et al., 2019a) for comparison. All methods obtained by fusing the shallow SqueezeNet-6 CNN and the MLP significantly improve over the single-modal branches. For instance, combining the MLP and SqueezeNet-6 with modality attention allows to obtain an accuracy of 89.00%, which is about 13% higher than the one obtained using the MLP alone (76.04%) and about 8% higher than the one obtained using SqueezeNet-6 alone (81.27%). Improvements are observed also for the full SqueezeNet (95.00% with modality attention vs 76.04% with MLP and 91.91% with the CNN alone) and for DenseNet121 (95.94% with modality attention vs 95.07% with the

Table 4: Accuracy of the single-modal and multi-modal methods on the 9 Subcontexts scenario. The performance scores of the methods marked with the ”*” symbol are reported from (Milotta et al., 2019a). Per-fold results have not been made available by the authors of (Milotta et al., 2019a) for these methods.

Model	Fold 0	Fold 1	Fold 2	AVG
MLP	59.66%	49.01%	51.72%	53.46%
AlexNet + DCT (weighted late fusion)*	-	-	-	83.68%
SqueezeNet	83.96%	88.06%	84.21%	85.41%
SqueezeNet + DCT (weighted late fusion)*	-	-	-	85.89%
SqueezeNet+MLP (modality attention)	86.51%	88.10%	86.28%	86.96%
VGG16 + DCT (weighted late fusion)*	-	-	-	90.01%
DenseNet121	91.15%	92.11%	89.35%	90.87%
DenseNet121+MLP (modality attention)	**92.23%**	**92.29%**	**91.40%**	**91.97%**

Table 5: Accuracy of the single-modal and multi-modal methods on the 17 Contexts scenario. The performance scores of the methods marked with the ”*” symbol are reported from (Milotta et al., 2019a). Per-fold results have not been made available by the authors of (Milotta et al., 2019a) for these methods.

Model	Fold 0	Fold 1	Fold 2	AVG
MLP	66.97%	48.58%	66.32%	60.62%
SqueezeNet	87.94%	91.07%	87.42%	88.81%
AlexNet + DCT (weighted late fusion)*	-	-	-	86.59%
SqueezeNet + DCT (weighted late fusion)*	-	-	-	89.37%
SqueezeNet+MLP (modality attention)	91.24%	91.12%	91.46%	91.27%
DenseNet121	91.42%	94.25%	91.58%	92.41%
VGG16 + DCT (weighted late fusion)*	-	-	-	92.43%
DenseNet121+MLP (modality attention)	**94.44%**	**94.59%**	**92.85%**	**93.96%**

CNN alone). It is worth noting that for deeper CNNs the relative improvement of the fusion with respect to the CNN become smaller. This is due to the fact that CNNs can already solve the problem with high accuracy without GPS. Interestingly, even in such cases, the proposed modality attention approach always brings improvements, even if marginal, and never leads to overfitting. It can also been observed that modality attention outperforms both weighted late fusion and learned late fusion with both SqueezeNet-5 and SqueezeNet. Indeed, weighted late fusion leads to minimal improvements, whereas learned late fusion performs even worse than weighted late fusion. Finally, even the relatively small model fusing SqueezeNet and MLP with modality attention outperforms all approaches previously reported in (Milotta et al., 2019a), which are indicated by a ”*” symbol in the table.

4.3 Results on the 9 Subcontexts and 17 Contexts Scenarios

Table 4 and Table 5 compare the best performing models based on the fusion of SqueezeNet and DenseNet121 with MLP through modality attention with the respective single-branch performances and with previous methods proposed in (Milotta et al., 2019a) (marked with the ’*’ symbol). In the 9 subcontexts scenario (Table 4) modality attention allows to obtain large improvements over the MLP both when SqueezeNet (86.96% vs 53.46%) and DenseNet121 (91.97% vs 53.46%) are considered. The methods also improve over the performances achieved by the respective CNNs. As can be noted, the proposed approaches outperform the methods proposed in (Milotta et al., 2019a).

Similar trends can be observed in the case of the 17 contexts (Table 5), where the proposed SqueezeNet+MLP model with modality attention outperforms the SqueezeNet+DCT model proposed in (Milotta et al., 2019a) by $+1.9\%$ and similarly, DenseNet121+MLP outperforms VGG16+DCT by $+1.53\%$ and AlexNet+DCT by $+7.37\%$. These results highlight the flexibility of the proposed modality attention approach for multi-modal fusion.

4.4 Computational Resources Analysis

Table 6 reports the required time and memory of the considered methods. Times have been computed on CPU using a four-cores Intel® Core$^{\text{TM}}$ i7-7700HQ @ 3.80GHz, averaging over 400 predictions. For reference, we also report the accuracy of such methods on the 9 Contexts scenario. While methods based on GPS are fast and require very little memory, their performances are limited. Using even small CNNs such

Table 6: Execution time, required memory and average accuracy on the 6 Contexts scenario of the different methods considered in this study.

Model	Time (ms)	Mem (MB)	Accuracy
DCT	0.01	0.07	67.37%
MLP	0.49	0.006	76.04%
SqueezeNet-6	7.88	0.31	81.27%
SqueezeNet-6 + MLP (modality attention)	9.23	0.50	89.00%
SqueezeNet	18.30	2.78	91.91%
AlexNet	20.25	217.60	90.45%
SqueezeNet + MLP (modality attention)	24.40	3.19	95.05%
VGG16	434.86	512.32	94.64%
DenseNet121	559.09	27.685	95.07%
DenseNet121 + MLP (modality attention)	563.71	30.434	95.94%

as SqueezeNet-6 and SqueezeNet, allows to improve performance with a small computational overhead. In particular, the SequeezeNet-6 + MLP (modality attention) model achieves significantly better performance with respect to DCT, MLP and SqueezeNet-6 still mantaining a fast inference and a small memory footprint. Fusing a full SqueezeNet model with an MLP using modality attention allows to achieve performance comparable to the one obtained by much larger model such as DenseNet121 still maintaining a fast inference (about 25ms) and a very small memory footprint (about 3MB). This suggests that fusing predictions based on images and GPS with modality attention can bring a significant boost in performance with a very efficient inference.

5 CONCLUSION

We have investigated the use of different fusion techniques to improve the localization of visitors in a natural site from egocentric images and noisy GPS measurements. The experiments have highlighted that: 1) GPS data alone allows to achieve only limited performance, which suggests that such data is not accurate enough in the considered context, 2) localization from egocentric images can achieve much more accurate results, 3) fusing image- and GPS-based predictions generally allows to improve results and in particular, 4) the proposed modality attention fusion mechanism allows to achieve good localization performance at a very low computational budget, which makes the investigated methodologies suitable for implementation in embedded settings. Future works can investigate the use of the considered fusion techniques to perform a more fine-grained localization based on camera pose estimation.

ACKNOWLEDGEMENTS

This research is part of the project VALUE - Visual Analysis for Localization and Understanding of Environments (N.08CT6209090207, CUP G69J18001060007) supported by PO FESR 2014/2020 - Azione 1.1.5. - “Sostegno allavanzamento tecnologico delle imprese attraverso il finanziamento di linee pilota e azioni di validazione precoce dei prodotti e di dimostrazioni su larga scala” del PO FESR Sicilia 2014/2020, and Piano della Ricerca 2016-2018 linea di Intervento 2 of DMI, University of Catania.

REFERENCES

Alkhafaji, A., Fallahkhair, S., and Cocea, M. (2019). Design challenges for mobile and wearable systems to support learning on-the-move at outdoor cultural heritage sites. In Lamas, D., Loizides, F., Nacke, L., Petrie, H., Winckler, M., and Zaphiris, P., editors, *Human-Computer Interaction – INTERACT 2019*, pages 185–207, Cham. Springer International Publishing.

Alletto, S., Abati, D., Serra, G., and Cucchiara, R. (2016). Exploring architectural details through a wearable egocentric vision device. *Sensors*, 16(2):237.

Battiato, S., Farinella, G. M., Gallo, G., and Ravi, D. (2008). Scene categorization using bag of textons on spatial hierarchy. In *2008 15th IEEE International Conference on Image Processing*, pages 2536–2539.

Cucchiara, R. and Del Bimbo, A. (2014). Visions for augmented cultural heritage experience. *IEEE MultiMedia*, 21(1):74–82.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Farinella, G., Signorello, G., Battiato, S., Furnari, A., Ragusa, F., Leonardi, R., Ragusa, E., Scuderi, E., Lopes, A., Santo, L., et al. (2019). VEDI: Vision Exploitation for Data Interpretation. In *International Conference on Image Analysis and Processing*, pages 753–763. Springer.

Furnari, A. and Farinella, G. M. (2019). What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision*.

Giuliano, R., Marzovillo, M., Mazzenga, F., and Vari, M. (2014). Visitors localization in cultural heritages for experience enhancement. In *2014 Euro Med Telco Conference (EMTC)*, pages 1–6. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Horn, G. V., Mac Aodha, O., Song, Y., Shepard, A., Adam, H., Perona, P., and Belongie, S. J. (2017). The inaturalist challenge 2017 dataset. *CoRR*, abs/1707.06642.

Huang, G., Liu, Z., v. d. Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

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 <0.5mb model size.

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 Jones, G. J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., and Ferro, N., editors, *CLEF: Cross-Language Evaluation Forum*, volume LNCS of *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 255–274, Dublin, Ireland. Springer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. B. (2012). Leafsnap: A computer vision system for automatic plant species identification. In Fitzgibbon, A. W., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *ECCV (2)*, volume 7573 of *Lecture Notes in Computer Science*, pages 502–516. Springer.

Milotta, F. L., Furnari, A., Battiato, S., Signorello, G., and Farinella, G. M. (2019a). Egocentric visitors localization in natural sites. *Journal of Visual Communication and Image Representation*, page 102664.

Milotta, F. L. M., Furnari, A., Battiato, S., Salvo, M. D., Signorello, G., and Farinella, G. M. (2019b). Visitors localization in natural sites exploiting egovision and gps. In *International Conference on Computer Vision Theory and Applications (VISAPP)*.

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 (JOCCH)*, 12(2):11.

Seidenari, L., Baecchi, C., Uricchio, T., Ferracani, A., Bertini, M., and Bimbo, A. D. (2017). Deep artwork detection and retrieval for automatic context-aware audio guides. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3s):35.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6014–6023.