Egocentric Shopping Cart Localization

Emiliano Spera*† Antonino Furnari* Sebastiano Battiato* Giovanni Maria Farinella*
*Department of Mathematics and Computer Science - University of Catania, Italy
†Centro Studi S.r.l., Zona Industriale Buccino, Italy
Email: {spera, furnari, battiato, gfarinella}@dmi.unict.it

***Abstract*—This work investigates the new problem of image-based egocentric shopping cart localization in retail stores. The contribution of our work is two-fold. First, we propose a novel large-scale dataset for image-based egocentric shopping cart localization. The dataset has been collected using cameras placed on shopping carts in a large retail store. It contains a total of $19,531$ image frames, each labelled with its six Degrees Of Freedom pose. We study the localization problem by analysing how cart locations should be represented and estimated, and how to assess the localization results. Second, we benchmark different image-based techniques to address the task. Specifically, we investigate two families of algorithms: classic methods based on image retrieval and emerging methods based on regression. Experimental results show that methods based on image retrieval largely outperform regression-based approaches. We also point out that deep metric learning techniques allow to learn better visual representations w.r.t. other architectures, and are useful to improve the localization results of both retrieval-based and regression-based approaches. Our findings suggest that deep metric learning techniques can help bridge the gap between retrieval-based and regression-based methods.**

I. Introduction

The ability to estimate the position and orientation of a mobile object (e.g., a robot) from egocentric images is crucial to many industrial applications [1], [3], [4]. As it has been investigated by Santarcangelo et al. [27], in the context of retail stores, the position of shopping carts equipped with a camera can be obtained exploiting computer vision pipelines for scene classification. Such information can be used to analyse the customer behaviours, trying to infer, for instance, where they spend more time, which areas of the store are preferred (e.g., fruit, gastronomy, etc.) and how the placement of products can affect sales. Image-based localization abilities are also necessary to allow a robot to navigate and monitor the store or assist the costumers [10].

In this paper, we consider the problem of localizing shopping carts in retail stores from egocentric images acquired by cameras mounted on shopping carts. Differently from other indoor environments, retail stores present unique properties and challenges: 1) they are often large scale environments, 2) they usually present repetitive structures (e.g., the different shelves as well as some products may look very similar), 3) the visual appearance of the different parts of a store is always changing (e.g., products are moved by customers and sometimes are moved from one shelf to another). Fig. 1 shows some examples of the typical variability of egocentric images acquired in a retail store.

Fig. 1: Visual variability of acquired egocentric images.

Despite different datasets are available in the literature to study the problem of image-based indoor localization [3], [4], a large dataset to address the task of shopping cart localization in a retail store is still missing. Hence, we propose a new large scale dataset of images acquired in a retail store using cameras mounted on shopping carts Fig. 2. By means of careful semi-automatic 3D reconstruction and registration procedures, each image has been labelled with a six Degrees Of Freedom (6-DOF) pose summarizing the 3D position of the shopping cart, as well as its orientation in the 3D space.

Our data analysis points out that most of the variance of the collected shopping cart positions is explained by their first two principal components. This leads us to frame the egocentric shopping cart localization problem as a three Degrees Of Freedom (3-DOF) pose estimation task. Therefore, we create a 3-DOF version of the dataset by projecting the 6-DOF poses onto a 2D plane parallel to the floor of the store. In this version of the dataset, each frame is associated with the 2D coordinates and angle describing the position and orientation of the shopping cart.

Hence, we benchmark different approaches to address egocentric shopping cart localization in a retail store. Since algorithms for shopping cart (or robot) localization should be easy to deploy in embedded settings, we put an emphasis on the comparison between regression-based methods which allow for fast inference and are potentially compact, and classic approaches based on image-retrieval, which are shown to provide superior performance but require to keep the whole training set in memory and are generally slower at inference time. Our analysis shows that image representations learned using deep metric learning techniques allow to improve the results of both retrieval-based and regression-based approaches. This suggests that the gap between the two families of approaches can be bridged by enforcing that images of nearby locations are mapped close to each other in the latent feature space.

In sum, the contribution of our work is two-fold: 1) we propose a dataset to study the problem of egocentric shopping cart localization. The dataset is intended to foster research on the problem and it is publicly available at our web

Fig. 2: The hardware setup employed to collect the dataset using shopping carts.

page[1]; 2) we benchmark retrieval-based and regression-based localization techniques in the proposed application domain and hypothesize that the gap between the two approaches can be bridged by imposing a structured latent feature space.

The remainder of the paper is organized as follows: In Section 2, we review the state of the art approaches for camera localization. In Section 3, we present the proposed shopping chart localization dataset. Section 4 discusses the approaches investigated in this study. Section 5 presents the experimental setting, whereas Section 6 discusses the results. Section 7 concludes the paper and reports insights for future research.

II. Related work

Camera localization can be tackled either as a classification or camera pose estimation problem. Classification-based approaches [11], [27], [28] quantize the space into different regions (e.g., different rooms of a building) and tackle localization as an image classification problem, i.e., they assign one of the possible locations to the image under consideration. Approaches based on classification are unable to provide fine-grade positional information (e.g., accurate 2D or 3D coordinates) or to estimate the 6-DOF pose of the camera. However, they are useful in application contexts in which coarse localization is already sufficient, and computational costs need to be minimized.

Camera pose estimation approaches do not assume the space to be quantized and aim to predict the 6-DOF pose of the camera which acquired the considered image. Given a new image to be localized, the most straightforward method consists in looking for the most similar image in a labelled training set through a content-based image search. The 6-DOF label of the retrieved image is hence attached to the query image in order to obtain the 6-DOF prediction [21]–[23]. Image retrieval can be performed mapping the images in a chosen feature space and performing a nearest neighbour search [31]. Images can be encoded using hand-crafted local features (e.g. SIFT or SURF [12], [13]). Alternatively, learned representations such as Convolutional Neural Networks features can be used. Some approaches based on learned representations extract features using a model trained on a large dataset to address a different task [14]. Other methods represent images by extracting the activations of an intermediate layer of a model trained for classification/regression on the target dataset [15].

Some approaches address 3D object pose estimation (task strictly related to camera pose estimation) employing features learned using a siamese or triplet network [19], [20]. Such architectures are trained using a contrastive loss [16] on image pairs labelled either as similar or dissimilar [17], [18].

Another family of camera pose estimation methods leverage the 3D structure of the scene [4] [5] [24]. Such approaches compute matchings between 2D local features extracted from the images and the 3D points of a point cloud of the environment created beforehand. The inferred 2D-3D matchings are hence used to recover the 6-DOF camera pose [25] [26] [30]. These algorithms can reach very accurate results [24] but they are generally complex and hence inconvenient for embedded systems with low computational resources.

Recent works have investigated the possibility to regress the 6-DOF camera pose directly from images by employing Convolutional Neural Networks (CNN). The authors of [1] proposed POSENET, the first CNN-based model for end-to-end camera pose estimation. The POSENET model is obtained by replacing the classification components of the GoogleNet architecture [29] with a regressor. To balance the contribution of position prediction and orientation estimation, in [2] different loss functions are compared for the same network. In [3] a model based on the combination of a CNN and a Long-Short-Term-Memory (LSTM) architecture is proposed for camera pose regression. These methods are generally less accurate than those based on 3D models, but they have the advantage to be compact and fast to compute, which makes them appealing, especially when working in embedded settings.

III. Dataset Collection and Analysis

We collected a large-scale indoor dataset to address the task of egocentric shopping cart localization. The standard procedure for the collection of a dataset suitable for image-based localization involves the acquisition of large sets of images of the environment (this is generally achieved acquiring continuous videos) and the construction of a 3D model using Structure-from-Motion (SfM) algorithms [1].

6-DOF camera pose labels in the form of 3D coordinates and quaternions are then associated to the images which have been registered to the 3D model. In accordance with [3], we found that the implementation of this procedure is a challenging task in our settings due to the presence of repetitive structure elements (e.g., the shelves and products of the store) with nearly identical appearance which tend to create ambiguities. Most of the datasets for image-based indoor localization are generally collected in small environments typically spanning the extension of a single room. Very few studies, such as the one in [3], propose a large scale indoor dataset. In particular, the dataset proposed in [3] is acquired and labelled using a

[1] http://iplab.dmi.unict.it/EgocentricShoppingCartLocalization/

Fig. 3: Sample images from the proposed dataset. Each row shows images with similar visual elements acquired in different parts of the store.

system equipped with six high resolution cameras and three laser range finders.

To deal with the aforementioned problems and to keep a lower computational cost, we first built several 3D models of different areas of the store running SfM algorithms on subsets of the collected video dataset. The models have been hence registered together in order to obtain an overall 3D model.

The proposed dataset has been built using frames extracted from nine different videos acquired in a retail store of the south of Italy with an extension of 782 m^2. The videos have been acquired with two different zed-cameras [2] mounted on a shopping cart with their focal axes parallel to the floor of the store (see Fig. 2). Each video has been temporally sub-sampled at 3 fps. The 3D models and 6-DOF camera positions have been obtained using the 3D ZEPHIR software [6].

The overall dataset consists of $19,531$ labelled images. We divided the dataset into train and test sets. The two subsets are obtained considering respectively 6 videos training ($13,360$ frames) and 3 videos for test ($6,171$ frames). Each of the subsets contains images covering the entire store. Fig.3 shows some samples from the proposed dataset. Each row shows similar images acquired in different parts of the store. The visual similarity between the sample pairs highlights that the considered problem is a challenging one.

A. 3-DOF Labels

Since the cameras are fixed to the shopping carts, all images are acquired from a point of view which moves in accordance to the cart. This limits the number of degrees of freedom of the collected camera poses. Analysing all camera position coordinates through Principal Component Analysis (PCA), it is simple to demonstrate that almost 100% of the variability is retained within the first two principal components, which identify the position of the shopping cart on a 2D plane parallel to the floor of the store. This is depicted in Fig. 4, which shows the 2D projection of all camera positions contained in the dataset. We argue that such representation is the most appropriate for the considered application domain. Hence, we frame egocentric shopping cart localization as a 3-DOF camera pose estimation problem, where images are labelled with their 2D position and an angle indicating the orientation of the shopping cart in the 2D plane. Specifically, a camera pose is represented as $p = (\mathbf{x}, \mathbf{u})$, where $\mathbf{x} = (x, y)$ is a 2D vector representing the position of the shopping cart and $\mathbf{u} = (u, v)$ is a unit vector representing the orientation of the cart.

[2]http: www.stereolabs.com

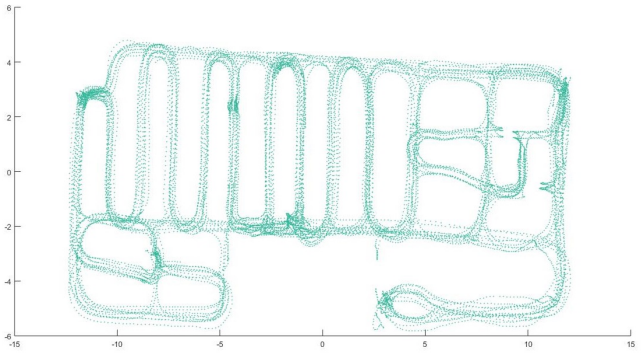

Fig. 4: 2D projection of camera positions of the proposed dataset.

B. Error Analysis

We characterize the proposed dataset by studying what is the minimum error achievable when localization is addressed as an image-retrieval task. Such value is the error committed by an optimal nearest neighbour search which always associates a test image to the most correct pose in the training set. In the considered 3-DOF scenario, a position is defined as a set of 2D coordinates measured in meters and an angle measured in degrees. Due to the different measure units, pose estimation errors needs to be computed separately for positions and orientations [1], [2]. Ideally, given a test pose p_i, we would like to be able to determine the training pose p_j such that the pair (p_i, p_j) constitutes an optimal match. To define such a matching criterion, given two 3-DOF poses p_i and p_j, we define the following parametric distance measure:

$$d(p_i, p_j; \alpha) = \alpha \cdot d_p(p_i, p_j) + (1 - \alpha) \cdot d_o(p_i, p_j) \quad (1)$$

where $d_p(p_i, p_j)$ represents the Euclidean distance between the positions of the poses p_i and p_j, $d_o(p_i, p_j)$ represents the angular distance between the orientations of the poses p_i and p_j, and α is a parameter regulating the trade-off between position and orientation errors.

By choosing a specific value for α, the parametric measure defined in Eq. (1) allows to define a criterion to perform optimal nearest neighbour matching. Given a test sample s_i with ground truth pose p_i, optimal nearest neighbour search is carried out predicting form p_i the training pose p_j such

TABLE I: Mean and median position and orientation lower-bound errors obtained with optimal nearest neighbour search.

	Mean		Median	
α	P.E.(m)	O.E.(°)	P.E.(m)	O.E.(°)
0	9.89	0.54	9.00	0.31
0.1	0.32	1.73	0.27	1.34
0.2	0.25	2.48	0.21	1.87
0.3	0.21	3.20	0.18	2.35
0.4	0.18	4.03	0.16	2.84
0.5	0.16	4.99	0.14	3.44
0.6	0.14	6.31	0.12	4.22
0.7	0.12	7.98	0.11	5.28
0.8	0.11	10.47	0.10	6.63
0.9	0.09	17.31	0.08	10.08
1	0.05	90.45	0.04	90.47

that $d(p_i, p_j; \alpha)$ is minimized. To measure the minimum error committed when choosing a given value for α, we then compute the error on position and orientation separately.

TABLE I reports the mean and median Position Errors (P.E.) and Orientation Errors (O.E.) committed by an optimal nearest neighbour search for different values of α over the whole test set. For instance, when $\alpha = 0$ optimal nearest neighbour is performed choosing the training pose which minimizes the orientation error, regardless its position. In this case, we obtain a large lower-bound error of $9.89\ m$ and a small orientation error of $0.54°$. Conversely, choosing larger values for α leads to larger orientation errors (up to $90.45°$) and lower position errors (up to $0.05\ m$).

It should be noted that the lower-bound errors reported in TABLE I have a practical interpretation. Indeed, they indicate the best performance expected by retrieval-based localization methods when a given trade-off between position and orientation is required. For instance, a method achieving a small position error close to $0.05\ m$ should perform poorly on orientation estimation, where the lower-bound mean position error is $90.45°$ (see TABLE I, last row). Similarly, a method scoring mean and median errors close to the values reported in one of the rows of TABLE I are intuitively showing good performance.

IV. Methods

We compare the performances of two different families of approaches: methods based on image retrieval and methods based on regression.

A. Retrieval-Based Methods

Methods based on image retrieval map training and test images to a chosen feature space. New test images are associated with the pose of the closest training image in the considered feature space. The amount of memory required by such approaches grows linearly with the number of training images. While we show that retrieval-based methods allow to obtain low errors, their memory footprint can easily get large, which makes them less convenient than more compact models based on regression in embedded settings.

For all our experiments, we consider a 1-NN decision rule based on the Euclidean distance, and investigate the contribution of different image representations. As a representative of shallow image representations, we consider the spatially enhanced Improved Fisher Vector based on densely extracted SIFT features described in [9]. To compute such image representations we employed a Gaussian Mixture Model with 256 components and reduced the dimensionality of SIFT descriptors to 80 components using PCA as suggested in [8]. To leverage the transfer learning capabilities of CNN features, we consider off-the-shelf 4096-dimensional features extracted from the fc7 layer of the VGG16 network pretrained on the ImageNet dataset [7]. We also explore the possibility of learning feature representations suitable for the localization task by fine-tuning the pre-trained VGG16 model using a triplet network architecture [17]. To learn a feature representation which maps images of nearby locations close to each other in the feature space, image triplets have been formed considering as "similar" each pair of training frames which poses are characterized by a spatial distance smaller than $30cm$ and an orientation distance smaller than $45°$.

To investigate whether imposing a temporal constraint can improve results of sequential camera pose estimation when processing a video, we experiment the combination of 1-NN search and the following heuristic. Given a new test frame f_i, we restrict the nearest neighbour search to training set samples which positions are in a neighbourhood of the last predicted position p_{i-1}. Restricting the nearest neighbour search as suggested above should reduce the influence of ambiguous representations in the training set. This is achieved by excluding from the search training samples which are located far away from the current location but may have a similar representation due the ambiguities arising from repeated structure elements. It should be noted that the proposed temporal constraint can only be applied when sequential localization is a feasible option. We tested different neighbourhood sizes, observing localization drifting for neighbourhood sizes smaller than $4m$. Therefore we chose $4m$ as neighbourhood sizes in our experiments.

B. Regression-Based Methods

These kind of methods address camera pose estimation as a regression problem. Such methods do not require to keep the whole training set in memory and hence they allow for fast inference and are generally more compact than retrieval-based methods.

To study the performance of compact models based on regression, we consider the POSENET architecture proposed in [1]. We adapt the POSENET architecture to perform 3-DOF camera pose estimation by requiring it to produce a 2D vector representing the 2D position of carts and a 2D unit vector representing its orientation. To train our 3-DOF version of POSENET, we set $\alpha = 125$ to weigh the position-related and orientation-related losses. In our experiments, this value allowed to obtain the best results over the following choices:

TABLE II: Mean and median position and orientation errors results

	Mean		Median	
Methods	P.E.(m)	O.E.(°)	P.E.(m)	O.E.(°)
1-NN FISHER	1.63	13.48	0.31	3.32
1-NN VGG16	0.72	7.32	**0.28**	**3.11**
1-NN TRIPLET	0.55	6.52	**0.28**	3.17
1-NN POSENET	2.17	11.53	1.38	7.07
1-NN TRIPLET TC	**0.44**	**5.76**	0.29	3.2
1-NN VGG16 TC	0.52	7.09	**0.28**	3.13
POSENET	1.62	7.52	1.23	4.63
SVR POSENET	1.96	10.1	1.54	6.14
SVR TRIPLET	1.46	8.04	0.9	4.39

$\alpha = \{500, 250, 125, 62.5\}$. The reader is referred to [1] for more details on the role of the α parameter when training the POSENET architecture. The model has been optimised using ADAM with a learning rate of $10-3$.

We also train Support Vector Regressors on images encoded according to two different feature spaces: 1) image representations learned fine-tuning an ImageNet pre-trained VGG16 model, using a triplet architecture as detailed in the previous section; 2) the internal representation learned by the POSENET architecture. Experiments with Support Vector Regressors have been conducted using an RBF kernel. The optimal values for the C and γ parameters have been found performing a grid search with cross-validation.

V. Results

TABLE II reports the mean and median position and orientation errors of the investigated methods. Methods denoted by "TC" are combined with the temporal constraint for sequential localization discussed in Section IV-A. Best per-column results are reported in bold numbers. We present the same results in graphical format in Fig. 5(a) and (b). In the plots shown in Fig. 5, each method is identified by two coordinates in the Cartesian space: the mean or median error on position (x axis) and the mean or median error on orientation (y axis). Intuitively, methods closest to the origin are the best performing. The figure also reports the different lower-bound values shown in TABLE I.

As shown in TABLE II and Fig. 5, a simple 1-NN search on images represented using off-the-shelf VGG16 features pre-trained on ImageNet allows to obtain remarkable results, both in terms of mean ($0.72\ m$ and $7.32°$) and median errors ($0.28\ m$ and $3.11°$). As it can be expected, the more expressive semantics of the hierarchical VGG16 features, allows to easily outperform classic "shallow" representation based on Fisher Vector encoding of local SIFT features (1-NN FISHER - $1.63\ m$ and $13.48°$ mean errors and $0.31\ m$ and $3.32°$ median errors). Such results can be further improved fine-tuning the VGG16 image representation with the deep metric learning criterion imposed by the triplet architecture (1-NN TRIPLET - $0.55\ m$ and $6.52°$ mean errors and $0.28\ m$ and $3.17°$ median errors). It should be noted that the triplet architecture allows to learn an embedding space in which it is explicitly enforces a desirable property for nearest neighbour search, i.e., images of nearby locations are mapped close to each other in the representation space.

Assuming that sequential localization is feasible, imposing the temporal constraint discussed in Section IV-A allows to improve 1-NN search results with both VGG16 (1-NN VGG16 TC - $0.52\ m$ and $7.09°$ mean errors and $0.28\ m$ and $3.13°$ median errors) and triplet representations (1-NN TRIPLET TC - $0.44\ m$ and $5.76°$ mean errors and $0.29\ m$ and $3.2°$ median errors). It should be noted that the results obtained imposing the temporal constraint with VGG16 features (1-NN VGG16 TC) are comparable to the ones obtained with the triplet representation without enforcing any temporal constraint. Moreover, the improvements obtained by 1-NN based on triplet when the temporal constraint is applied are modest (compare 1-NN TRIPLET with 1-NN TRIPLET TC both in TABLE II and Fig. 5). This suggests that the triplet-based learning effectively allows to reduce ambiguities in the representation space, mitigating the effect of repeated structure elements with similar appearance. We also would like to note that sequential pose estimation may not always be feasible, especially when dealing with low-power devices which can only work at a low frame rate.

Regression-based approaches generally perform worse than methods based on 1-NN search (compare top and bottom parts of TABLE II). The significant difference in performance between the "1-NN TRIPLET" and "SVR TRIPLET" methods (e.g., the mean position error increases from 0.55 to 1.46, while mean orientation error increases from 6.52 to 8.04) suggests that regression-based methods struggle to solve the inherent ambiguities in representation space, while simple 1-NN search benefits from the possibility to compare new test examples with the many different training samples which are kept in memory.

To assess whether the Triplet criterion allows to learn better image representations (as compared to regression methods such as POSENET) we also performed a 1-NN search using the internal representation of POSENET (1-NN POSENET) and trained an SVR regressor on triplet features (SVR TRIPLET). To obtain results comparable with the "SVR TRIPLET" method, we also trained an SVR regressor on top of POSENET features (SVR POSENET). Interestingly, the "1-NN POSENET" approach reports the worst results among the nearest neighbour methods ($2.17\ m$ and $11.53°$ mean errors and $1.38\ m$ and $7.07°$ median errors), while "SVR TRIPLET" outperforms all others regression-based methods ($1.46\ m$ and $8.04°$ mean errors and $0.9\ m$ and $4.39°$ median errors). These observations lead us to three conclusions: 1) feature representations learned using the regression objective imposed by the POSENET architecture are sub-optimal for retrieval-based methods. This suggests that images of nearby locations are not mapped close to each other in the learned feature space; 2) the deep metric learning criterion imposed by the triplet architecture allows to learn feature representations which improve the results of both retrieval-based and regression-based methods; 3) deep metric learning techniques could

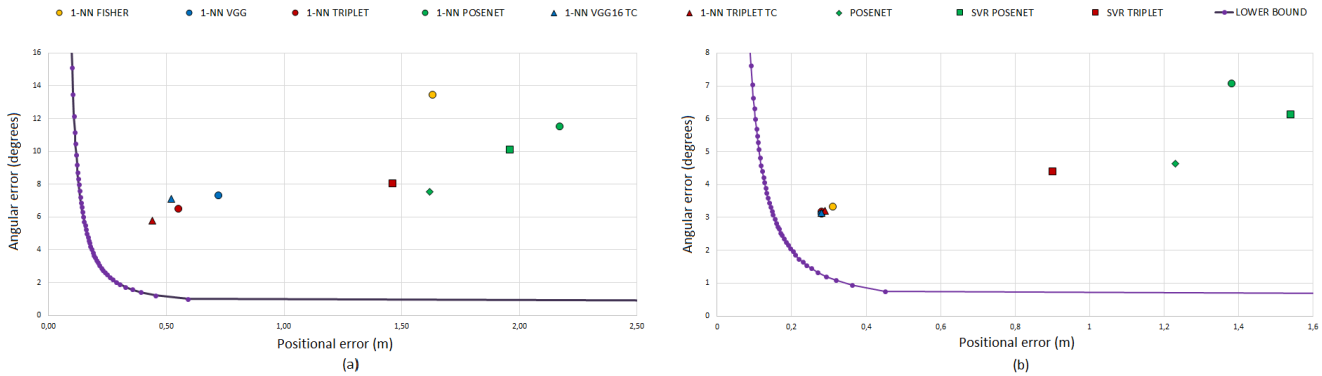

Fig. 5: Graphical representation of mean (a) and median (b) position and orientation errors of the considered methods.

help bridge the gap between retrieval-based and regression-based methods by allowing regression methods to learn more structured representations.

VI. Conclusion

In this paper, we considered the problem of shopping cart localization in retail stores. We proposed the first large scale labelled indoor dataset to address the task of image-based egocentric shopping cart localization. Following an analysis on the collected data, we framed the task as a 3-DOF camera pose estimation problem. To study the considered task, we benchmarked two families of approaches: retrieval-based techniques and regression-based methods. The considered experiments pointed out that retrieval-based approaches outperform regression-based methods. We also note that learning a feature space through deep metric learning techniques allows to improve the results of both retrieval-based and regression-based techniques.

References

[1] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A Convolutional Network for Real Time 6-DOF Camera Relocalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[2] A. Kendall, R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[3] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, D. Cremers. Image-based localization with spatial lstms. In arXiv, 2017.

[4] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In Computer Vision and Pattern Recognition (CVPR), 2013.

[5] Van Opdenbosch, G. Schroth, R. Huitl, S. Hilsenbeck, A. Garcea, and E. Steinbach. Camera-based indoor positioning using scalable streaming of compressed binary image signatures. In IEEE International Conference on Image Processing (ICIP), 2014.

[6] R. Gherardi, M. Farenzena, A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

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

[8] F. Perronnin, J. Sanchez, and T. Mensink, Improving the Fisher kernel for large-scale image classification In Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2010.

[9] J. Sanchez, F. Perronnin, T. Emdio de Campos. Modeling the spatial layout of images beyond spatial pyramids. In P.R. Letters, 2012.

[10] http://www.fellowrobots.com

[11] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. In arXiv preprint arXiv:1602.05314, 2016.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011

[13] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006.

[14] G. Baatz, K. Koser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Leveraging 3D city models for rotation invariant place-of-interest recognition. In International Journal of Computer Vision (IJCV), 2011

[15] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen. Deep learning of binary hash codes for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[16] R. Hadsell, S. Chopra, Y.LeCun. Dimensionality reduction by learning an invariant mapping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[17] E. Hoffer, N. Ailon. Deep metric learning using triplet network. In International Workshop on Similarity-Based Pattern Recognition, 2015.

[18] A. Gordo, J. Almazn, J. Revaud, and D. Larlus. Deep Image Retrieval: Learning Global Representations for Image Search. In Proceedings of the European Conference on Computer Vision (ECCV), 2016.

[19] A. Doumanoglou, V. Balntas, R. Kouskouridas, and T.-K. Kim. Siamese Regression Networks with Efficient mid-level Feature Extraction for 3D Object Pose Estimation. In arXiv preprint arXiv:1607.02257, 2016.

[20] P. Wohlhart and V. Lepetit. Learning descriptors for object recognition and 3d pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[21] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T.Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[22] A. R. Zamirand M. Shah. Accurate image localization based on google maps street view. In European Conference on Computer Vision, 2010

[23] W. Zhang and J. Kosecka. Image based localization in urban environments. In International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT), 2006.

[24] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2016.

[25] R. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. In International Journal of Computer Vision (IJCV), 1994.

[26] Z. Kukelova, M. Bujnak, and T. Pajdla. Real-Time Solution to the Absolute Pose Problem with Unknown Radial Distortion and Focal Length. In IEEE International Conference on Computer Vision, 2013.

[27] V. Santarcangelo, G. M. Farinella, S. Battiato. Egocentric Vision for Visual Market Basket Analysis. In European Conference on Computer Vision, 2016.

[28] A. Furnari, G. M. Farinella, S. Battiato. Recognizing Personal Locations From Egocentric Videos. In IEEE Transactions on Human-Machine Systems, 2017.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovi. Going deeper with convolutions.In IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[30] Y. Feng, L. Fan, Y. Wu. Fast Localization in Large Scale Environments Using Supervised Indexing of Bynary Features. In Transaction on Image Processing, 2016.

[31] D. Nistr, H. Stewnius, Scalable Recognition with a Vocabolary tree. In IEEE Conference on Computer Vision and Pattern Recognition, 2006.