CNN-BASED INITIAL LOCALIZATION IMPROVED BY DATA AUGMENTATION

M. S. Mueller*, A. Metzger, B. Jutzi

Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT) - Karlsruhe, Germany
{markus.mueller5, alexander.metzger, boris.jutzi}@kit.edu

Commission I, ICWG I/IV

KEY WORDS: Convolutional Neural Networks, Data Augmentation, Localization, Navigation, Pose Regression

ABSTRACT:

Image-based localization or camera re-localization is a fundamental task in computer vision and mandatory in the fields of navigation for robotics and autonomous driving or for virtual and augmented reality. Such image pose regression in 6 Degrees of Freedom (DoF) is recently solved by Convolutional Neural Networks (CNNs). However, already well-established methods based on feature matching still score higher accuracies so far. Therefore, we want to investigate how data augmentation could further improve CNN-based pose regression. Data augmentation is a valuable technique to boost performance on training based methods and wide spread in the computer vision community. Our aim in this paper is to show the benefit of data augmentation for pose regression by CNNs. For this purpose images are rendered from a 3D model of the actual test environment. This model again is generated by the original training data set, whereas no additional information nor data is required. Furthermore we introduce different training sets composed of rendered and real images. It is shown that the enhanced training of CNNs by utilizing 3D models of the environment improves the image localization accuracy. The accuracy of pose regression could be improved up to 69.37% for the position component and 61.61% for the rotation component on our investigated data set.

1 INTRODUCTION

Image localization or re-localization is an important and popular task in the computer vision community. In this work, we tackle the problem of such pose estimation in 6 Degrees of Freedom (DoF) by utilizing Convolutional Neural Networks (CNNs). Further, data augmentation is introduced to support the training process, whereas a 3D model of the test environment is utilized to render arbitrary images.

Localization or pose estimation is of high interest in the fields of autonomous navigation, robotics and augmented or virtual reality. Recent vehicle navigation systems obtain the localization mainly based on Global Navigation Satellite Systems (GNSSs). Multipath effects or shadowing make them vulnerable for safe and continuous navigation. Therefore, such navigation frameworks are often fused with local navigation methods, like Inertial Navigation Systems (INSs) or Visual Odometry (VO) to overcome such drawbacks. Alternative methods could increase safety and integrity of such navigation frameworks additionally. A promising extension or supplement to existing navigation networks could be introduced by utilizing Convolutional Neural Networks (CNNs).

Convolutional Neural Networks showed great success on computer vision tasks like classification, object detection, segmentation, human pose regression or image matching in the last years. In many fields, CNNs outperform conventional methods. Recently also camera pose estimation by CNNs showed promising results. After training on a set of images of a specific environment, such CNNs estimate poses from unseen query images of the same environment. Even though classic image matching performs better in most scenarios, the potential of CNNs in this field is of high interest. Besides, developments on small CNNs tackle pose regression with similar accuracy. However, the accuracy needs to be further improved to compete with existing solutions. Therefore, we introduce data augmentation on CNNs for pose regression.

*Corresponding author.

Data Augmentation could potentially improve pose regression. Improving the performance of learning-based methods and CNNs by augmenting the underlying training data sets is widely known and well established. Augmenting training data is a popular tool to overcome problems in several fields of computer vision. Such data augmentation includes the modification of existing training data as well as the simulation of purely new data to expand training sets. Common methods are to shift, rotate, scale, flip, crop, transform, compress or blur the training images to extend the training database. In this paper purely new images are rendered in the target environment to augment a training image data set. Therewith, training is carried out on a set of training images enriched with simulated images. CNNs and other learning-based methods benefit from a high variety of training data. The more different representations are included in the training samples, the more robust and accurate is a latter determination during test time. Additionally an equal distribution of training data is mandatory to train a well-adjusted network. Therefore, we introduce data augmentation to overcome recent drawbacks on pose estimation.

3D Models are utilized for our demands on data augmentation. The amount of 3D models or 3D city models raised in the last years and covers large parts of our environment nowadays. Simultaneously such models got more realistic concerning geometry and texture, and are updated more frequently. Taken this into account navigation methods for future ground and aerial vehicles could utilize this knowledge of the environment for localization and subsequent for navigational purposes. We underscore that a 3D model is not utilized directly for our demands of localization, since that would presume to provide the 3D model at runtime onboard a vehicle. This is mandatory for navigation frameworks on small Unmanned Aerial Vehicles (UAVs) or Micro Aerial Vehicles (MAVs) due to limited storage. Rather we utilize a 3D model to render additional images for training off-line. The on-board processing is restricted to the execution of the CNN. Therefore, a navigation application could run on a small on-board device, enabling autonomous navigation with no need of a ground or base

station. However, this approach is not limited to UAV-based navigation. Such data augmentation could be introduced to ground-based navigation frameworks as well. The reconstruction of the 3D environment is realized by only accessing the existing training data. Therefore, no additional data is necessary for our approach on data augmentation.

After reviewing the related work in Section 2 we focus on image-based pose regression by CNNs and feature matching in Section 3, whereas the training process of the CNNs is described subsequently in Section 3.2. In Section 4 the data utilized for this work is presented. The experiments and their results are depicted in Section 5. Subsequently the results are discussed in Section 6. Finally we conclude in Section 7 and give an outlook that provides ideas for future work and research.

2 RELATED WORK

The focus on related work is set on image-based localization, Convolutional Neural Networks to solve such localization, methods of data augmentation and how 3D models aid such augmentation.

Localization and pose estimation based on imagery are fundamental tasks in computer vision and robotics. Pose determination in a global coordinate frame is provided by correlation based methods that match aerial and UAV images to further localize the aerial vehicle (Conte and Doherty, 2011). Feature-based methods match remotely sensed data (Ma et al., 2015) or oblique images (Huachao et al., 2012). Drawbacks of such methods, like mismatches, are successfully tackled by Locality Preserving Matching (LPM) (Ma et al., 2017). Camera poses are also determined by model-based approaches (Unger et al., 2016). Moreover real-time indoor camera pose determination is solved by CAD model matching (Urban et al., 2013). Convolutional Neural Networks are utilized for matching aerial and UAV images (Altwaijry et al., 2016) or terrestrial and UAV images (Lin et al., 2015).

Besides the absolute estimation of poses in a global reference frame, local navigation methods determine relative poses in a local frame. While the lack of global referencing is a drawback, these methods still offer valuable relative positioning. Such local navigation can be conducted by visual Simultaneous Localization and Mapping (SLAM) or Visual Odometry (VO) approaches, which reconstruct a trajectory based on image sequences. State-of-the-Art solutions like ORB-SLAM (Mur-Artal et al., 2015), Large-Scale Direct Monocular SLAM (LSD-SLAM) (Engel et al., 2014) or Direct Sparse Odometry (DSO) (Engel et al., 2016) provide satisfying solutions according to accuracy and real time capability. Even though SLAM or VO solutions show impressive results and do drift only slightly for short distances, they will drift over long trajectories particularly if there are no loop closures. Additionally SLAM or VO fail if the track is lost. Restoring a lost track is impossible without moving back to a known or mapped position. Besides, latest results on CNN-based egomotion estimation provide satisfying results matching or outperforming ORB-SLAM (Mahjourian et al., 2018).

Convolutional Neural Networks recently became very popular and scored impressive results in fields of computer vision like classification or segmentation. Solving camera re-localization was successfully introduced with PoseNet (Kendall et al., 2015), a CNN to acquire a 6 Degrees of Freedom (DoF) camera pose within a known environment. For this purpose a CNN is trained with images and their corresponding poses in order to estimate the pose of an unknown image during runtime. An enhancement is the Bayesian PoseNet (Kendall and Cipolla, 2016) which provides re-localization uncertainty by adding dropout layers after each convolutional layer and improves accuracy by averaging over multiple forward passes. Enhanced accuracies in the task of estimating poses were derived by further improvements (Walch et al., 2016) using Long Short-Term Memory layers (LSTM)(Hochreiter and Schmidhuber, 1997), a type of recurrent neural net which was combined with CNNs in the past. LSTM handles the problem of a dissolving gradient during the back-propagation using so-called gates. A CNN for localization on omnidirectional images is introduced with O-CNN which finds a closest place exemplar in a data base and computes the relative distance (Wang et al., 2018). Combining RGB data and depth data in a dual stream CNN showed further improvements of the localization results (Li et al., 2017). CNNs are also capable of estimating 3D positions per pixel and subsequently estimating the cameras pose (Li et al., 2018).

Data Augmentation is a well established technique in computer vision (Gharbi et al., 2016; Lemley et al., 2017). It is shown to boost performance in fields of classification (Tu, 2005; Karpathy et al., 2014; Ng et al., 2015), segmentation (Rajpura et al., 2017), object recognition (Maturana and Scherer, 2015), object detection (Peng et al., 2015), hand gesture estimation (Molchanov et al., 2015) or human pose estimation (Rogez and Schmid, 2016). Data augmentation further supports learning based methods and CNNs to handle invariance to e.g. shift and rotation which helps to generalize and boost accuracy (Parkhi et al., 2015; Cui et al., 2015). Recently data augmentation with Generative Adversarial Networks (GANs) showed promising results (Sharma and Namboodiri, 2018). Furthermore augmenting training data by generating synthetic images is a valuable process of data augmentation. Synthetic images of text in clutter were generated to train a Fully-Convolutional Regression Network (FCRN) (Gupta et al., 2016).

3D models have a high potential to serve for data augmentation. Simulated or synthetic images rendered from 3D models have long been used in computer vision to generate extensive training data (Stark et al., 2010; Michels et al., 2005). Rendering images from 3D objects is also practiced to expand training data and improve performance of CNNs (Su et al., 2015; Gupta et al., 2015). 3D models support the learning process for deep object detectors (Peng et al., 2015) or serve for data augmentation for segmentation (Rajpura et al., 2017). Furthermore such models are utilized to augment data sets for dense 3D object reconstructions (Yang et al., 2018) or human 3D pose estimation (Rogez and Schmid, 2016). It is also shown that CNNs trained on artificial images generalize well to real images (Rogez and Schmid, 2016). In addition hand-gesture estimation is also supported by data augmentation with 3D models (Molchanov et al., 2015; Limonchik and Amdur, 2017). For our demands on pose regression a 3D model of the target environment is utilized to render images. Generating 3D models is of high interest in researches communities like photogrammetry, computer vision or geo-information sciences (Se and Jasiobedzki, 2006; Poullis and You, 2009; Ivarsson, 2014).

3D Models or images with known 6 DoF poses are the basis to train CNNs for pose regression. Therefore, pose estimation with CNNs is limited by the coverage and possible lack of training data. It was shown that pose regression in areas with less training data scores worse compared to areas with a dense distribution of training samples (Mueller et al., 2017). Utilizing a photo-realistic model for data augmentation showed improvements regarding estimation accuracy (Mueller and Jutzi, 2018). However, photo-realistic models are not as wide distributed or available as triangulated 3D model. Furthermore generating simple triangulated 3D models is a fully automated process. Therefore, it is of interest whether or not triangulated 3D models may serve for

data augmentation as well. Such models are often available for city scale environments. In addition they are simple to generate with automatic and open source Structure-from-Motion (SfM) pipelines. Researches focus on the reconstruction of such models and its automation (Singh et al., 2013; Se and Jasiobedzki, 2006; Pollefeys et al., 2000). Moreover, recently various benchmark data sets for visual localization (Kendall et al., 2015) and with varying conditions were published (Sattler et al., 2017).

3 METHODOLOGY

We apply CNN-based pose regression in Section 3.1 and describe the training procedure subsequently in Section 3.2. For comparison conventional feature matching is applied and described in Section 3.3.

3.1 Convolutional Neural Networks for pose regression

For the demands on investigating data augmentation, we utilize two different CNNs for reasons of generalized comparison. An adaption of the VGG16-Net (Simonyan and Zisserman, 2014), modified to solve for pose regression and SqueezePoseNet (Mueller et al., 2017) an adaption of SqueezeNet (Iandola et al., 2016), also modified to output poses. The original VGG16-Net and SqueezeNet are designed to solve classification tasks, whereas the modified nets solve pose regression. Both modified nets differ especially in the number of weighting parameters. The modified VGG16-Net has a model size of 527 MB and SqueezePoseNet only 3.85 MB, which is therefore roughly a hundred times smaller. Deeper or bigger networks usually tend to be more accurate than small networks (Iandola et al., 2016). Whereas the VGG16-Net mostly consists of sets of convolutional layers followed by pooling layers, SqueezeNet is built of so-called fire modules. The fire modules first decrease the number of input channels from the previous layer by 1×1 convolutions in a so-called squeeze operation. Thereafter, an expand operation that is a combination of 1×1 and 3×3 filters increases the number of activation maps while keeping the number of parameters low (Figure 1). Further for the modification of both nets a fully

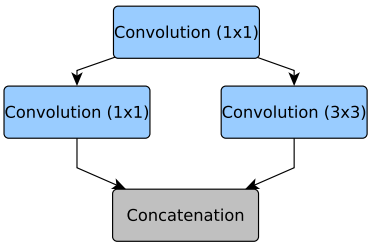


Figure 1. Architecture of a Fire module. A so-called squeeze operation is performed by the 1x1 convolutional layer. Subsequently an expand operation that is a combination of 1x1 and 3x3 filters increases the number of activation maps while keeping the number of parameters low.

connected layer and another two dense layers for actual pose determination are added to the end of the net. These layers increase the model size of the original nets enormous. SqueezeNet originally has a model size of less than 0.5 MB, whereas the addition of these layers, especially the fully connected layer, increases the model size to 3.85 MB. This is a rise of 770%. However, this is still considered as a small CNN. Furthermore the activation functions used are set to Leaky Rectified Linear Units (Leaky ReLU) (Maas et al., 2013) as this helps convergence. Additionally batch normalization (Ioffe and Szegedy, 2015) is added after each convolutional layer making higher learning rates possible. CNNs are optimized by iteratively adjusting the weighting parameters using back propagation. Therefore, the following loss function is utilized for training (Kendall et al., 2015):

$$\text{Loss}_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2 + \beta \left\|\mathbf{q}_i - \frac{\hat{\mathbf{q}}_i}{\|\hat{\mathbf{q}}_i\|}\right\|_2$$

Whereas the loss is calculated as the sum of the position error (in meters) and rotation error (in quaternions). $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$ are ground truth and estimated position. $\hat{\mathbf{q}}_i$ and $\mathbf{q}_i$ are ground truth and estimated orientation. Since position and orientation do not share the same unit space a weighting parameter β is utilized. Therefore, the CNN does not tend to optimize for only one of the two error values. The weighting parameter is helpful to scale the error for indoor and outdoor environments. Empirically β should be set between 120 to 750 for indoor and between 250 to 2000 for outdoor environments (Kendall et al., 2015). Since our environment has outdoor scale we set β to 500 in this work.

3.2 Training for CNNs

Convolutional Neural Networks usually need to be trained on a huge amount of training data to assure robust and accurate performance. The lack of training data is a major problem in many fields of learning based methods. This also applies for pose regression. Therefore, we apply transfer learning as a valuable process to overcome issues of sparse and unequally distributed training data. Since the nets are originally designed to solve for classification, the pre-existing layers – before the modification for pose regression – are initialized with weights obtained by transfer learning on the *Places* data set (Zhou et al., 2016), a data set for classification tasks. After modifying the CNNs for pose regression transfer learning is reapplied on the *Shop Façade* benchmark data set to obtain suitable initial weights for estimating poses. This re-localization data set is part of the Cambridge Landmarks data, a large urban re-localization data set with images and their labeled 6-DoF camera poses (Kendall et al., 2015). The weights obtained by transfer learning are subsequent deployed to start training on the original data set of our environment. Even though transfer learning is a valuable process to help convergence and to speed up the training process, drawbacks caused by sparse training data or simple lack of training data can not be covered by this method. Thus data augmentation is introduced to overcome recent drawbacks and improve the accuracy of pose regression. Therewith, we train on eight further data sets enhanced by rendered images of the environment. Training on these data sets is carried out with initial weights obtained by the training on the standard data set for faster convergence. However, a detailed description of all data sets is introduced in Section 4.

3.3 Feature Matching for pose estimation

A further method to estimate poses from images is the long established feature matching. We perform such feature matching for comparison to the introduced CNN-based approaches. Implicitly, we investigate feature matching based on the number of inlier matches between images from a training data set to images from evaluation data sets. However, to save computational effort and avoid matching every test image to each single training image of the training data set, we employ Bag of Visual Words (BoVW). The goal is to compare a single test image to a data base of training images. As a first step a visual vocabulary with 50 visual words is created by utilizing Speeded Up Robust Features (SURF) to extract features and descriptors from every training image. All features are clustered by using k-means with 50

clusters, whereby every cluster represents a visual word. Based on these visual words a histogram for every training image is derived. Subsequently the features and descriptors of the test images are derived and added to one of the 50 clusters by using a simple Nearest Neighbor approach. Adjacent a histogram of visual words of the test image is derived and compared to the BoVW by using histogram intersection. Therewith, the best matching images are obtained and classical feature matching between every test image and its top three closest training images is performed. As a measure of quality we take the number of inlier matches between test images and training images into account.

4 DATA SETS

The Atrium data set serves for the investigation on improving accuracy for pose regression by CNNs with data augmentation. The utilized images are part of the *LaFiDa*[1] benchmark data set (Urban and Jutzi, 2017). Figure 2 shows a side view of the Atrium as a 3D model. The dimension of the area is 39 m $\times$ 36 m $\times$ 18 m. The data set consists of 852 high resolution images[2]. The poses of these images are determined by Agisofts Structure from Motion (SfM) routine (Agisoft, 2017). Therewith, these images and their corresponding poses build the standard training data set, which we refer to as *Real* data set in this work. For evaluation two image sequences are utilized, the *medium coverage sequence* and the *low coverage sequence*. The *medium coverage sequence* contains images of the Atrium captured on ground level and spatially close to the training data. This sequence consists of 145 images, which show a medium coverage to the training data. High coverage can be stated, if training and testing data show very similar poses and share similar perspectives. The images of the *low coverage sequence* are spatially far away from the training data and have high discrepancies in perspectives compared to the training data set. This sequence contains 198 images and is captured with a higher altitude than the *medium coverage sequence*. The images show a low coverage to the training data, since the positions as well as the orientations are very different from the training poses. However, the *low coverage sequence* is a challenging evaluation data set for pose regression, which is mainly caused by the difference of training and evaluation data and the sparse distribution of images in the *Real* training data set.

Figure 2. Side view of the Atrium. The dimension of the area is 39 m $\times$ 36 m $\times$ 18 m.

For investigating data augmentation to improve the accuracy on pose regression, we add rendered images to the original training data set to aid the training process. In particular, we render images for the *medium* and the *low coverage sequence*. The *medium coverage sequence* is aided with 1,153 images rendered near the 'expected' evaluation images. The actual evaluation poses were used to interpolate images and add Gaussian noise to each pose of an image (1 m standard deviation on each position component and 0.1 on each rotation component, which is denoted in quaternions). The *low coverage sequence* was aided with 1,460 rendered images, which are generated analog to the process for the *medium coverage sequence*. The data sets with these rendered images are named *Diverge*. Figure 3a shows the poses of the rendered training images for the *medium coverage sequence*. For experimental intension, sets of images which correspond to the exact evaluation images are additionally rendered. This data sets are named *Coincide*. Furthermore we generate the *Real-Coincide* and *Real-Diverge* data sets, which are joints of the *Real* data set and the *Coincide* respectively the *Diverge* data sets. Therewith, the following types of training data sets are introduced for this work:

- *Real*: This data set contains 852 real training images from the original *Atrium* data set. A single *Real* data set is used for training and evaluated on both evaluation data sets.
- *Coincide*: This data sets contain rendered images which share the exact poses as the actual evaluation images from the evaluation sequences. The poses therefore coincide with the evaluation poses. This sets should give an idea of how well a CNN can transfer or recognize visual aspects of model images to real images.
- *Real-Coincide*: This data sets are the combination of the *Real* and the *Coincide* data sets. This sets therefore consists of real images and rendered images. Training on this data sets should give an idea if training with rendered images can be improved by adding real images.
- *Diverge*: This data sets contain rendered images with poses near the actual evaluation data. Gaussian noise is added to the original evaluation poses to create new data samples close to the original data. Since data sets like the *Coincide* or *Real-Coincide* are not applicable due to the assumption of prior knowledge of the exact evaluation poses, the *Diverge* data sets give a realistic example of actual training sets. However, when the expected trajectory of a vehicle or device is roughly known beforehand (e.g. by a given trajectory or within prior path planning), images on or near this path could be rendered a-priori and used for training. The *Diverge* data sets are utilized to simulate such a scenario. Figure 3a shows the poses of rendered images for this data set for the *medium coverage sequence*.
- *Real-Diverge*: This data sets are the combination of the *Real* and the *Diverge* data sets. This sets therefore consist of real images and rendered images. Training on this data sets should show the benefit of combining real and rendered images for training. Figure 3b shows the poses of the real and rendered images for this data set for the *medium coverage sequence*.

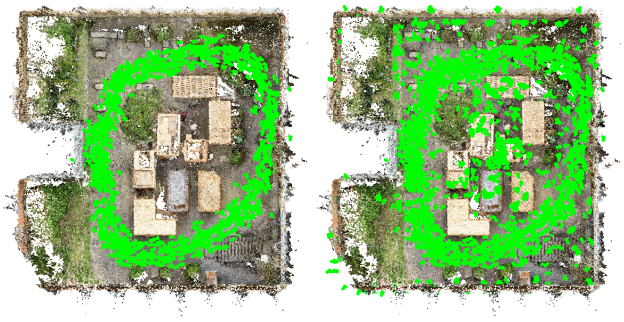
Notice, this data set types – except for *Real* – are created for each, the *medium coverage sequence* and the *low coverage sequence* leading to nine different data sets in total. An overview including the number of images separated in real and rendered images is presented in Table 1.

[1] https://www.ipf.kit.edu/lafida.php (last access 31st March 2018)

[2] https://www2.ipf.kit.edu/ pcv2016/downloads/photos_atrium_reconstruction.zip (last access 31st March 2018)

	medium coverage sequence			low coverage sequence		
	# real images	# rendered images	all images	# real images	# rendered images	all images
Real	852	–	852	852	–	852
Coincide	–	145	145	–	198	198
Real-Coincide	852	145	997	852	198	1050
Diverge	–	1150	1150	–	1458	1458
Real-Diverge	852	1150	2002	852	1458	2310

Table 1: Overview of the proposed data sets. The pre-existing *Real* data set contains 852 images of the Atrium. The *Coincide* and *Diverge* data sets are created by rendering images utilizing a 3D model created by Structure-from-Motion (SfM) on the base of the real images from the *Real* data set. The *Real-Coincide* and *Real-Diverge* data sets are joints of the fully synthethic data sets *Coincide* respectively *Diverge* and the real training images from the *Real* data set.



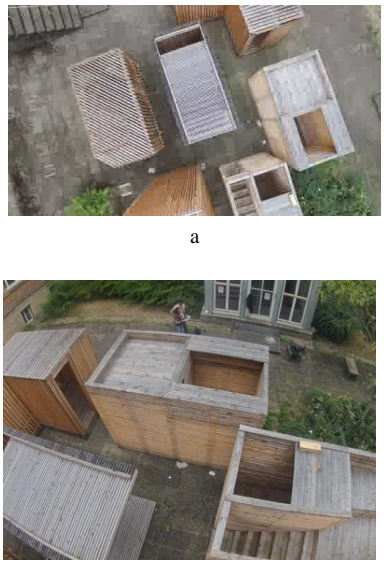
a b

Figure 3. Poses of training images of the *Diverge* data set (a) and the *Real-Diverge* data set (b) for the *medium coverage sqeuence* (depicted in green) visualized in top view of the Atrium.

5 EXPERIMENTS AND RESULTS

Experiments are carried out with the modified VGG16-Net and SqueezePoseNet to investigate the improvement of image localization by data augmentation. Both nets are trained on the different training data sets *Real*, *Coincide*, *Real-Coincide*, *Diverge* and *Real-Diverge* as introduced in Section 4 and evaluated on the *medium* respectively *low coverage sequence*. The results are depicted in Table 2. A visual representation of the evaluation errors separated by the utilized CNNs is depicted in Figure 5. The figure depicts the position and rotation errors separated by the *medium* and *low coverage sequence*.

Besides pose regression by CNNs, experiments on feature matching with a Bag of Visual Words (BoVW) approach are carried out. Therefore, test images are assigned to their top three nearest neighbors and subsequently feature matching is applied. The experiments shall expose the difficulty of the evaluation data sets regarding pose estimation. However, satisfying image matching between the training images of the *Real* data set and the evaluation data could not successfully be determined due to insufficient number of inlier matches. Explicitly the evaluation images of the *medium coverage sequence* have on average 152.2 matches between an evaluation image and its assigned nearest training images according to BoVW. After inlier test by RANSAC the confidential matches drop to 6.2 on average. The analogous test on the *low coverage sequence* shows 120.3 matches per image and 3.4 inlier on average. An overview including additionally the maximum number of matches and inlier is given in Table 3. A visualization of a test image of the *low coverage sequence* and its nearest neighbor from the training image set determined by BoVW is visualized in Figure 4. Due to wide baselines, perspective changes and low coverage sufficient image matching could not be determined. This circumstances appear all over the data sets.



a b

Figure 4. Evaluation image (a) of the *low coverage sequence* and its nearest training image (b) according to BoVW. A feature matching between these two images could not be determined sufficient due to low number of corresponding image features.

6 DISCUSSION

The results of the training processes enhanced by data augmentation show thoroughly positive outcome. The accuracy of pose estimation could be increased for the modified VGG16-Net and SqueezePoseNet on both evaluation data sets, the *medium coverage sequence* and the *low coverage sequence*. The improvements for the *medium coverage sequence* are up to 36.05% for the translation component and up to 44.74% for the orientation component. The improvements for the *low coverage sequence* are up to 69.37% for the translation component and up to 61.61% for the orientation component. Figure 6 visualizes the results obtained on the *medium coverage sequence* (6a – 6d) and the *low coverage sequence* (6e – 6h). Visualized are the results from training on the *Real* data set and corresponding the results from training on the augmented data sets with highest improvement (cf. Table 2). Figure 6a – 6d visualize the results on the *medium coverage sequence*. The figures depict the improvement and verify the numerical results from Table 2 as the pose estimates (red) move closer to the ground truth poses (blue). However, the knowledge transfer from real images to rendered images seems only moderate by the CNN due to dissimilarity in appearance. This can be stated since the *Coincide* data set, which shared the exact poses as the evaluation images did not push the accuracy too far. The same applies for the *Diverge* data set, which also contains solely rendered images. Fortunately a combination of real and rendered

	medium coverage sequence		low coverage sequence	
	VGG16-Net (modified)	SqueezePoseNet	VGG16-Net (modified)	SqueezePoseNet
Real	4.91 m, 33.30°	5.19 m, 29.28°	11.34 m, 37.33°	15.18 m, 65.02°
Coincide	3.36 m, 21.83°	5.18 m, 27.45°	**4.53 m**, **16.67**°	5.26 m, 24.90°
Real-Coincide	3.37 m, 19.63°	**3.91 m**, **19.01**°	4.46 m, 20.90°	6.6 m, 31.88°
Diverge	3.90 m, 21.79°	5.32 m, 25.99°	4.48 m, 26.76°	**4.65 m**, **24.96**°
Real-Diverge	**3.14 m, 18.40**°	3.89 m, 19.90°	6.38 m, 19.02°	6.86 m, 26.43°
Improvement	36.05%, 44.74%	24.66%, 35.08%	60.05%, 55.34%	69.37%, 61.61%

Table 2: Position and rotation evaluation errors on the *medium* and *low coverage sequence*. The improvement corresponds to the best result per CNN, which is marked in bold.

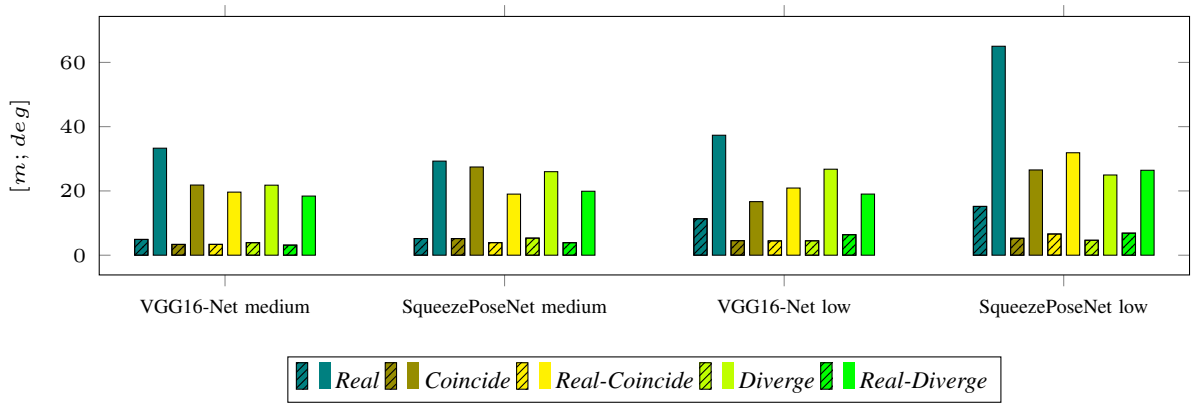


Figure 5. Visualization of mean evaluation errors. The Figure depicts the position (striped bars) and rotational (untextured bars) evaluation errors separated by the *medium* and *low coverage sequence* evaluated on the modified VGG16-Net and SqueezePoseNet. The experiments were carried out on the different training data sets *Real*, *Coincide*, *Real-Coincide*, *Diverge* and *Real-Diverge*. The striped bars show the mean position error (meter), the untextured bars show the mean rotational error (degree).

Figure 6. Visualization of estimated poses (red) and ground truth poses (blue). a) VGG16-Net trained on the *Real* data set. b) VGG16-Net trained on the *Real+Diverge* data set. c) SqueezePoseNet trained on *Real* data set. d) SqueezePoseNet trained on *Real-Coincide* data set. e) VGG16-Net trained on the *Real* data set. f) VGG16-Net trained on the *Coincide* data set. g) SqueezePoseNet trained on *Real* data set. h) SqueezePoseNet trained on *Diverge* data set.

Feature matching	*medium coverage sequence*	*low coverage sequence*
# matches (max)	254	208
# matches (avg)	152.2	120.3
# inlier (max)	38	13
# inlier (avg)	6.2	3.4

Table 3: Evaluation of feature matching on the *medium* and *low coverage sequence*. The number of matches by SURF and inlier matches between test images and nearest training images according to BoVW from the *Real* data set are depicted. The test images are assigned to their nearest training images by BoVW. On average 152.2 respectively 120.3 matches between a test image and its nearest training image could be found. However, the number of average inlier with 6.2 respectively 3.4 is unsatisfying for subsequently image matching.

images improves the accuracy further, leading to clearly better results than training on real images only. Reviewing the numerical results on the *low coverage sequence* in Table 2 shows also clear improvement. Whereas training on the *Real* data set scored insufficient results, the training on the proposed data sets by data augmentation improved the pose determination. However, by visualizing the pose estimates (Figure 6e – 6h) it is shown that the poses are still not determined satisfactorily. The numerical improvement is mainly caused by the fact of better distributed training data as the estimated image poses are shifted to the center of the actual evaluation poses.

Even though the evaluation sequences seem to represent odometry data, we investigate single pose determination in this work and make no use of image sequence analysis methods.

7 CONCLUSION AND OUTLOOK

In this work we show that CNN-based pose regression can benefit from data augmentation. The method shows promising results on two tested CNNs and on different data sets. In future research a transfer to state-of-the-art CNNs is of high interest. Additionally the investigation of complex environments with sparse training data should be further tackled to assure robust pose estimation by CNNs in arbitrary environments. In addition an integration to a navigation filter and carrying out test flights with an Unmanned Aerial Vehicle would be of interest. However, the absolute accuracy has to be improved afore to operate an UAV by such CNN solutions.

References

Agisoft, 2017. Agisoft LLC, http://www.agisoft.com/, (last access 31th March 2018).

Altwaijry, H., Trulls, E., Hays, J., Fua, P. and Belongie, S., 2016. Learning to match aerial images with deep attentive architectures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3539–3547.

Conte, G. and Doherty, P., 2011. A visual navigation system for uas based on geo-referenced imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Cui, X., Goel, V. and Kingsbury, B., 2015. Data augmentation for deep neural network acoustic modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(9), pp. 1469–1477.

Engel, J., Koltun, V. and Cremers, D., 2016. Direct sparse odometry. arXiv:1607.02565.

Engel, J., Schöps, T. and Cremers, D., 2014. LSD-SLAM: Large-Scale Direct Monocular SLAM. In: D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars (eds), Computer Vision – ECCV 2014, Lecture Notes in Computer Science, Springer International Publishing, pp. 834–849. DOI: 10.1007/978-3-319-10605-2_54.

Gharbi, M., Chaurasia, G., Paris, S. and Durand, F., 2016. Deep joint demosaicking and denoising. ACM Trans. Graph. 35(6), pp. 191:1–191:12.

Gupta, A., Vedaldi, A. and Zisserman, A., 2016. Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324.

Gupta, S., Arbeláez, P., Girshick, R. and Malik, J., 2015. Inferring 3d object pose in rgb-d images. arXiv preprint arXiv:1502.04652.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation 9(8), pp. 1735–1780.

Huachao, Y., Shubi, Z. and Yongbo, W., 2012. Robust and precise registration of oblique images based on scale-invariant feature transformation algorithm. IEEE Geoscience and Remote Sensing Letters 9(4), pp. 783–787.

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. arXiv:1602.07360 [cs]. arXiv: 1602.07360.

Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of The 32nd International Conference on Machine Learning, pp. 448–456.

Ivarsson, C., 2014. Combining street view and aerial images to create photo-realistic 3d city models.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Kendall, A. and Cipolla, R., 2016. Modelling uncertainty in deep learning for camera relocalization. In: Robotics and Automation (ICRA), 2016 IEEE International Conference on, IEEE, pp. 4762–4769.

Kendall, A., Grimes, M. and Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision, pp. 2938–2946.

Lemley, J., Bazrafkan, S. and Corcoran, P., 2017. Smart augmentation learning an optimal data augmentation strategy. IEEE Access 5, pp. 5858–5869.

Li, R., Liu, Q., Gui, J., Gu, D. and Hu, H., 2017. Indoor relocalization in challenging environments with dual-stream convolutional neural networks. PP(99), pp. 1–12.

Li, X., Ylioinas, J. and Kannala, J., 2018. Full-Frame Scene Coordinate Regression for Image-Based Localization. arXiv:1802.03237 [cs]. arXiv: 1802.03237.

Limonchik, B. and Amdur, G., 2017. 3d model-based data augmentation for hand gesture recognition.

Lin, T.-Y., Yin Cui, Belongie, S. and Hays, J., 2015. Learning deep representations for ground-to-aerial geolocalization. IEEE, pp. 5007–5015.

Ma, J., Zhao, J., Guo, H., Jiang, J., Zhou, H. and Gao, Y., 2017. Locality preserving matching. In: International Joint Conference on Artificial Intelligence, pp. 4492–4498.

Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J. and Tian, J., 2015. Robust feature matching for remote sensing image registration via locally linear transforming. IEEE Transactions on Geoscience and Remote Sensing 53(12), pp. 6469–6481.

Maas, A. L., Hannun, A. Y. and Ng, A. Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, Citeseer.

Mahjourian, R., Wicke, M. and Angelova, A., 2018. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. arXiv preprint arXiv:1802.05522.

Maturana, D. and Scherer, S., 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE, pp. 922–928.

Michels, J., Saxena, A. and Ng, A. Y., 2005. High speed obstacle avoidance using monocular vision and reinforcement learning. In: Proceedings of the 22nd international conference on Machine learning, ACM, pp. 593–600.

Molchanov, P., Gupta, S., Kim, K. and Kautz, J., 2015. Hand gesture recognition with 3d convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 1–7.

Mueller, M. S. and Jutzi, B., 2018. UAS Navigation with SqueezePoseNet—Accuracy Boosting for Pose Regression by Data Augmentation. Drones, 2(1), 7.

Mueller, M. S., Urban, S. and Jutzi, B., 2017. SqueezePoseNet: Image based pose regression with small convolutional neural networks for real time uas navigation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4, pp. 49.

Mur-Artal, R., Montiel, J. M. M. and Tardós, J. D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics 31(5), pp. 1147–1163.

Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G., 2015. Beyond short snippets: Deep networks for video classification. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE, pp. 4694–4702.

Parkhi, O. M., Vedaldi, A. and Zisserman, A., 2015. Deep face recognition. In: BMVC, Vol. 1number 3, p. 6.

Peng, X., Sun, B., Ali, K. and Saenko, K., 2015. Learning deep object detectors from 3D models. In: Computer Vision (ICCV), 2015 IEEE International Conference on, IEEE, pp. 1278–1286.

Pollefeys, M., Koch, R., Vergauwen, M. and Van Gool, L., 2000. Automated reconstruction of 3d scenes from sequences of images. ISPRS Journal of Photogrammetry and Remote Sensing 55(4), pp. 251–267.

Poullis, C. and You, S., 2009. Photorealistic large-scale urban city model reconstruction. IEEE transactions on visualization and computer graphics 15(4), pp. 654–669.

Rajpura, P., Goyal, M., Hegde, R. and Bojinov, H., 2017. Dataset augmentation with synthetic images improves semantic segmentation. arXiv preprint arXiv:1709.00849.

Rogez, G. and Schmid, C., 2016. Mocap-guided data augmentation for 3d pose estimation in the wild. In: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (eds), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., pp. 3108–3116.

Sattler, T., Maddern, W., Torii, A., Sivic, J., Pajdla, T., Pollefeys, M. and Okutomi, M., 2017. Benchmarking 6dof urban visual localization in changing conditions. arXiv preprint arXiv:1707.09092.

Se, S. and Jasiobedzki, P., 2006. Photo-realistic 3d model reconstruction. In: Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on, IEEE, pp. 3076–3082.

Sharma, S. and Namboodiri, V. P., 2018. No Modes left behind: Capturing the data distribution effectively using GANs. arXiv:1802.00771 [cs]. arXiv: 1802.00771.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Singh, S. P., Jain, K. and Mandla, V. R., 2013. Virtual 3d City Modeling: Techniques and Applications. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2), pp. 73–91.

Stark, M., Goesele, M. and Schiele, B., 2010. Back to the future: Learning shape models from 3d cad data. In: Bmvc, Vol. 2number 4, Citeseer, p. 5.

Su, H., Qi, C. R., Li, Y. and Guibas, L. J., 2015. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2686–2694.

Tu, Z., 2005. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, Vol. 2, IEEE, pp. 1589–1596.

Unger, J., Rottensteiner, F. and Heipke, C., 2016. Integration of a generalised building model into the pose estimation of uas images. In: 23rd International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences Congress, ISPRS 2016, 12–19 July 2016, Prague, Czech Republic.

Urban, S. and Jutzi, B., 2017. Lafida—a laserscanner multi-fisheye camera dataset. Journal of Imaging 3(1), pp. 5.

Urban, S., Leitloff, J., Wursthorn, S. and Hinz, S., 2013. Self-Localization of a Multi-Fisheye Camera Based Augmented Reality System in Textureless 3D Building Models. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Science 2, pp. 43–48.

Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S. and Cremers, D., 2016. Image-based localization with spatial lstms. arXiv:1611.07890.

Wang, T.-H., Huang, H.-J., Lin, J.-T., Hu, C.-W., Zeng, K.-H. and Sun, M., 2018. Omnidirectional cnn for visual place recognition and navigation. arXiv preprint arXiv:1803.04228.

Yang, B., Rosa, S., Markham, A., Trigoni, N. and Wen, H., 2018. 3d Object Dense Reconstruction from a Single Depth View. arXiv:1802.00411 [cs]. arXiv: 1802.00411.

Zhou, B., Khosla, A., Lapedriza, A., Torralba, A. and Oliva, A., 2016. Places: An image database for deep scene understanding. arXiv:1610.02055.