Deep Learning of Convolutional Auto-encoder for Image Matching and 3D Object Reconstruction in the Infrared Range

Vladimir A. Knyaz[1,2], Oleg Vygolov[1], Vladimir V. Kniaz[1,2], Yury Vizilter[1], Vladimir Gorbatsevich[1]
[1] State Res. Institute of Aviation Systems (GosNIIAS), 7 Victorenko str., Moscow, Russia
[2] Moscow Institute of Physics and Technology (MIPT), Russia
{knyaz, o.vygolov, vl.kniaz, viz, gvs}@gosniias.ru

Thomas Luhmann, Niklas Conen
Jade University of Applied Sciences Oldenburg
Ofener Straße 16/19, Oldenburg, Germany
{thomas.luhmann, niklas.conen}@jade-hs.de

Abstract

Performing image matching in thermal images is challenging due to an absence of distinctive features and presence of thermal reflections. Still, in many applications, infrared imagery is an attractive solution for 3D object reconstruction that is robust against low light conditions. We present an image patch matching method based on deep learning. For image matching in the infrared range, we use codes generated by a convolutional auto-encoder. We evaluate the method in a full 3D object reconstruction pipeline that uses infrared imagery as an input. Image matches found using the proposed method are used for estimation of the camera pose. Dense 3D object reconstruction is performed using semi-global block matching. We evaluate on a dataset with real and synthetic images to show that our method outperforms existing image matching methods on the infrared imagery. We also evaluate the geometry of generated 3D models to demonstrate the increased reconstruction accuracy.

1. Introduction

Object detection in the infrared range proves to be a robust solution for such applications as pedestrian detection [22, 21], face recognition [49, 56] and autonomous driving [53]. Still, 3D object pose estimation and model reconstruction with infrared images are challenging due to low image contrast or absence of feature points.

Despite these disadvantages, thermal cameras have several benefits that make them attractive for 3D object reconstruction and pose estimation. Firstly, they are robust against degraded visual environments such as dust, fog, and low light conditions. Secondly, infrared cameras are used for 3D reconstruction of objects that have a distinctive texture only in the infrared range, *e.g.* an aerial survey of geysers [33]. Finally, rich, multi-view, multispectral infrared image datasets are highly demanded nowadays to train deep learning based object recognition algorithms [60, 53, 2]. While such datasets are readily available for the visible spectrum [34, 9, 13, 55] only a few small datasets for the infrared range [60, 22, 53, 3, 27] can be found to date. One way to easily obtain large datasets with infrared imagery is to generate it synthetically using reconstructed 3D models with real infrared textures.

1.1. 3D reconstruction and thermal imaging

3D object reconstruction techniques such as Structure from Motion (SfM) [40], simultaneous localization and mapping (SLAM) [7], Semi-global Matching (SGM) [19, 4], silhouette-based 3D reconstruction [48] and Shape from Interaction [38] prove to be fast and robust techniques for 3D model generation from the imagery captured in the visible range. SfM requires sparse image matching using keypoint descriptors for preliminary orientation of each image. Evaluation of SfM on the infrared imagery shows that commonly used key point descriptors like SIFT [36] or SURF [1] fail to obtain feature points [15, 57]. Still, SfM provides a convenient pipeline for generation of digital elevation models with thermal textures [33]. Another approach for scene reconstruction with a thermal camera is an LSD-SLAM algorithm [7]. A recent evaluation showed that it also fails to match the infrared imagery due to low image contrast [57]. Thus a feature extraction method that is robust to low contrast details is required for 3D object reconstruction in the thermal range.

Recently, new approaches for feature matching based on deep learning methods [43, 26, 10, 25] have demonstrated excellent performance.The patch matching problem could not be solved directly by image classification using deep neural networks as the number of possible image patches is unlimited. In [26] it is proposed to use a convolutional auto-encoder (CAE) to overcome this problem. The CAE is trained under an unsupervised learning approach to compress an image patch into a low dimensional code and restore the original image from that code. If a good restoration quality is achieved, the CAE has learned to extract the most informative bits of information from the original image. Hence, the code could be used to perform sparse image matching.

In this paper, we propose a new method for image patch matching based on a CAE using an approach inspired by [26]. We evaluate our method in a full 3D object reconstruction pipeline and use infrared images as an input (fig. 1). Firstly, we perform feature matching using a CAE codebook. Secondly, we use image patch correspondences to perform estimation of the camera pose. Finally, we use SGM for dense point cloud reconstruction.

We evaluate our method on real data to show that it outperforms previous methods such as SIFT [36] and other deep convolutional feature point descriptors [43, 58] on the infrared images. To perform the evaluation, we created a multi-view stereo infrared imagery dataset (MVSIR) with accurate ground truth 3D models of test objects and a camera calibration data. The dataset could be used for evaluation of 3D object reconstruction methods on infrared imagery and training of feature matching algorithms on patches of thermal images. The MVSIR dataset is publicly available from `http://www.zefirus.org/mvsir17/`.

We provide a comparison of the accuracy of 3D models generated using CAE-based image matching and SGM and well-established 3D object reconstruction techniques to show that our matching method helps to achieve a better accuracy on the complex infrared imagery.

1.2. Contributions

In this paper, we present three key technical contributions:

1. New image patch matching method based on deep learning,
2. Pipeline for 3D reconstruction from thermal images based on the proposed patch matching,
3. Thermal image dataset MVSIR with ground truth data for evaluation of 3D reconstruction quality.

First two contributions achieve state-of-the-art results on the created thermal image dataset. Our main contribution is a new image patch matching method based on deep learning. The method uses a CAE to build unique patch codes that condense discriminative features of an image patch. To perform the matching an additional table (codebook) is used. The codebook defines correspondence between CAE code and patch class ID, that defines an abstract patch type (blob, line, etc.)

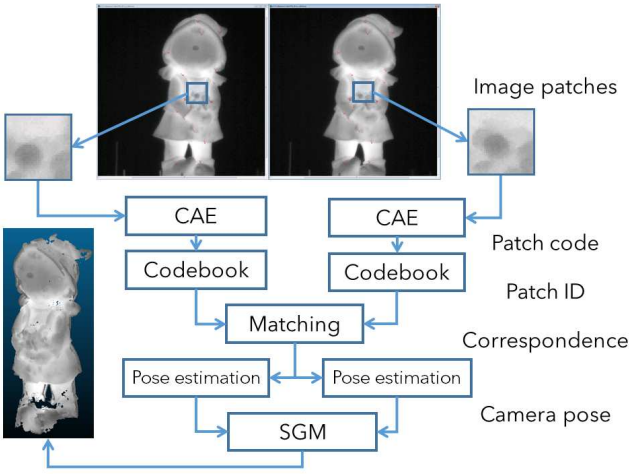


Figure 1. Infrared image matching and 3D reconstruction using the CAE. We detect feature points and create local image patches. We find patch correspondences using the CAE codebook trained for abstract image patch classes. Patch point correspondences are used for camera pose estimation. We perform dense 3D point reconstruction using semi global matching.

2. Related work

3D object reconstruction from imagery has a history of more than 50 years. In recent years an intense research activity is focused on 3D object reconstruction, pose estimation and scene understanding with a monocular camera. Robust image matching is an essential element of most of the proposed approaches. Most of the modern software for 3D object reconstruction use approaches based on analytically developed feature descriptors [36, 1].

The availability of low-cost thermal cameras stimulated an active research in the field of 3D object reconstruction and pose estimation in the infrared range [16, 52, 39, 33, 24, 57, 42, 44]. An evaluation of hand-crafted feature descriptors on the infrared imagery [16] outlined complexities of image matching in the infrared range such as infrared reflections, infrared halo effects, saturation and history effects. Still, the research proved that the 3D object reconstruction and image matching in the infrared range are possible. The combination of Harris detector [17] and normalized correlation for image matching demonstrated the best performance among classic image matching approaches.The evaluation

of methods that do not use feature points such as the LSD-SLAM [7, 57] showed that they also could not recover scene geometry due to lack of contrast in features in thermal images. Thus, the main problem of 3D reconstruction and pose estimation in the infrared range is the poor performance of existing image matching methods on the thermal imagery.

Standard approaches like SfM compute the parameters of interior and exterior orientation of the camera based on the robust matching of keypoints. Hence, if the object does not provide sufficient texture, the quality of orientation and calibration drops significantly. As an alternative, thermal images can be oriented by means of given 3D control points, *e.g*. targeted in a way that they can be measured in thermal imagery as well. As shown by [37], a thermal camera can be calibrated with high accuracy and used for subsequent 3D reconstruction in a classical photogrammetric workflow.

Image matching methods that use finite object planes such as plane sweep matching or PatchMatch [11, 6, 12] seem to be robust on low-textured areas. Still, such methods require diffuse Lambertian reflection properties of the observed surface and regularization conditions to provide smoothness between adjacent local planes. To our knowledge, there is no previous work for applying these methods to the thermal imagery.

Recent research on image matching has proven that deep learning based feature descriptors [54, 43, 26] outperform their hand-crafted predecessors in matching accuracy. Deep learning based descriptors can be divided into two broad groups. The first group is based on classical deep convolutional neural networks (CNN) for image classification [30, 45]. To perform matching top layers of the network are removed. The output of the remaining layers is used as a code to find feature correspondences.

The second group of deep learning based descriptors is based on the unsupervised learning approach. As the number of possible image points in a dataset could reach billions of classes, it is often impossible to choose good classes at the training stage. In [26] it is proposed to use CAE to overcome this difficulty. The usage of CAE for 6D pose estimation with RGB-D data have shown the state-of-the-art results on various datasets. The other benefit of CAE is their robustness to previously unseen data. All in all, deep learning based architectures provide a robust solution for local patch matching that can adapt to arbitrary kind of features and spectral range.

The final stage of modern 3D object reconstruction pipelines is the dense image matching and point cloud generation. Most of the well-known matching algorithms rely on some features based on pixel brightness. Silhouette image features [8] and volumetric graph-cut based approaches [51] require sharp edges on image for robust performance. Hence, they could not be directly applied for dense image matching in the infrared range. SGM algorithm has proven to be a robust solution for dense image matching in images. As SGM methods provide a reliable performance on low or non-textured image areas, they seem to be a promising solution for dense image matching in infrared imagery.

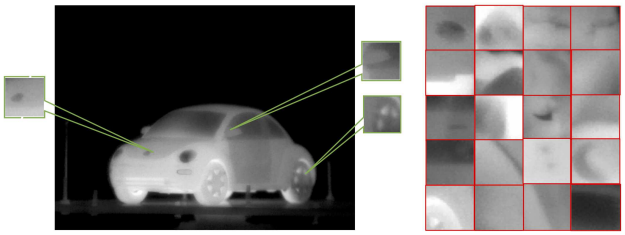


Figure 2. Example of patch selection from an infrared image.

3. Methodology

This section presents all stages of the proposed 3D object reconstruction pipeline. Firstly, we discuss the proposed CAE for local feature representation and the training process. Secondly, we present feature correspondence based camera orientation estimation. Finally, we discuss the 3D model generation using SGM.

3.1. Local Patch Extraction

A standard approach for local patch selection from an image is based on feature point detectors. However, they perform unstable on infrared imagery [16]. To obtain the local patch representation, we use a uniform sampling of the image (fig. 2). Usually, the resolution of thermal cameras is lower than for cameras of the visible range. Hence, the patch has to be small in pixel dimensions to be invariant to perspective transformations. Inspired by the previous research [26] with RGB-D cameras of comparable resolution we have selected the patch size of 28×28 pixels.

3.2. Convolutional Auto-encoder

An auto-encoder (AE) can be considered a special case of a feed-forward neural network. AE accepts some input $\boldsymbol{x}$ and attempts to copy it at its output $\boldsymbol{y}$ [14]. The network consists of two main parts. The first part is an encoder function $\boldsymbol{h} = f(\boldsymbol{x})$ that condenses the input $\boldsymbol{x}$ to a hidden layer $\boldsymbol{h}$ that produces a code F containing all values required to perform the reconstruction of the input. The second part is a decoder that attempts to reconstruct the input $\hat{\boldsymbol{y}} = g(F)$. Since the code F has a lower dimension than the original image, during the training AE tries to capture the most salient features of the training dataset. After the training stage, the output of a hidden layer could be used as a code F. As the code condenses the most discriminative features of the input patch it could be used to perform an efficient search for a corresponding image patch. Recent research [47, 26] have shown that convolutional layers

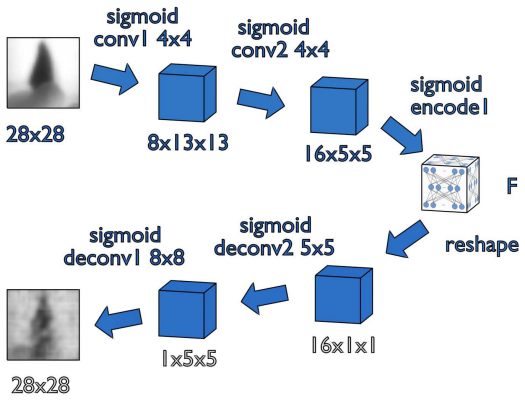


Figure 3. Architecture of the CAE.

increase the quality of AE reconstruction. AE with convolutional and deconvolutional layers are commonly called convolutional auto-encoders.

To develop an effective CAE architecture for infrared image matching, we have considered following qualities of the input data. Firstly, the CAE has to work with relatively small image patches due to the low resolution of infrared sensors. Secondly, it should have a small number of learning parameters to obtain good convergence properties on small datasets of infrared image patches.

Firstly, we have experimented with an architecture proposed in [47]. While it demonstrated an excellent performance on MNIST dataset [31], training on more complex datasets showed that it tends to converge to the mean value of the dataset. After analysis of training results, we concluded that architecture [47] did not have enough training parameters and decided to increase the number of convolutional filters. We adopted the base architecture proposed in [41]. Our contribution to the architecture was the following. Firstly, we scaled convolutional and deconvolutional layers to achieve the target image size of 28×28 pixels. Secondly, we replaced $\tanh$ activation layers with sigmoid layers to obtain more effective error back propagation [32] and increase the stability of training on datasets with low-textured images. We have experimented with three dimensions of code F to find a compromise between the reconstruction quality and the compression ratio. The final architecture of the CAE network is presented in figure 3 and table 1.

We also have experimented with two loss functions. A classic Euclidean distance based loss is not robust to CNN designs with deconvolutional layers [61]. We use cross-entropy loss (logistic loss) given by

$$L_{sc} = -\frac{1}{n} \sum_{i \in (w,h)} \Big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big), \quad (1)$$

Layer	Size out	Kernel	Stride
Input	$1 \times 28 \times 28$		
Convolution	$8 \times 13 \times 13$	4×4	2
Convolution	$16 \times 5 \times 5$	4×4	2
Inner Product	F		
Deconvolution	$8 \times 5 \times 5$	5×5	2
Deconvolution	$1 \times 28 \times 28$	8×8	5

Table 1. CAE architecture. F is the size of the trained feature code.

where w, h are layer dimensions, y is the pixel value of the target image at point (x, y), and $\hat{y}$ is the pixel value of the image reconstructed by CAE at point (x, y). We found out that an addition of a Gaussian noise just before the final sigmoid layer as proposed in [14] increases the stability of the training process. The standard deviation of the Gaussian noise uniformly increases from 0 to 0.5 during the training process.

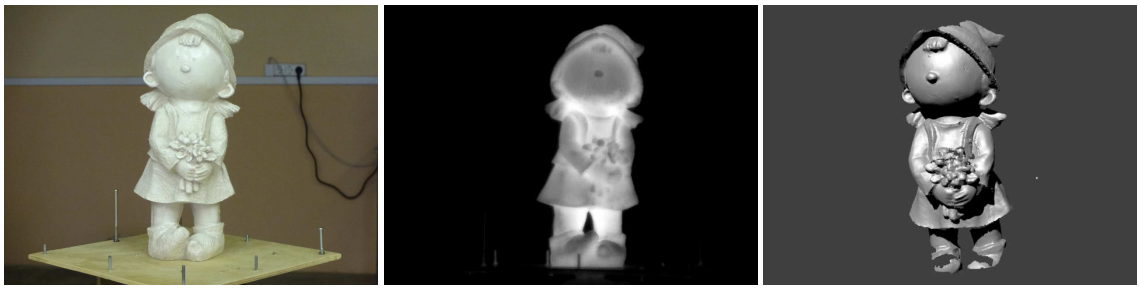
3.3. Training dataset

CAE training requires a large dataset with a high variance of random image patches. To train the developed CAE architecture, we generated a multi-view stereo infrared dataset (MVSIR) with 1 million of image patches of different feature points. We varied emission intensity with random uniform noise. The dataset includes synthetic image patches of three test objects. Test objects from the dataset are presented in figure 4. 3D models of objects were generated using a fringe projection 3D scanner [28] with the accuracy of 0.1 mm. Real infrared textures were captured using the FLIR P640 thermal camera with a lens of 130 mm.

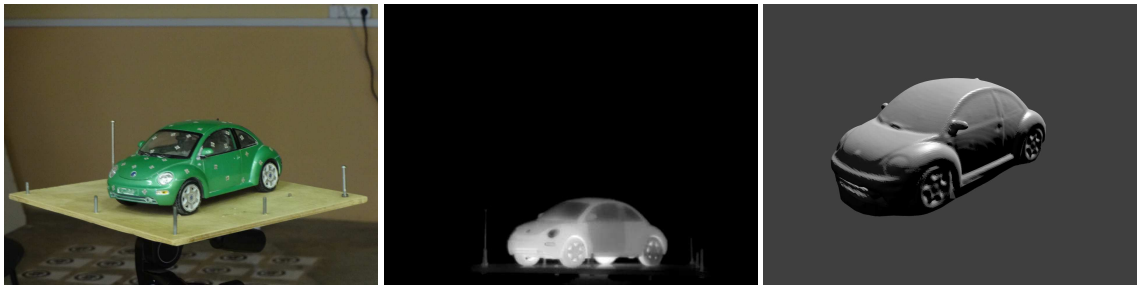
To create the training dataset we used a technique similar to [18]. We sample image patches by placing a virtual camera on an icosahedron and pointing it to a feature point. Feature point locations were selected by detecting distinctive feature using the Harris corner detector on original infrared textures. 3D coordinates of the feature points were obtained by back projection to 3D space. To perform patch matching, we assign each 3D feature point on a test object a unique patch ID. The patch ID is stored in the codebook with the patch code F generated by CAE. All in all, we sample a set of 3D points on test objects and assign each 3D point an unique patch ID. For each patch ID, we generated 7000 image patches sampled from different viewpoints.

3.4. Patch matching

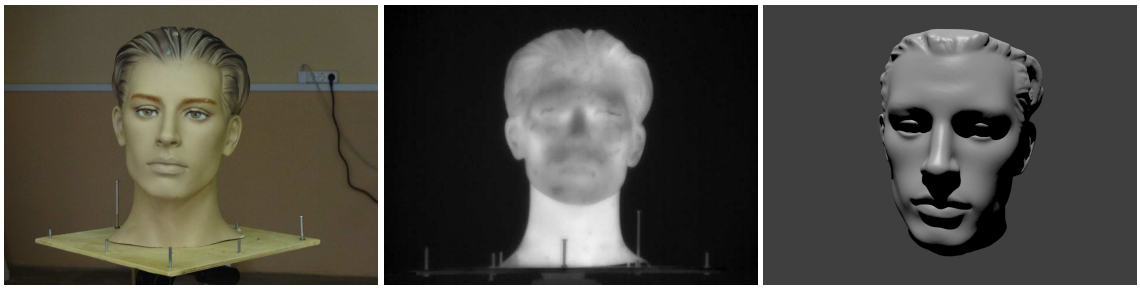
To perform patch matching, we prepare the codebook that establishes correspondences between the patch code and the patch ID of a unique 3D point on a test object. To generate the codebook, we process all image patches from the training dataset using the CAE to receive the code F.



(a) **Gnome** Visible (b) **Gnome** Infrared (c) **Gnome** Ground truth 3D model



(d) **Car** Visible (e) **Car** Infrared (f) **Car** Ground truth 3D model



(g) **Head** Visible (h) **Head** Infrared (i) **Head** Ground truth 3D model

Figure 4. Examples of test objects, infrared imagery and ground truth 3D models from the dataset.

We perform matching using a voting-based approach. For a given patch $\boldsymbol{I}$, we generate a code F using the CAE. After that, we query n nearest neighbors from the codebook. The patch ID $d(\boldsymbol{I})$ is defined by the patch ID of a majority vote of its neighbors. To filter false correspondences, we define the probability p that the patch $\boldsymbol{I}$ has a patch ID z, as a ratio of majority count to the selected number of nearest neighbors n

$$p = P(d(\boldsymbol{I}) = z) = \frac{|\{b \in B : b = z\}|}{n}, \qquad (2)$$

where B is a set of path IDs of nearest neighbors.

Let $\boldsymbol{I}_1$, $\boldsymbol{I}_2$ be two image patches to be matched. Then the probability p_{pair} that they have a similar patch ID z is defined as follows

$$p_{pair} = \begin{cases} P(d(\boldsymbol{I}_1) = z) \cdot P(d(\boldsymbol{I}_2) = z), & d(\boldsymbol{I}_1) = d(\boldsymbol{I}_2) \\ 0, & \text{otherwise} \end{cases}. \qquad (3)$$

While voting approach demonstrated a good performance for pose estimation of a known object [26], its usage for stereo correspondences matching requires a large codebook with patches of 3D points similar to 3D points of the selected object. To find out the robustness of our matching method to previously unseen data we perform matching using the codebook from a different test object.

4. Camera calibration and pose estimation

To perform 3D object reconstruction, we use patch codes produced by CAE. The patch matching is performed using nearest neighbor search. We filter correct matches using threshold and use them to estimate the camera pose (camera external orientation parameters).

4.1. Camera calibration

An accurate camera calibration is required as a preliminary stage of 3D object reconstruction. We calibrate the camera using a camera model in a form [5] and an original

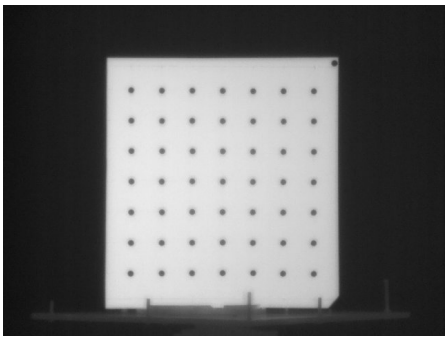


Figure 5. Testfield used for the infrared camera calibration.

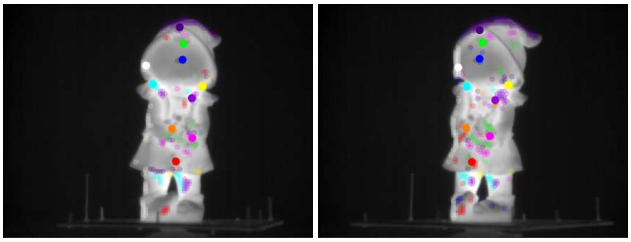


Figure 6. Image orientation. Point correspondences for ten random patch classes are shown. Patch ID probability p is shown using points transparency.

software [29]. We use a planar testfield (fig. 5) with known spatial coordinates of reference points to perform the calibration.

For calibration 20 images were acquired. The root-mean square error at reference points for the estimated parameters of the camera model was about 0.1 mm. The estimated interior parameters of the camera were used for camera pose estimation and 3D reconstruction.

4.2. Pose estimation

To determine the camera pose, we use the corresponding points which were roughly found using the CAE. While the CAE allows us to find corresponding features in two images, the accuracy of points coordinates is not sufficient for an accurate pose estimation. We refine the initial coordinates estimates to sub-pixel accuracy using the correlation technique. The results of image patches matching and subpixel measuring are shown in figure 6.

For camera poses estimation we perform robust nonlinear minimization of the measurement (re-projection) errors by bundle adjustment [46, 35]. We use a redundant number of corresponding points to minimize the squared re-projection error for the detected 2D points $\boldsymbol{x}_{ij}$ in an image j as a function of the unknown image pose parameters $(\boldsymbol{R}, \boldsymbol{X})$ and unknown 3D point positions $\boldsymbol{p}_i$ using nonlinear least squares. For the projection equations:

$$\boldsymbol{x}_{ij} = \boldsymbol{f}(\boldsymbol{p}_i, \boldsymbol{R}_j, \boldsymbol{X}_j), \tag{4}$$

the iteratively minimized re-projection errors are

$$\boldsymbol{E} = \sum_{i,j} \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \Delta \boldsymbol{x} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{R}} \Delta \boldsymbol{R} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{X}} \Delta \boldsymbol{X} - \boldsymbol{r}_{ij} \right), \tag{5}$$

where $\boldsymbol{r}_{ij} = \boldsymbol{x}_{ij} - \hat{\boldsymbol{x}}_{ij}$ is the current residual vector (2D error in the predicted position) and the partial derivatives are with respect to the unknown pose parameters (camera rotation and translation and 3D point coordinates).

To determine the real scale of the object we use the known distance between reference points presented in the working area.

It is worth mentioning that the quality of parameters estimation by bundle adjustment procedure depends on the distribution of used points in images of processing image set. To check the results of camera pose estimation, we use a set of control points with known 3D coordinates located in the working area. The relative differences in pose parameters determined by bundle adjustment and by control 3D points lay in 2% limits showing a reasonable accuracy of pose estimation.

4.3. Semi-Global Matching

The well-known semi-global matching approach [19, 20] uses rectified stereo pairs for pixel-wise matching based on a cost function that consists of a term for dissimilarity and two penalty functions for disparity distances. The aggregated cost space is searched in different paths providing the best matching disparity. SGM is usually combined with multi-view stereo (MVS) in order to model a complete 3D object from multiple views. The results in figure 8 have been computed with semi-global block matching provided by OpenCV which is a modified implementation of [20]. The point clouds computed from multiple views are finely registered using iterative closest point algorithm to provide a final 3D model.

As outlined in [4], SGM can be converted to object space (OSGM). In this case, a voxel space is created where the standard SGM cost function is replaced by a new cost function based on XYZ values directly. OSGM is able to process all given images simultaneously without the need of rectified normalized stereo pairs. The desired voxel resolution can be set arbitrarily, *e.g.* with respect to the given ground sampling distance of the camera. Besides the reconstructed 3D model, OSGM directly derives true orthophotos.

5. Evaluation

In this section, we present an evaluation of the developed patch matching method. Firstly, we evaluate the CAE ability to reconstruct arbitrary infrared image patches and ex-

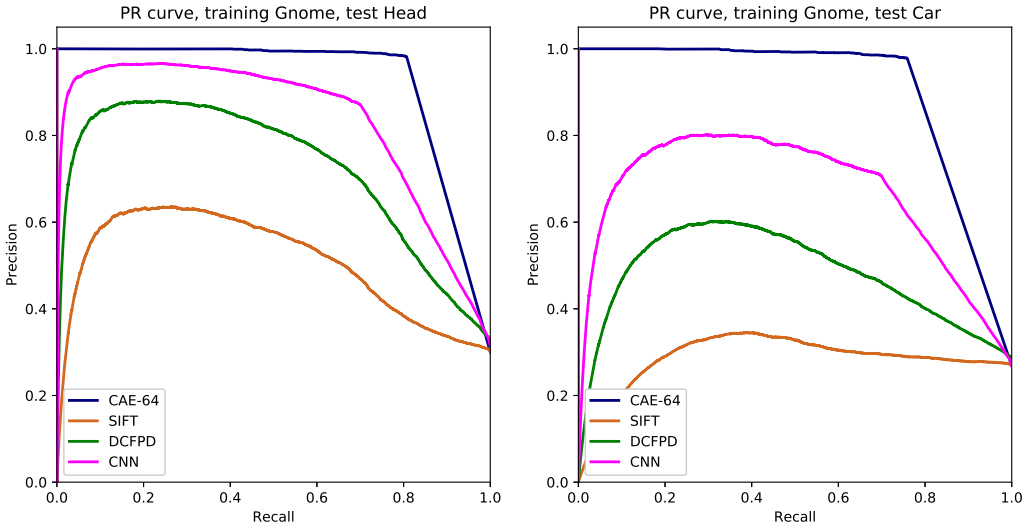


Figure 7. PR curves for the CAE, SIFT, DCFPD, and CNN on two splits of the MVSIR dataset.

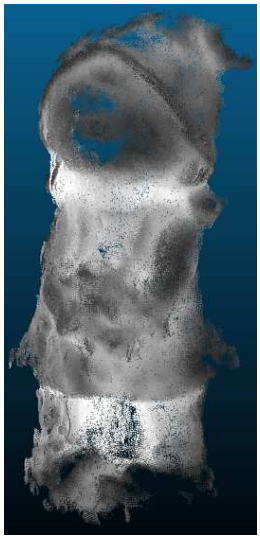


Figure 8. A 3D reconstruction (MVS and OpenCV semi-global block matching).

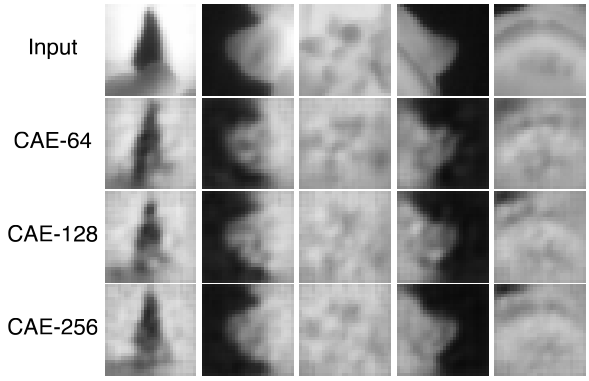


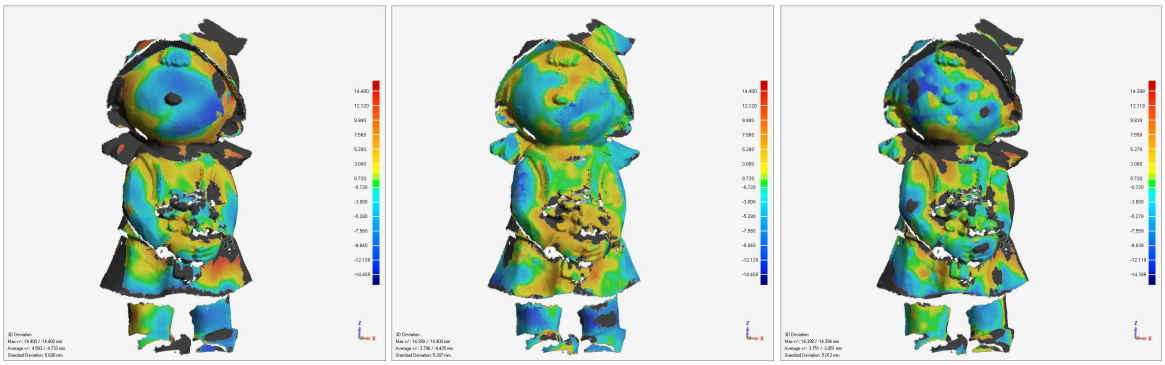
Figure 9. Result of CAE reconstruction with various dimensions of code F.

tract discriminative features for image matching. Secondly, we compare the matching accuracy to other feature descriptors. Finally, we evaluate our matching method in the full 3D object reconstruction pipeline and compare it to 3D reconstructions generated by well-established algorithms. For 3D model evaluation, we adopt methodology proposed in [40] to measure the accuracy of the generated surface.

5.1. CAE Evaluation

The developed architecture was implemented using Caffe [23] and trained using an NVIDIA Titan GPU. We used a fixed learning rate of $6 \cdot 10^{-3}$. The training was completed in 20000 iterations with a batch size of 100. The average error of the Euclidean loss function after training was about 1.5. Firstly, we evaluated the CAE reconstruction quality visually by processing infrared image patches that were not included in MVSIR dataset and presented on figure 9. The top row of the figure shows input images that were fed into a CAE. Bottom rows present the reconstruction produced by the CAE with different dimensions of the code F. All dimensions were sufficient to obtain an accurate reconstruction of the input image. Only the CAE with $F = 64$ shows few features of the input images.

To evaluate discriminative qualities of the code generated by a CAE we use the test part of the MVSIR dataset. We compare the discriminative quality of codes produced by the CAE with SIFT, deep convolutional feature point descriptors (DCFPD) [43] and stereo matching CNN [59]. We



(a) σ_{PS} = 5.69 mm (b) σ_{3DZ} = 5.21 mm (c) σ_{SGM} = 5.01

Figure 10. Evaluation result for the Gnome statue. Distances from the ground truth model are presented in false color.

Train	Test	SIFT	DCFPD	CNN	CAE-64
Gnome	Head	0.51	0.72	0.83	**0.93**
Gnome	Car	0.28	0.47	0.66	**0.89**

Table 2. PR AUC for the three MVSIR dataset splits.

Method	Gnome	Car	Head
PMVS	6.91	-	-
PS	5.69	6.12	6.71
3D Zephyr	5.21	6.11	5.34
Ours	**5.01**	**2.81**	**4.41**

Table 3. Standard deviation of distances in mm to the ground-truth 3D model of evaluated methods on the MVSIR dataset. '-' indicates that the method has failed during point matching stage.

use precision-recall curve (PR) and area under the curve (AUC) as performance metrics. The possible number of patch combinations in the test part of MVSIR dataset is excessive. We reduce the number of patch pairs to 10000 positive random pairs and 30000 random negative pairs. We compare pairs using the euclidean distance between codes F_1, F_2 for two patches. We use SIFT implementation from the VLFeat project [50]. The PR curve for a sample split of the MVSIR dataset is presented in figure 7. The detailed results for PR AUC are given in table 2.

5.2. Evaluation of 3D reconstruction

We compare the accuracy of reconstructed 3D models with three other algorithms implemented in open source and commercial software: Agisoft PhotoScan (PS), 3DF Zephyr and PMVS (Visual SfM). As a ground truth data, we use 3D models generated by a 3D scanner based on fringe projection. The 3D scanner [28] provides 0.1 mm accuracy for reconstructed reference 3D models. To evaluate the deviation of 3D models obtained by various techniques from the reference 3D model we transform them to a common coordinate system and display deviations using pseudo colors. The accuracy of the reconstructed surfaces is presented in figure 10 and table 3.

6. Conclusion

We showed that convolutional auto-encoders are capable of extracting features from low or non-textured objects to perform robust patch matching from multi view stereo infrared imagery. The CAE prove to generalize from training dataset to previously unseen data and are robust to image matching challenges specific to the infrared range such as high noise level and local changes of temperature contrast.

To compare the CAE-based image matching technique with the well-known state-of-the-art image matching algorithms, we designed a new MVSIR dataset with infrared images, ground truth point correspondences and reference 3D models. We showed that application of the CAE for feature matching on thermal imagery provides the better performance compared to other feature descriptors.

We evaluated a set of 3D reconstruction algorithms (SfM, Visual SfM, SGM, 3DF Zephyr) on MVSIR dataset to find out which one works better for thermal imagery. SGM (using the developed technique for patch matching) have demonstrated the best accuracy of 3D reconstruction.

We demonstrated that proposed 3D reconstruction pipeline allows obtaining 3D models based on thermal imagery with reasonable accuracy for tasks of infrared 3D modeling and pose estimation and for a creation of datasets for deep learning in the infrared spectra.

7. Acknowledgements

The reported study was funded by Russian Scientific Foundation (RSF) according to the project №16-11-00082 and Russian Foundation for Basic Research (RFBR) according to the projects №17-29-04509 and №16-38-00940.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer vision–ECCV 2006*, 2006.

[2] A. Berg. *Detection and Tracking in Thermal Infrared Imagery*. PhD thesis, Linköping University Electronic Press, Mar. 2016.

[3] A. Berg, J. Ahlberg, and M. Felsberg. A thermal infrared dataset for evaluation of short-term tracking methods. In *Swedish Symposium on Image Analysis*, Svenska sllskapet fr automatiserad bildanalys (SSBA), 2015.

[4] F. Bethmann and T. Luhmann. Semi-Global Matching in Object Space. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3/W2:23–30, 2015.

[5] H. Beyer. Advances in characterization and calibration of digital imaging systems. *nt. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XXIX*, pages 545–555, 1992.

[6] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In *British Machine Vision Conference 2011*, pages 14.1–14.11. British Machine Vision Association.

[7] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. *ECCV*, 8690(Chapter 54):834–849, 2014.

[8] C. H. Esteban and F. Schmitt. Silhouette and Stereo Fusion for 3D Object Modeling. *3DIM*, pages 46–53, 2003.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2009.

[10] T. Fäulhammer, R. Ambrus, C. Burbridge, M. Zillich, J. Folkesson, N. Hawes, P. Jensfelt, and M. Vincze. Autonomous learning of object models on a mobile robot. *Robotics and Automation Letters (IEEE Journal)*, Volume 2(Issue 1):26–33, 2016.

[11] S. GALLIANI, K. LASINGER, and K. SCHINDLER. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25:361–369, 2016.

[12] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions. *CVPR*, pages 1–8, 2007.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics - The KITTI dataset. *I. J. Robotics Res.*, 32(11):1231–1237, 2013.

[14] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Nov. 2016.

[15] K. Hajebi and J. S. Zelek. Structure from Infrared Stereo Images. In *2008 Canadian Conference on Computer and Robot Vision*, pages 105–112. IEEE, 2008.

[16] K. Hajebi and J. S. Zelek. Structure from Infrared Stereo Images. In *2008 Canadian Conference on Computer and Robot Vision*, pages 105–112. IEEE, 2008.

[17] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Alvey Vision Conference 1988*, pages 23.1–23.6. Alvey Vision Club.

[18] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. R. Bradski, K. Konolige, and N. Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. *ACCV*, 7724(Chapter 42):548–562, 2012.

[19] H. Hirschmuller. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 807–814. IEEE.

[20] H. Hirschmüller. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell. ()*, 30(2):328–341, 2008.

[21] D. A. Huckridge, R. Ebert, and S. T. Lee, editors. *Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs*. SPIE, Oct. 2016.

[22] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection - Benchmark dataset and baseline. *CVPR*, pages 1037–1045, 2015.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[24] V. John, S. Tsuchizawa, Z. Liu, and S. Mita. Fusion of thermal and visible cameras for the application of pedestrian detection. *Signal, Image and Video Processing*, 11(3):517–524, Oct. 2016.

[25] E. Johns, S. Leutenegger, and A. J. Davison. Pairwise decomposition of image sequences for active multi-view recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3813–3822, 2016.

[26] W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab. Deep Learning of Local RGB-D Patches for 3D Object Detection and 6D Pose Estimation. *ECCV*, 9907(7):205–220, 2016.

[27] V. V. Kniaz. A photogrammetric technique for generation of an accurate multispectral optical flow dataset. In F. Remondino and M. R. Shortis, editors, *SPIE Optical Metrology*, pages 103320G–12. SPIE, June 2017.

[28] V. A. Knyaz. Multi-media projector single camera photogrammetric system for fast 3d reconstruction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5:343–348, 2010.

[29] V. A. Knyaz and A. G. Chibunichev. Photogrammetric techniques for road surface analysis. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B5:515–520, 2016.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012.

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[32] Y. Lecun, L. Bottou, G. B. Orr, and K. R. Müller. Efficient BackProp. In *Neural Networks: Tricks of the Trade*, pages 9–50. Springer, Berlin, Heidelberg, Berlin, Heidelberg, 1998.

[33] A. Lewis, G. E. Hilley, and J. L. Lewicki. Integrated thermal infrared imaging and structure-from-motion photogrammetry to map apparent temperature and radiant hydrothermal heat flux at Mammoth Mountain, CA, USA. *Journal of Volcanology and Geothermal Research*, 303:16–24, Sept. 2015.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, Cham, Sept. 2014.

[35] M. I. A. Lourakis and A. A. Argyros. Sba: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software (TOMS)*, 36(1), 2009.

[36] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[37] T. Luhmann, J. Piechel, and T. Roelfs. Geometric calibration of thermographic cameras. In C. Kuenzer and S. Dech, editors, *Thermal Infrared Remote Sensing Sensors, Methods, Applications*, pages 27–42. Berlin, 2013.

[38] D. Michel, X. Zabulis, and A. A. Argyros. Shape from interaction. *Machine Vision Applications*, 25(4):1077–1087, May 2014.

[39] N. K. Negied, E. E. Hemayed, and M. B. Fayek. Pedestrians' detection in thermal bands – Critical survey. *Journal of Electrical Systems and Information Technology*, 2(2):141–148, Sept. 2015.

[40] F. Remondino, M. G. Spera, E. Nocerino, F. Menna, and F. Nex. State of the art in high density image matching. *The Photogrammetric Record*, 29(146):144–166, June 2014.

[41] K. Ridgeway, J. Snell, B. Roads, R. S. Zemel, and M. C. Mozer. Learning to generate images with perceptual similarity metrics. *CoRR*, abs/1511.06409, 2015.

[42] J.-F. Shi, S. Ulrich, and S. Ruel. Spacecraft Pose Estimation using Principal Component Analysis and a Monocular Camera. In *AIAA Guidance, Navigation, and Control Conference*, pages 131–24, Reston, Virginia, Jan. 2017. American Institute of Aeronautics and Astronautics.

[43] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. *ICCV*, pages 118–126, 2015.

[44] X. Sun, T. Xu, J. Zhang, and X. Li. A Hierarchical Framework Combining Motion and Feature Information for Infrared-Visible Video Registration. *Sensors*, 17(2):384–16, Feb. 2017.

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 2015.

[46] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[47] V. Turchenko and A. Luczak. Creation of a Deep Convolutional Auto-Encoder in Caffe. *CoRR abs/1501.02565*, 1512:arXiv:1512.01596, 2015.

[48] K. Tzevanidis, X. Zabulis, T. Sarmis, P. Koutlemanis, N. Kyriazis, and A. A. Argyros. From multiple views to textured 3d meshes: A gpu-powered approach. In *European Conference on Computer Vision Workshops (CVGPU 2010 - ECCVW 2010)*, pages 384–397, Heraklion, Crete, Greece, September 2010. Springer.

[49] H. M. Vazquez, C. S. Martín, J. Kittler, Y. Plasencia, and E. B. G. Reyes. Face Recognition with LWIR Imagery Using Local Binary Patterns. *ICB*, 5558(Chapter 34):327–336, 2009.

[50] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

[51] G. Vogiatzis, C. Hernandez, P. H. S. Torr, and R. Cipolla. Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2241–2246, 2007.

[52] M. Weinmann, J. Leitloff, L. Hoegner, B. Jutzi, U. Stilla, and S. Hinz. Thermal 3D mapping for object detection in dynamic scenes. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-1:53–60, 2014.

[53] Wenbin, Li, Cosker, Darren, Lv, Zhihan, and Brown, Matthew. Nonrigid Optical Flow Ground Truth for Real-World Scenes With Time-Varying Shading Effects. *IEEE Robotics and Automation Letters*, pages 231–238, 2017.

[54] P. Wohlhart and V. Lepetit. Learning Descriptors for Object Recognition and 3D Pose Estimation. *arXiv.org*, page arXiv:1502.05908, Feb. 2015.

[55] Y. Xiang, W. Kim, W. Chen, J. Ji, C. B. Choy, H. Su, R. Mottaghi, L. J. Guibas, and S. Savarese. ObjectNet3D - A Large Scale Database for 3D Object Recognition. *ECCV*, 2016.

[56] Z. Xie, P. Jiang, and S. Zhang. Fusion of LBP and HOG using multiple kernel learning for infrared face recognition. *ICIS*, 2017.

[57] M. Yamaguchi, H. Saito, and S. Yachida. Application of LSD-SLAM for Visualization Temperature in Wide-area Environment. *VISIGRAPP*, pages 216–223, 2017.

[58] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016.

[59] J. Zbontar and Y. Lecun. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 2016.

[60] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan. VAIS - A dataset for recognizing maritime imagery in the visible and infrared spectrums. *CVPR Workshops*, pages 10–16, 2015.

[61] R. Zhang, P. Isola, and A. A. Efros. Colorful Image Colorization. *ECCV*, 9907(Chapter 40):649–666, 2016.