EVALUATING THE ACCURACY OF 3D OBJECT RECONSTRUCTION FROM THERMAL IMAGES

V. A. Mizginov[a,*], V. V. Kniaz[a,b]

[a] State Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia
(vl.mizginov,vl.kniaz)@gosniias.ru
[b] Moscow Institute of Physics and Technology (MIPT), Russia

Technical Commission II

KEY WORDS: optical metrology, 3D scanner, thermal images, multispectral images

ABSTRACT:

Thermal cameras are increasingly used in many photogrammetric and computer vision tasks. Nowadays it is possible to detect and recognize objects in infrared images, to solve such tasks as pedestrian detection (Huckridge et al., 2016), security applications, and autonomous driving (Wenbin, Li et al., 2017). Nevertheless, some tasks that are easily solved in the visible range data are still challenging to achieve in the infrared range. Reconstruction of a 3D object model from infrared images is challenging due to the low contrast of the original infrared image, noise of the sensor, and the absence of feature points on the image. Nevertheless, thermal cameras have their advantages, which make them popular for solving practical problems. Firstly, thermal cameras can be used in degraded environments (smoke, fog, precipitation, low light conditions). Secondly, infrared images can be fused with color images (Gao et al., 2013) to increases the system's performance.
This paper is focused on the evaluation of accuracy of 3D object reconstruction from thermal images. The evaluation of the accuracy is threefold. Firstly, we train four stereo matching methods (CAE, LF-Net, SURF, and SIFT) on the MVSIR dataset (Knyaz et al., 2017) and our new ThermalPatches dataset. We used two RTX 2080 Ti GPUs and the PyTorch library for the training. Secondly, we evaluate the matching score for the selected methods. Finally, we perform 3D object reconstruction using the SfM (Remondino et al., 2014) approach and matches for each method. We compare the object space accuracy of the resulting surfaces to the ground-truth 3D models generated with a structured light 3D scanner.

1. INTRODUCTION

Thermal cameras are increasingly used in many photogrammetric and computer vision tasks. Nowadays it is possible to detect and recognize objects in infrared images, to solve such tasks as pedestrian detection (Huckridge et al., 2016), security applications, and autonomous driving (Wenbin, Li et al., 2017). Nevertheless, some tasks that are easily solved in the visible range data are still challenging to achieve in the infrared range. Reconstruction of a 3D object model from infrared images is challenging due to the low contrast of the original infrared image, noise of the sensor, and the absence of feature points on the image. Nevertheless, thermal cameras have their advantages, which make them popular for solving practical problems. Firstly, thermal cameras can be used in degraded environments (smoke, fog, precipitation, low light conditions). Secondly, infrared images can be fused with color images (Gao et al., 2013) to increases the system's performance (Kniaz et al., 2019a, Kniaz and Knyaz, 2019). Finally, multispectral image datasets are highly demanded nowadays. While many large datasets of images captured in the visible range can be found in the public domain (Lin et al., 2014, Everingham et al., 2015), nowadays only a few small thermal images datasets are available.

Moreover, such datasets are limited in terms of object classes and imaging conditions. 3D object reconstruction techniques such as Structure from Motion (SfM) (Remondino et al., 2014), simultaneous localization and mapping (SLAM) (Engel et al., 2014), Semi-global Matching (SGM) (Hirschmuller, n.d., Bethmann and Luhmann, 2015), silhouette-based 3D reconstruction (Tzevanidis et al., 2010), Shape from Interaction (Michel et al., 2014), and deep learning-based methods (Knyaz et al., 2019, Kniaz et al., 2019b) prove to be fast and robust techniques for 3D model generation from the imagery captured in the visible range.But these methods are not accurate for reconstructing infrared models (Figure 1). Thus a feature extraction method that is robust to low contrast details is required for 3D object reconstruction in the thermal range. 3D modeling is a perspective approach for the generation of large datasets of thermal images. Nevertheless, real thermal textures are required to generate realistic 3D models. Therefore, thermal 3D scanning systems are needed to create 3D models with thermal textures. Moreover, 3D reconstruction of objects from thermal images is required in such applications as thermal texturing of 3D models and reconstruction of hot air streams.

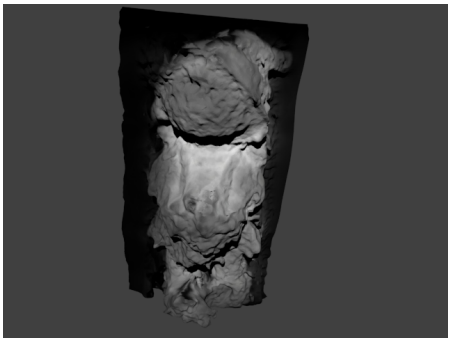

Figure 1. Example of 3D reconstruction in the thermal range.

*Corresponding author

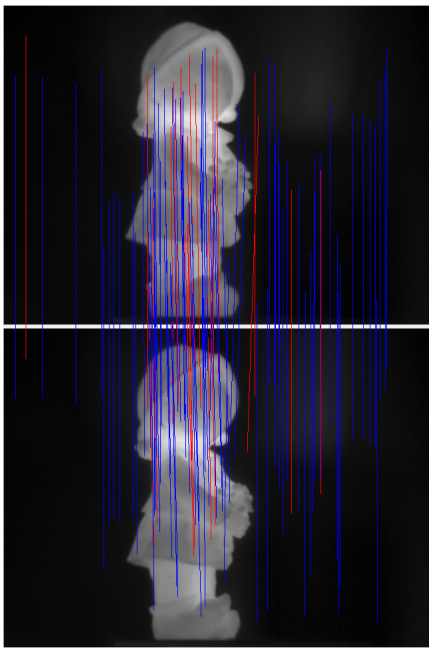

Figure 2. Example of point matching in thermal images using SIFT algorithm.

2. RELATED WORK

The reconstruction of three-dimensional models of objects from images has been developing successfully for a long time. Recently the possibility of of solving this task using a monocular camera has been actively investigated. Robust image matching is the main element in modern reconstruction techniques. Such methods of 3D object reconstruction are based on feature descriptors (Bay et al., 2006).

The development of infrared cameras has increased interest in the reconstruction of three-dimensional models in the infrared range (Hajebi and Zelek, 2008, Weinmann et al., 2014, Negied et al., 2015, Lewis et al., 2015). However, reconstruction of objects in the infrared range is a very difficult task due to the presence of different distortions such as infrared reflections, infrared halo effects, saturation. Also, there are low contrast in the infrared images and such methods as SIFT (Lowe, 1999) and SURF (Bay et al., 2006) fail to obtain feature points (Figure 2). The evaluation of methods that do not use feature points such as the LSD-SLAM (Engel et al., 2014, Yamaguchi et al., 2017) 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.

Image matching methods that use finite object planes such as plane sweep matching or PatchMatch (GALLIANI et al., 2016, Bleyer et al., n.d., Gallup et al., 2007) 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.

Active development of convolutional neural networks has led to the appearance of deep learning based feature descriptors (Wohlhart and Lepetit, 2015). They outperform their hand-crafted predecessors in matching accuracy.Recently many deep learning methods were proposed for robust stereo matching such as the convolutional autoencoder (CAE) for stereo matching (Knyaz et al., 2017), LF-Net (Ono et al., 2018), and LIFT (Yi et al., 2016). The first group is based on classical deep convolutional neural networks (CNN) for image classification (Krizhevsky et al., 2012, Szegedy et al., 2015). 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. One such method is LIFT(Learned Invariant Feature Transform). This is a novel deep network architecture that implements the full feature point handling pipeline, that is, detection, orientation estimation, and feature description in a unified manner. LIFT is based on the four-branch Siamese architecture. Each branch contains three distinct CNNs, a detector, an orientation estimator, and a descriptor. Experimental results demonstrate that such approach outperforms the state-of-the-art methods on visible range,but there are no evaluations for infrared images. Another algorithm from this group is LF-Net(a Local Feature Network). LF-Net has two main components. The first one is a dense, multi-scale, fully convolutional network that returns keypoint locations, scales, and orientations. It is designed to achieve fast inference time, and to be agnostic to image size. The second is a network that outputs local descriptors given patches cropped around the keypoints produced by the first network.

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 (Kehl et al., 2016) 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. The performance of machine learning methods can be tuned by training on the dataset containing local patches of target objects. However, there are few training datasets (Knyaz et al., 2017) that include local patches of thermal images. 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.

3. METHOD

Our evaluation of 3D Object Reconstruction accuracy is twofold. Firstly, we want to compare the accuracy of keypoint matching for four algorithms: SIFT (Lowe, 1999), SURF (Bay et al., 2006), LF-Net (Ono et al., 2018), and CAE-64 (Knyaz et al., 2017). Secondly, we would like to perform 3D Object reconstruction using the single SfM approach and various keypoint matches provided by the four algorithms. We perform evaluation on two datasets: MVSIR (Knyaz et al., 2017) and our new ThermalPatches dataset.

3.1 ThermalPatches Dataset Generation

MVSIR dataset was generating using two FLIR P640 cameras with a resolution of 640×480 pixels and focal length of 130 mm. Our new ThermalPatches dataset contains images from a low cost thermal 3D reconstruction system that includes two FLIR ONE Pro cameras with an internal resolution of 160×120 pixels upsampled to 640×480 using super-resolution chip and focal length of 40 mm. The dataset consists of thermal stereo pairs of three solid objects: Gnome statue, Head, and Car (Figure 3).The dataset includes 540 thermal images.Multiview pictures were taken in increments of 2 degrees. The FLIR ONE camera produces as a standard output thermal preview images that present temperature of

captured objects as monochrome (or pseudocolors) images with a reference temperature scale bar. Also the FLIR ONE camera provides raw 16bit data and the EXIF information for acquired images. Values of the raw data represent the object emission in the wavelengths 8–14 µ.

We generated ground-truth 3D models of the test objects using a structured light scanner. Also, we generated scans of a hot water stream, to evaluate the scanners' performance on the task of 3D reconstruction of liquid objects. We generated ground-truth correspondences for the thermal stereo pairs using the ground-truth 3D models and raytracing. We provide the ground-truth correspondences both as the cropped image patches and a dense optical flow from the left to the right image in the stereo pair. Our ThermalPatches dataset includes 1200 cropped image patches (Figure 5), three ground-truth 3D models, and camera poses in the object coordinate system for each stereo pair.

4. EXPERIMENTS

We present the results of the evaluation and demonstrate the quality of 3D object reconstruction from different image datasets.We reconstructed the ground truth 3D models of each object in visible range. After that we compare the matching accuracy four feature descriptors.In the end, we demonstrate the evaluate of full 3D object reconstruction for MVSIR and ThermalPatches. As a ground truth data, we use 3D models generated by a 3D scanner based on fringe projection. The 3D scanner (Knyaz, 2010) provides 0.1 mm accuracy for reconstructed reference 3D models.

The first part of the evaluation of three-dimensional reconstruction of objects is the matching feather points on thermal images. Using thermal images of an object of the same class from different datasets, we present that the traditional feature descriptors methods are not robust and effective especially for low-cost cameras (Figure 6).

We use precision-recall curve (PR) and area under the curve (AUC) as performance metrics. The detailed results for PR AUC are given in Table 1

Dataset	**SIFT**	**SURF**	**LF-NET**	**CAE-64**
ThermalPatches	0.2	0.4	0.7	**0.85**
MSVIR	0.35	0.62	0.81	**0.9**

Table 1. PR AUC for MVSIR and ThermalPatches dataset.

The second step in evaluation of three-dimensional reconstruction objects has begun the demonstrate the evaluate of full 3D object reconstruction for MVSIR and ThermalPatches. We compare the accuracy of reconstructed 3D models using open source and commercial software: Agisoft PhotoScan (PS) and GeoMagic Design X. 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. Our pipeline failed to reconstruct three-dimensional model of the Gnome model from ThermalPatches datasets. The accuracy of the reconstructed surfaces and examples of comparing some objects are presented in Figures 7–9 and Table 2.

Method	**Head**	**Gnome**	**Car**
ThermalPatches	10.1	-	9.8
MSVIR	6.07	5.25	5.8

Table 2. Standard deviation of distances in mm for MVSIR and ThermalPatches dataset.

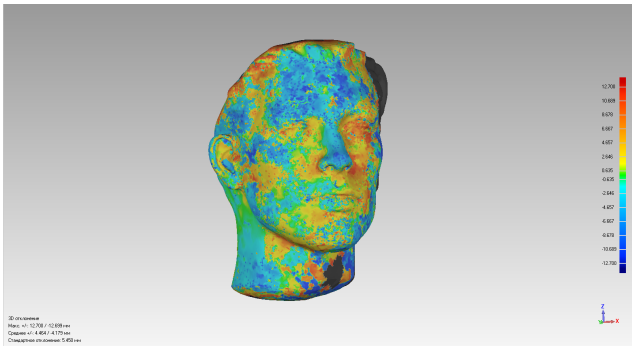

Figure 7. Evaluation result for the Head statue for images from MSVIR dataset.

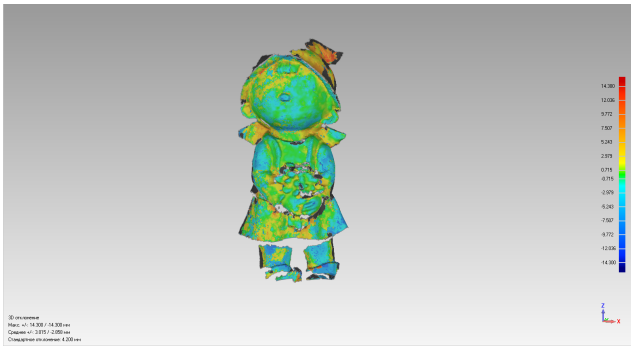

Figure 8. Evaluation result for the Gnome statue for images from MSVIR dataset.

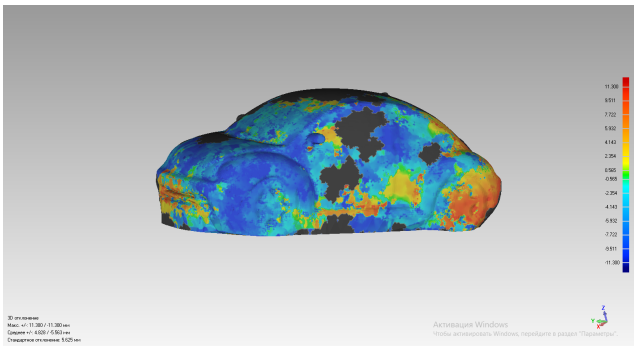

Figure 9. Evaluation result for the Car for images from ThermalPatches dataset.

5. CONCLUSION

We evaluated the accuracy of four feature point matching algorithms on the MVSIR and our new ThermalPatches datasets. We

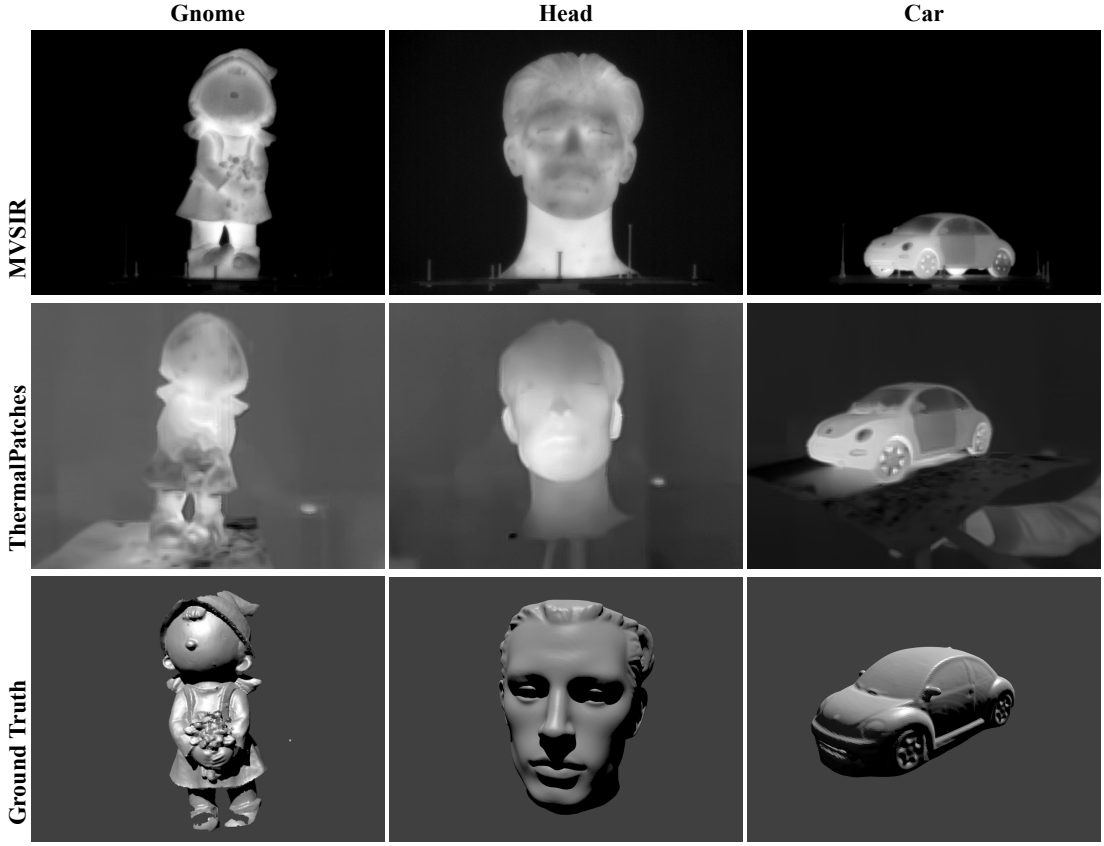

Figure 3. Comparing images from MVSIR and ThermalPatches datasets.

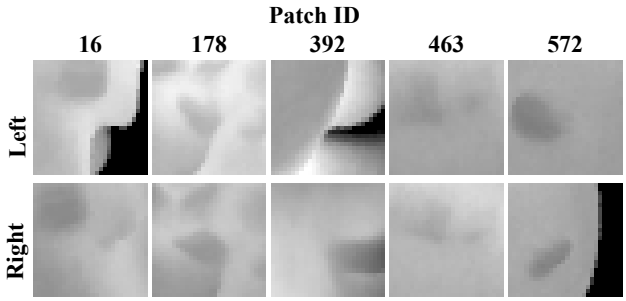

Figure 4. Examples of cropped images patches from MVSIR dataset.

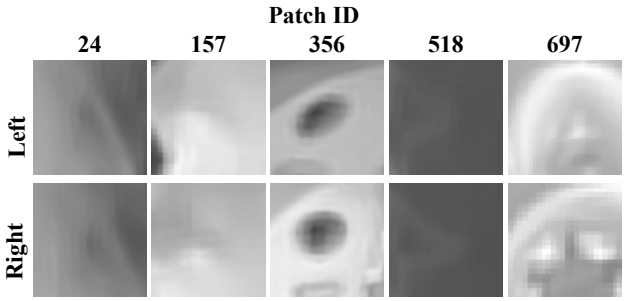

Figure 5. Examples of cropped images patches from ThermalPatches dataset.

generated a new ThermalPatches dataset that includes low resolution infrared images of three objects.The size of the dataset is 540 images. Evaluation of matching feature points demonstrates that traditional methods cannot obtain feature points on low resolution thermal images. 3D models reconstructed from low-cost thermal scanner show limited details and have the object space accuracy of about 50 mm for the working volume of $100 \times 100 \times 100$ mm. While such accuracy is insufficient for texture generation, it is still possible to track the flow of the hot liquid using a low-cost scanner. 3D reconstruction using the moderate resolution images allows acheiving the object space accuracy of about 7 mm for the working volume of $100 \times 100 \times 100$ mm.

ACKNOWLEDGEMENTS

The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project N^{o} 17-29-04410.

REFERENCES

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

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

Bleyer, M., Rhemann, C. and Rother, C., n.d. PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In: *British Machine Vision Conference 2011*, British Machine Vision Association, pp. 14.1–14.11.

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

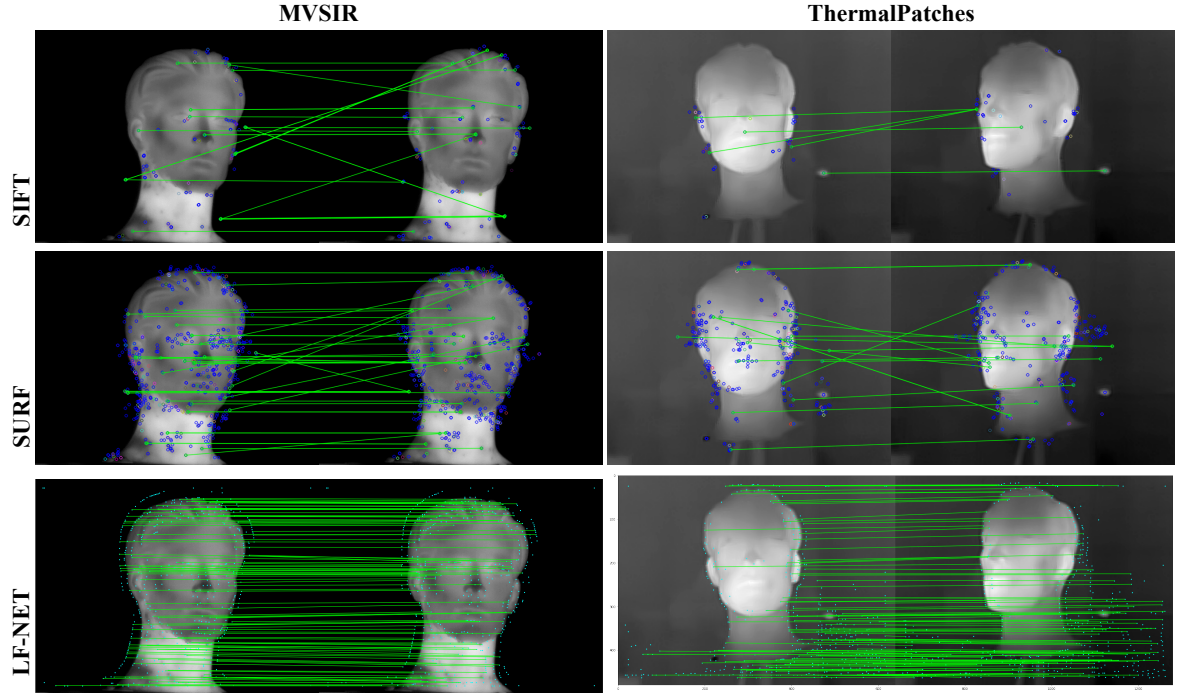

Figure 6. Evaluation of three feature point matching algorithms on two datasets.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision* 111(1), pp. 98–136.

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

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

Gao, S., Cheng, Y. and Zhao, Y., 2013. Method of visual and infrared fusion for moving object detection. *Optics letters* 38, pp. 1981–3.

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

Hirschmuller, H., n.d. 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)*, IEEE, pp. 807–814.

Huckridge, D. A., Ebert, R. and Lee, S. T. (eds), 2016. Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs. SPIE.

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

Kniaz, V. V. and Knyaz, V. A., 2019. Chapter 6 - multispectral person re-identification using gan for color-to-thermal image translation. In: M. Y. Yang, B. Rosenhahn and V. Murino (eds), *Multimodal Scene Understanding*, Academic Press, pp. 135 – 158.

Kniaz, V. V., Knyaz, V. A., Hladůvka, J., Kropatsch, W. G. and Mizginov, V., 2019a. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: L. Leal-Taixé and S. Roth (eds), *Computer Vision – ECCV 2018 Workshops*, Springer International Publishing, Cham, pp. 606–624.

Kniaz, V. V., Remondino, F. and Knyaz, V. A., 2019b. Generative adversarial networks for single photo 3d reconstruction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W9, pp. 403–408.

Knyaz, V. A., 2010. 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, pp. 343–348.

Knyaz, V. A., Kniaz, V. V. and Remondino, F., 2019. Image-to-voxel model translation with conditional adversarial networks. In: L. Leal-Taixé and S. Roth (eds), *Computer Vision – ECCV 2018 Workshops*, Springer International Publishing, Cham, pp. 601–618.

Knyaz, V. A., Vygolov, O., Kniaz, V. V., Vizilter, Y., Gorbatsevich, V., Luhmann, T. and Conen, N., 2017. Deep learning of convolutional auto-encoder for image matching and 3d object reconstruction in the infrared range. In: *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pp. 2155–2164.

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

Lewis, A., Hilley, G. E. and Lewicki, J. L., 2015. 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, pp. 16–24.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P., 2014. Microsoft COCO: Common Objects in Context. *ArXiv e-prints*.

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

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

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

Ono, Y., Trulls, E., Fua, P. and Yi, K. M., 2018. Lf-net: Learning local features from images. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 6237–6247.

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

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

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

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

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

Wohlhart, P. and Lepetit, V., 2015. Learning Descriptors for Object Recognition and 3D Pose Estimation. *arXiv.org* p. arXiv:1502.05908.

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

Yi, K. M., Trulls, E., Lepetit, V. and Fua, P., 2016. LIFT: learned invariant feature transform. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pp. 467–483.