Multimodal Dense Stereo Matching

Max Mehltretter[1], Sebastian P. Kleinschmidt[2],
Bernardo Wagner[2], and Christian Heipke[1]

[1] Institute of Photogrammetry and GeoInformation
Leibniz Universität Hannover, Germany
{mehltretter,heipke}@ipi.uni-hannover.de
[2] Institute of Systems Engineering - Real Time Systems Group
Leibniz Universität Hannover, Germany
{kleinschmidt,wagner}@rts.uni-hannover.de

Abstract. In this paper, we propose a new approach for dense depth estimation based on multimodal stereo images. Our approach employs a combined cost function utilizing robust metrics and a transformation to an illumination independent representation. Additionally, we present a confidence based weighting scheme which allows a pixel-wise weight adjustment within the cost function. We demonstrate the capabilities of our approach using RGB- and thermal images. The resulting depth maps are evaluated by comparing them to depth measurements of a Velodyne HDL-64E LiDAR sensor. We show that our method outperforms current state of the art dense matching methods regarding depth estimation based on multimodal input images.

1 Introduction

The reconstruction of depth information from a set of images is a well-known problem in the fields of photogrammetry and computer vision. For this task, the identification of image correspondences is an essential prerequisite. However, in the presence of difficult environmental conditions such as lighting or weather changes, the performance of state of the art correspondence identification techniques is limited. This task becomes even more challenging if cameras of different imaging modalities are used, which are operating in different parts of the electromagnetic spectrum. In this case, correspondences of image features of one modality may be represented differently or may be absent altogether in images of other modalities. However, multimodal imaging may help to make current computer vision algorithms more robust due to additional spectral information: Thermal imaging is insensitive to lighting variations and therefore still works when RGB-cameras fail. Consequently, establishing spatial relations between images of different modalities is relevant for a variety of applications. In medical diagnosis for example, multimodal image fusion algorithms have shown notable achievements in improving the clinical accuracy of decisions based on medical images [13]. Whereas most multimodal approaches for medical applications are only confronted with objects of a limited and known anatomic atlas,

more general computer vision applications have to deal with a larger variety of objects and environments. Therefore, they have a different focus and varying requirements. The presented work uses the advantages of multimodal imaging for dense depth estimation using a multimodal stereo setup.

The overall aim of the current work is to investigate the feasibility of dense depth estimation from multimodal stereo images. For this purpose, we examine the performance of a dense image matching approach using one thermal- and one RGB image as input data. We demonstrate the possibility to reconstruct an environment densely even if its representation in the images differs greatly. Our main contributions on this topic are:

- A combined cost function utilizing robust metrics and a transformation to an illumination independent representation.
- A confidence based weighting scheme which allows adjusting the weights within the cost function pixel-wise.

2 Related Work

2.1 Multimodal Image Fusion

The combination of RGB and thermal imaging is useful for a series of applications as pedestrian detection and tracking as well as silhouette extraction (e.g. [2, 4, 7, 20, 33]), agricultural applications (e.g. [21, 26]), maintenance (e.g. [28]) and traffic monitoring (e.g. [1]). Multimodal image registration has been realized using contours [8], Harris Corners [10], Hough lines [12], wavelet transformations [27] or the intrinsic and extrinsic camera calibration [15]. Beside merging image features of different modalities only using two-dimensional information, depth information of a depth sensor [16] or structure from motion techniques [17] can be used for feature association. Based on spatially aligned multimodal images, the authors of [15] perform an analysis of the statistical and spatial distribution of sparse image features for RGB, IR and thermal imaging and conclude, that only a relatively small quantity of sparse image features has corresponding image features across modalities for the evaluated scenario. The results indicate a potential limit for the number of transferable image features across imaging modalities.

For multimodal dense matching, the performance of current state of the art techniques is limited as well: Because area-based cross-correlation only works insufficiently for a thermal stereo setup, [6] obtains phase congruency maps for two thermal images before correlation is performed. The authors conclude, that their method is more robust than matching the greyscale images directly. The work presented in [25] investigates the feasibility of matching multimodal features in a stereo setup consisting of an RGB and a thermal camera. The authors introduce a novel feature descriptor based on the combination of the phase congruency and the spatial distribution of the contours in a window around the extracted point.

2.2 Dense Image Matching

Finding correspondences between images of different modalities (e.g. RGB and thermal) and between those taken under different lighting conditions leads to a similar problem: In general, the grey- or RGB-values of such an image pair cannot be transformed to each other in a global linear manner. Rather, the transformations depend on the different depicted objects and therefore vary locally. However, most of the published approaches focus on image pairs which were taken under similar conditions and with the same type of sensor (e.g. [9, 31]). Hence, they assume the scene to appear similarly in all images. Consequently, these methods are not robust against influences which have an impact on the imaging process, as changing illumination or contrast. Only a few methods like [14, 23, 24] address this problem. Nevertheless, most of them rely on strong assumptions and therefore have a quite limited range of application: The assumptions made in [23] for example, are only valid if the sun is the only significant light source. Besides, [14] can only handle specific changes in illumination. Consequently, common dense matching approaches are neither sufficient for solving multimodal fusion nor for matching under varying illumination conditions, because they are based on assumptions which do not hold under these conditions.

3 Method

3.1 Imaging Process

Infrared (IR) thermal imaging, captures radiation in the electromagnetic spectrum from approximately 0.9 to 14 μm. All objects with a temperature above 0 K emit thermal radiation. Because most thermal cameras operate surrounded by atmospheric gases, only radiation can be reasonably used for thermal imaging, which line up with the atmospheric window, i.e. is not absorbed by the atmosphere. As a result, there are only two ranges of IR wavelength which are typically used for thermal imaging: The short or medium wavelength band (SW/MW) and the long wavelength band (LW). The general setup of a thermal camera is very similar to the one of a typical RGB camera which also consists of a lens focussing radiation on a detector arranged as a focal plane array. Consequently, also similar mathematical models are used to describe the imaging process of a thermal camera with respect to reflection, refraction, and transmission. As summarized in [22], there are multiple reasons why many state of the art computer vision algorithms have a worse performance on thermal images than on RGB images. One major problem is more significant Gaussian image noise: In RGB imaging most objects only reflect light, and the resulting brightness level can typically be covered by a common exposure/gain level for the dynamic range of the camera. In contrast, in thermal imaging all objects with a temperature above 0 K emit thermal radiation, and thermal imaging typically has to deal with a much higher dynamic range. Additionally, the emission of an object in the thermal spectrum depends on the object's temperature. Finally, the appearance of an

object changes over time due to thermal balancing effects causing significant changes in the appearance of the object in the thermal image.

As described in this section, RGB and thermal imaging capture different wavelengths of the electromagnetic spectrum and therefore represent different aspects of their environment. Moreover, material transitions often provide different appearances in the visual spectrum as well as in the thermal image due to different thermal capacity and emissivity. Even though, because of the similar properties of visual and thermal radiation regarding reflection, refraction, and transmission, visual and thermal images not only depend on the reflected or emitted radiation in the specific ranges only, but also on the scene geometry from which light is reflected or emitted. These common factors are used in our method to establish multimodal image correspondences.

3.2 Multimodal Dense Matching

In general, it is advantageous for dense image matching if the images are stereo-rectified beforehand, as then the task of correspondence determination is reduced from a two-dimensional problem to one where homologous features lie in horizontal lines. To perform this kind of rectification, the cameras' intrinsic parameters as well as their relative orientation are assumed to be known. Subsequently, an approach based on Semi-Global Matching [9] and a combined cost function [24] is applied to estimate disparity maps for the thermal as well as the RGB image.

Combined Cost Function In order to densely reconstruct an environment from images that depict it in greatly different manners, a robust matching approach is crucial. For RGB and thermal images, it will typically not be possible to linearly transform the grey values from one domain to the other. Consequently, either a metric is needed which is robust against these kinds of differences or a transformation has to be used which brings both images into a common representation. In this work, a cost function is introduced which considers both approaches. For this purpose, various metrics are combined:

$$C(x, d) = \sum_{n} \lambda_n \cdot C_n(x, d)\,. \tag{1}$$

The combined cost function is computed for each pixel x and disparity d and is defined as the weighted sum of the individual functions. The weights λ_n represent the confidence of the corresponding cost functions in the current pixel. All weights add up to one. More information on the weights is provided later in this section.

In order to combine the response of the individual metrics to a consistent cost function, the value range of the responses has to be considered. In this context, not only the size of the support region for cost aggregation is relevant, but also the type of similarity measure used. For example, when comparing a Census filter plus Hamming distance with a SAD metric, the responses are not only within various intervals but also show different distributions over these

intervals. Consequently, a sum of these responses, weighted or not, can lead to a situation in which the influence of one metric is canceled out by the dominating one. Hence, the metric responses are normalized within $[0;1]$ and spread over the whole interval.

Metrics The combined cost function consists of four elements: a modified Census transformation (MC), Zero-mean Normalized Cross-Correlation (ZNCC), Normalized Sum of Squared Differences (NSSD) and a triangle-based depth prediction approach. All of them are applied to the images transformed by phase congruency [18] instead of the original images. This procedure is described in detail in the subsequent section.

Proposed in [34], the modified Census transformation extends the original concept [30] by intensifying the use of cross-correlation information. For this purpose, the pixels within the support region are not only compared to the center pixel, but to the mean value of the region as well. Keeping in mind that the image pair consists of a thermal and a RGB image, the grey value distribution in both images may differ greatly. Consequently, the corresponding maps obtained by phase congruency may vary also. Therefore, this modified version of the Census transformation better suits our application.

With respect to these properties, the sum of squared differences has to be modified as well, in order to be able to apply this metric to these kind of image pairs effectively. For this purpose, the results of the metric are assumed to be normally distributed and the standard score is utilized for normalization:

$$NSSD(x,d) = \frac{1}{\gamma \cdot |X_L|} \sum_{\tilde{x} \in X_L} \left(\frac{I_L(\tilde{x}) - \mu_L}{\sigma_L} - \frac{I_R(\tilde{x}-d) - \mu_L}{\sigma_R} \right)^2 , \tag{2}$$

$$C_{NSSD}(x,d) = min\left(NSSD(x,d), 1\right) , \tag{3}$$

where $|X_L|$ is the number of pixels within the support region of x in the left image and $\tilde{x}$ are the single elements. The mean and standard deviation of the support region in the corresponding image are denoted as μ and σ, respectively. Finally, the parameter γ is used to normalize the resulting value before truncation in Equation 3.

The last element of the combined cost function is a triangle-based depth prediction which is based on the approach originally proposed in [3]. In order to adjust this process to the specified data, the images are transformed via phase congruency prior to feature detection. Consequently, the moments of this transformation are used for the detection step [19]. Afterwards, the feature points are matched using the edge histogram descriptor proposed in [25] and the matching strategy suggested in [24]. Subsequently, a Delaunay triangulation is applied to the features for both images individually. In a final step, a disparity is predicted for every pixel within the triangles by interpolating the disparity values of the corresponding vertices. Thus, the corresponding cost function is defined as:

$$C_T(x,d) = min\left(\frac{|d - d_{T,x}|}{d_0}, 1 \right) , \tag{4}$$

Fig. 1: RGB and thermal images with their corresponding phase congruency.

where $d_{T,x}$ is the interpolated disparity at pixel x within triangle T and d_0 is the threshold for the maximum distance to the prediction. To avoid wrong predictions, triangles that are not surface consistent must be filtered out. Here, this is done via three criteria: The number of pixels within a triangle, the maximum edge length and the inclination of the triangle relative to the image plane.

Phase Congruency Originally proposed in [18], phase congruency is an image transformation which allows a representation which is invariant to differences in illumination and contrast. The image is then analyzed in the frequency domain within a local context and is based on the concept of the local energy model. In contrast to many other approaches that operate in the frequency domain, phase congruency utilizes phase information, not amplitudes. While [25] has already shown its capability to operate on thermal RGB image pairs in order to find sparse correspondences, within this work the transformation is used to convert the images into a representation that allows the use of common dense matching approaches.

Due to the Time-Frequency Uncertainty Principle it is not possible to accurately determine the spatial position and frequency simultaneously. Thus, the conventional approach is applied, utilizing a bank of Log-Gabor filters to approximate the phase congruency. The bank contains filters for various scales n and orientations Θ which are applied on an image $I(x)$ with:

$$PC(x) = \frac{\sum_\Theta \sqrt{(\sum_n (I(x) * M^e_{n\Theta}))^2 + (\sum_n (I(x) * M^o_{n\Theta}))^2}}{\sum_\Theta \sum_n (A_{n\Theta}(x)) + \epsilon}\,, \tag{5}$$

$$A_{n\Theta}(x) = \sqrt{(I(x) * M^e_{n\Theta})^2 + (I(x) * M^\Theta_{n\Theta})^2}\,, \tag{6}$$

where $PC(x)$ is the phase congruency value and $A_{n\Theta}(x)$ the amplitude of the response in pixel x. The even and odd symmetric components of the Log-Gabor filters are denoted as $M^e_{n\Theta}$ and $M^o_{n\Theta}$, respectively. In order to prevent a division by zero, ϵ is added to the denominator in Equation 5.

After phase congruency was applied on two images, a matching metric $F(*,*)$ is used to compute the similarity of the transformation results I^L_{PC} and I^R_{PC}. The cost function is then constructed as follows:

$$C_{PC}(x,d) = F(I^L_{PC}(x), I^R_{PC}(x-d))\,. \tag{7}$$

Besides, the result of the phase congruency itself, the maximum and minimum moment of the transformation can be utilized as well. By filtering these values with certain thresholds, edges and corners can be extracted [19]. As it is directly based on the transformation, the invariance against changes in illumination and contrast also applies here.

$$M = \frac{1}{2}\left(c + a + \sqrt{b^2 + (a-c)^2}\right), \tag{8}$$

$$m = \frac{1}{2}\left(c + a - \sqrt{b^2 + (a-c)^2}\right). \tag{9}$$

In order to determine edges and corners, the values of the moments are compared to two thresholds ϵ_E and ϵ_C: If the maximum moment $M > \epsilon_E$ in a certain point, this point is labeled as 'edge' and if $M > \epsilon_E$ and the minimum moment $m > \epsilon_C$ the point is labeled as 'corner'. The corresponding coefficients are defined as:

$$a = \sum_{\Theta} \left(PC(\Theta)cos(\Theta)\right)^2, \tag{10}$$

$$b = 2\sum_{\Theta} \left(PC(\Theta)cos(\Theta) \cdot PC(\Theta)sin(\Theta)\right), \tag{11}$$

$$c = \sum_{\Theta} \left(PC(\Theta)sin(\Theta)\right)^2, \tag{12}$$

where $PC(\Theta)$ represents the phase congruency value regarding only orientation Θ, but all scales. The sum is then calculated for all the orientations used.

Confidence based weighting The influence of the different metrics utilized within the combined cost function is controlled by their weights. Thus, the selection of suitable weights is a crucial task to obtain accurate results. While the approach proposed in [24] uses constant weights, the significance of the individual metrics can vary greatly, when operating on different parts of an image. Consequently, constant weights can only be a compromise over all pixels and in general, they will not be optimal.

To overcome this problem, a dynamic weighting scheme is proposed in this work. For this purpose, the weights of the single metrics are adjusted pixel-wise based on their confidence for the current pixel. To do so, in [11] different approaches were evaluated: Some are based on local attributes of the curve corresponding to the cost function, others analyze the entire curve and a third group utilizes the left-right consistency assumption. However, most of these concepts are not suitable to compare different kinds of metrics. Furthermore, in general, a more complex analysis leads to more reliable results. Thus, we propose to not only consider characteristics of the resulting cost curve but to introduce expectations on an 'ideal' cost curve, as well. For this purpose, the confidence measure is based on the difference between the curve of the cost function and a predefined reference curve:

$$\rho(x) = \begin{cases} \sqrt{\frac{1}{n}\sum_{d}^{n}(R(d) - C(x,d))^2}, & \text{if } |c_0 - c_1| > \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

$$R(d) = 1 - \exp(-\omega \cdot (d - c_0)^2)\,. \tag{14}$$

To ensures the uniqueness of the solution, the confidence ρ of a metric is only computed if the difference between the smallest (c_0) and the second smallest (c_1) value of the cost function is bigger than a predefined threshold ϵ. Otherwise, it is set to zero. The confidence itself is defined as the RMS-Error between the cost function C and the reference function R, for which R is specified by its extension ω and minimum, which is placed at c_0. The proposed confidence measure is used for all matching metrics except for the triangle-based depth prediction. Based on the definition of the prediction approach, the corresponding cost curve always has the same shape. Consequently, the comparison to a reference curve would result in a constant value. Instead, the distance $g(x, T)$ between the current pixel and the closest vertex of triangle T is used to measure the confidence as proposed in [3]:

$$\rho_T(x) = \exp\left(-\frac{g(x,T)}{\sigma}\right)\,, \tag{15}$$

where σ is a non-negative constant that controls the degree to which the confidence value descends. Finally, the weights λ are the confidence values normalized over the sum of all confidences:

$$\lambda_n = \frac{\rho_n}{\sum \rho}\,. \tag{16}$$

Optimization and Post-processing In the final step, the optimal disparity value for every pixel is determined by optimizing the cost volume produced by the combined cost function. For this purpose, Semi-Global Matching [9] is used. As suggested in the original work, gradient information is introduced to adjust the penalties. For this purpose, the edge map extracted from the maximum moment of the phase congruency is utilized. The resulting disparity maps are post-processed by filtering for speckles first and applying a left-right consistency check afterwards. To be able to demonstrate the capability of the proposed approach only, no interpolation is applied subsequently.

4 Evaluation

4.1 Experimental Setup

To evaluate our approach, we use a *FLIR A655sc* thermal camera, a *FLIR Grasshopper3 GS3-U3-23S6C* RGB-camera, and a *Velodyne HDL64E S2* LiDAR sensor. All sensors are rigidly connected. The baseline between the camera

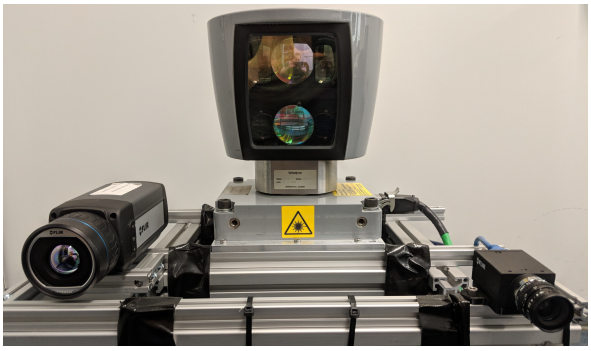

Fig. 2: Setup to evaluate our multimodal dense matching approach: FLIR A655sc *(left)*, Velodyne HDL64E S2 *(center)*, FLIR Grasshopper3 GS3-U3-23S6C *(right)*.

centers of the RGB and the thermal camera has a length of 0.393 m. The resulting setup is shown in Fig. 2. The RGB camera works at a framerate of 7 Hz whereas the thermal camera captures images at 50 Hz. Because both cameras are working untriggered at different measurement rates, time synchronization is performed according to [16] using global timestamps. The resulting maximum time difference between the images can be computed to be less than 10 ms. The extrinsic calibration between the cameras and the laser scanner was performed according to [32]. To determine the extrinsic calibration between the cameras, we used a multimodal chessboard with patterns with specific color and thermal emissivity properties, which can be clearly recognized in both imaging modalities as described in [16]. Furthermore, to better distinguish the patterns of the chessboard in the thermal image, the chessboard is actively heated using heat pads applied on the backside of the patterns. Additionally, the patterns are coated by materials with varying emissivity.

The presented approach is evaluated using a dataset consisting of 15 multimodal stereo images taken outside. The dataset covers artificial structures as buildings and cars as well as natural vegetation. Besides minor variations in the scene caused by environmental influences as wind, the scenes are assumed to be static. Furthermore, the distance to the objects in the scene varies between 2 m and 50 m, with a median of 5 m.

4.2 Error Metric

In order to evaluate the depth estimation error of our approach, lidar data is utilized as ground truth. For this purpose, the acquired 3D point clouds are projected into image space using the results of the intrinsic and extrinsic calibration:

$$ {}_{(RGB)}\mathbf{x} = {}_{(RGB)}\mathbf{M} \cdot {}^{(RGB)}\mathbf{T}_{(L)} \cdot {}_{(L)}\mathbf{X} \,, \tag{17} $$

where $\mathbf{X}$ is a 3D point and $\mathbf{x}$ its 2D correspondence. ${}^{(RGB)}\mathbf{T}_{(L)}$ is the transformations from the lidar to the RGB camera coordinate system. Finally, ${}_{(RGB)}\mathbf{M}$

is the projection matrix of the camera. To compute the 2D image coordinates, Euclidean normalization is applied. The error is then estimated for every pixel that is set in the ground truth image obtained by the laser scanner. Furthermore, a pixel is considered as correct if the distance between the dense image matching result and the ground truth is lower than a specified error bound. Lastly, the overall error is defined as the percentage of pixels which are considered as incorrect.

In general, an error bound of 20 % of the depth was used. This value was chosen due to the challenging setup and associated lower expected accuracy compared to traditional dense image matching applications. It is mainly used to compare our approach against others (see Table 1) and to demonstrate the contribution of the different components (see Fig. 4 on the right). Additionally, a more differentiated evaluation of the error is given in Fig. 4 on the left, providing results with varying error bounds in the range of 5 % - 25 %.

4.3 Results

Based on the error metric described above, the results were evaluated in a qualitative as well as a quantitative manner. On average 63.4 % of the estimated disparity values are classified as correct. As can be seen in Fig. 3, especially for structured areas and borders between different materials depth can be reconstructed accurately. This is based on the fact, that both are clearly visible in thermal as well as in RGB images. Conversely, textured but flat surfaces like facades and asphalt are challenging for this kind of application. Because of their varying appearance, they are not or only insufficiently visible in the spectrum covered by thermal imaging. This can also be observed when comparing the phase congruency maps obtained from both modalities shown in Fig. 1: The structure of the wall, as well as the ground is only visible in the transformed RGB image. Moreover, Fig. 3 and Table 1 show that the estimated disparity maps do not cover the complete image and consequently, not all ground truth points. This is based on the fact that no interpolation is performed at the end. Consequently, depth cannot be reconstructed in occluded areas as well as for pixels at edges which are only visible in one image.

Based on the error metric described in the previous section, we have also evaluated our approach with different error bounds for the depth. As can be seen in Fig. 4 on the left side, utilizing an error bound of 20 %, 63.4 % of the available ground truth points can be reconstructed correctly, and more than 40 % of the points have a difference between estimation and ground truth which is smaller than 10 % of their distance to the camera. Furthermore, the right side of Fig. 4 shows the importance of every single metric for the overall approach: Using Zero-mean Normalized Cross-Correlation (ZNCC) exclusively, only 51.5 % of the depth estimations are correct. Compared to this baseline, the combined cost function approach increases the performance by more than 10 % to 63.4 %.

To demonstrate the advantages of our approach regarding multimodal images, we compared our method against conventional dense matching approaches based on the error bound of 20 %. As shown in Table 1, conventional methods

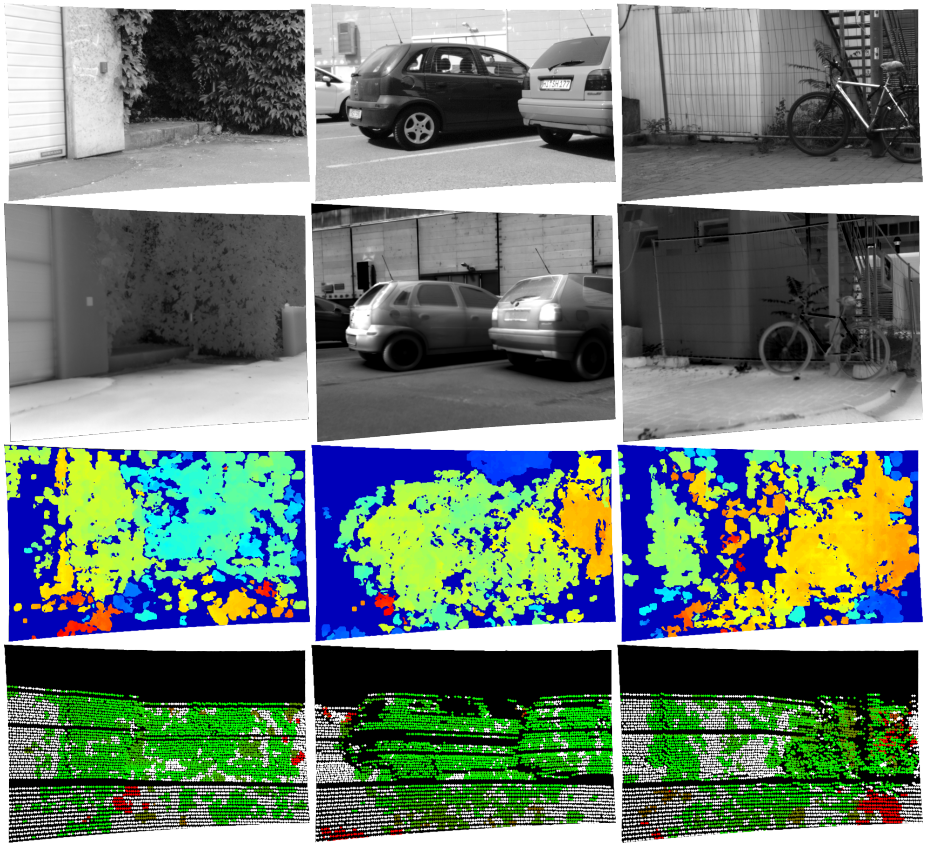

Fig. 3: Qualitative results of the proposed approach. From top to bottom: RGB image, thermal image, disparity map (from large values in red to small values in blue) and error map (from a small error in green to a large one in red - white points are not covered by the estimation). The disparity and error maps are related to the RGB image.

Table 1: Comparison against conventional dense matching approaches

Method	**Correctly reconstructed pixels in %**	**Coverage of ground truth points in %**
Elas Matcher [5]	15.1	0.3
SPS Stereo [29]	15.4	94.8
Illum. invariant matching [24]	54.5	23.2
Ours	63.4	60.5

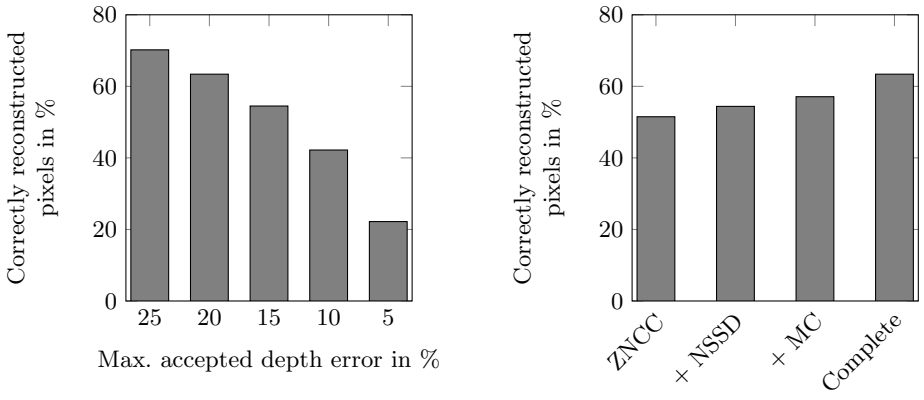

Fig. 4: Left: Percentage of pixels with a correct depth estimation according to varying thresholds which are based on the ground truth data. Right: Comparison of the performance of our approach using subsets of the proposed metrics.

are not suitable for this kind of data in general. With a coverage of ground truth points of less than 1 %, the Elas Matcher [5] even failed completely. This clarifies once again the challenge that the combination of RGB and thermal images presents to dense image matching.

5 Conclusion

In this paper, we have presented a new approach for multimodal dense depth estimation based on stereo images of an RGB- and a thermal camera. To solve this task we developed a cost function combining different robust metrics which are applied to an illumination independent image representation. Furthermore, we introduced a new confidence based weighting scheme which allows a pixel-wise weight adjustment within a cost function. We evaluated our approach based on 15 multimodal image pairs including ground truth data of a 64 channel lidar sensor and demonstrated, that our approach correctly estimates disparity values for on average 63.4 % of the estimated disparity values.

To further examine the performance of the presented approach, a more diverse dataset including additional lighting and temperature variations would be beneficial. Additionally, a common public multimodal dataset including ground truth data would allow a more competitive comparison.

As far as the authors know, we presented the first dense matching approach designed for the special requirements resulting from multimodal stereo setups. As shown in our experiments, the presented approach outperforms current state of the art dense matching methods regarding depth estimation based on multimodal images.

Acknowledgements

This work was supported by the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens [GRK2159] and the MOBILISE initiative of the Leibniz Universität Hannover and TU Braunschweig.

References

1. Alldieck, T., Bahnsen, C.H., Moeslund, T.B.: Context-Aware Fusion of RGB and Thermal Imagery for Traffic Monitoring. Sensors **16**(11) (2016). https://doi.org/10.3390/s16111947
2. Bhanu, B., Han, J.: Kinematic-based Human Motion Analysis in Infrared Sequences. Sixth IEEE Workshop on Applications of Computer Vision - Proceedings (2002)
3. Bulatov, D., Wernerus, P., Heipke, C.: Multi-view Dense Matching supported by Triangular Meshes. ISPRS Journal of Photogrammetry and Remote Sensing **66**(6), 907–918 (2011)
4. Conaire, C., Cooke, E., O'Connor, N., Murphy, N., Smearson, A.: Background Modelling in Infrared and Visible Spectrum Video for People Tracking. IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (2005)
5. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale Stereo Matching. In: Asian Conference on Computer Vision. pp. 25–38. Springer (2010)
6. Guo, L., Zhang, G., Wu, J.: Infrared Image Area Correlation Matching Method Based on Phase Congruency. International Conference on Artificial Intelligence and Computational Intelligence (2010)
7. Han, J., Bhanu, B.: Fusion of Color and Infrared Video for Moving Human Detection. Pattern Recognition **40**(6), 1771–1784 (2007). https://doi.org/10.1016/j.patcog.2006.11.010
8. Heather, J.P., Smith, M.I.: Multimodal Image Registration with Applications to Image Fusion. In: 7th International Conference on Information Fusion. vol. 1, pp. 8 pp.– (2005). https://doi.org/10.1109/ICIF.2005.1591879
9. Hirschmuller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2), 328–341 (2008)
10. Hrkac, T., Kalafatic, Z., Krapac, J.: Infrared-visual Image Registration Based on Corners and Hausdorff Distance. In: Proceedings of the 15th Scandinavian Conference on Image Analysis. pp. 383–392. SCIA'07, Springer-Verlag, Berlin, Heidelberg (2007)
11. Hu, X., Mordohai, P.: A Quantitative Evaluation of Confidence Measures for Stereo Vision. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2121–2133 (2012)
12. Istenic, R., Heric, D., Ribaric, S., Zazula, D.: Thermal and Visual Image Registration in Hough Parameter Space. In: 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. pp. 106–109 (2007). https://doi.org/10.1109/IWSSIP.2007.4381164
13. James, A.P., Dasarathy, B.V.: Medical Image Fusion: A survey of the state of the art. CoRR **abs/1401.0166** (2014)

14. Kim, S., Ham, B., Kim, B., Sohn, K.: Mahalanobis Distance Cross-Correlation for Illumination-Invariant Stereo Matching. IEEE Transactions on Circuits and Systems for Video Technology **24**(11), 1844–1859 (2014)
15. Kleinschmidt, S.P., Wagner, B.: Probabilistic Fusion and Analysis of Multimodal Image Features. 18th International Conference on Advanced Robotics pp. 498–504 (2017)
16. Kleinschmidt, S.P., Wagner, B.: Spatial Fusion of Different Imaging Technologies Using a Virtual Multimodal Camera. In: Madani, K., Peaucelle, D., Gusikhin, O. (eds.) Informatics in Control, Automation and Robotics, pp. 153–174. Springer (2018)
17. Kleinschmidt, S.P., Wagner, B.: Visual Multimodal Odometry: Robust Visual Odometry in Harsh Environments. In: IEEE International Symposium on Safety, Security and Rescue Robotics (2018)
18. Kovesi, P.: Image Features from Phase Congruency. Videre: Journal of Computer Vision Research **1**(3), 1–26 (1999)
19. Kovesi, P.: Phase Congruency detects Corners and Edges. In: The australian pattern recognition society conference: DICTA. pp. 309–318 (2003)
20. Krotosky, S.J., Trivedi, M.M.: Mutual Information Based Registration of Multimodal Stereo Videos for Person Tracking. Computer Vision and Image Understanding **106**(2-3), 270–287 (2007). https://doi.org/10.1016/j.cviu.2006.10.008
21. Leinonen, I., Jones, H.G.: Combining Thermal and Visible Imagery for Estimating Canopy Temperature and Identifying Plant Stress. Journal of Experimental Botany **55**(401), 1423–1431 (2004)
22. Lin, S.S.: Review: Extending Visible Band Computer Vision Techniques to Infrared Band Images. Tech. Rep. MS-CIS-01-04, GRASP Laboratory, Computer Vision and Information Science Department, University of Pennsylvania (2001)
23. Maddern, W., Stewart, A., McManus, C., Upcroft, B., Churchill, W., Newman, P.: Illumination Invariant Imaging: Applications in Robust Vision-based Localisation, Mapping and Classification for Autonomous Vehicles. In: IEEE International Conference on Robotics and Automation - Workshop Proceedings. vol. 2, p. 3 (2014)
24. Mehltretter, M., Heipke, C.: Illumination Invariant Dense Image Matching based on Sparse Features. In: 38. Wissenschaftlich-Technische Jahrestagung der DGPF und PFGK18 Tagung in München. vol. 27, pp. 584–596 (2018)
25. Mouats, T., Aouf, N.: Multimodal Stereo Correspondence based on Phase Congruency and Edge Histogram Descriptor. In: Proceedings of the 16th International Conference on Information Fusion. pp. 1981–1987. IEEE (2013)
26. Raza, S., Sanchez, V., Prince, G., Clarkson, J., Rajpoot, N.M.: Registration of Thermal and Visible Light Images of Diseased Plants using Silhouette Extraction in the Wavelet Domain. Pattern Recognition **7**(48) (2015)
27. Shah, P., Merchant, S.N., Desai, U.B.: Fusion of Surveillance Images in infrared and visible Band using Curvelet, Wavelet and Wavelet Packet Transform. International Journal of Wavelets, Multiresolution and Information Processing **8**(2) (2010)
28. Vidas, S., Moghadam, P.: HeatWave: A handheld 3D Thermography System for Energy Auditing. In: Energy and Buildings. vol. 66, pp. 445 – 460 (2013)
29. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In: European Conference on Computer Vision. pp. 756–771. Springer (2014)
30. Zabih, R., Woodfill, J.: Non-Parametric Local Transforms for Computing Visual Correspondence. In: European Conference on Computer Vision. pp. 151–158. Springer (1994)

31. Zbontar, J., LeCun, Y.: Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. Journal of Machine Learning Research **17**(1-32), 2 (2016)
32. Zhang, Q., Pless, R.: Extrinsic Calibration of a Camera and Laser Range Finder. In: International Conference on Intelligent Robots and Systems. pp. 2301–2306 (2004)
33. Zhao, J., Cheung, S.c.S.: Human Segmentation by geometrically fusing Visible-Light and Thermal Imageries. Multimedia Tools and Applications **73**(1), 61–89 (2014). https://doi.org/10.1007/s11042-012-1299-2
34. Zhu, S., Yan, L.: Local Stereo Matching Algorithm with Efficient Matching Cost and Adaptive Guided Image Filter. The Visual Computer **33**(9), 1087–1102 (2017)