Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

HighTech and Innovation Journal

Vol. 4, No. 3, September, 2023

Trainable Regularization in Dense Image Matching Problems

Vladimir Zh. Kuklin [1], Aslan A. Tatarkanov [1*], Alexander A. Umyskov [3]

[1] *Institute of Design and Technology Informatics of RAS, Russian Federation.*

Received 28 May 2023; Revised 11 August 2023; Accepted 22 August 2023; Published 01 September 2023

Abstract

This study examines the development of specialized models designed to solve image-matching problems. The purpose of this research is to develop a technique based on energy tensor aggregation for dense image matching. This task is relevant within the framework of computer systems since image comparison makes it possible to solve current problems such as reconstructing a three-dimensional model of an object, creating a panorama scene, ensuring object recognition, etc. This paper examines in detail the key features of the image matching process based on the use of binocular stereo reconstruction and the features of calculating energies during this process, and establishes the main parts of the proposed method in the form of diagrams and formulas. This research develops a machine learning model that provides solutions to image matching problems for real data using parallel programming tools. A detailed description of the architecture of the convolutional recurrent neural network that underlies this method is given. Appropriate computational experiments were conducted to compare the results obtained with the methods proposed in the scientific literature. The method discussed in this article is characterized by better efficiency, both in terms of the speed of work execution and the number of possible errors.

Keywords: Image Matching; Convolutional Recurrent Neural Network; Stereo Reconstruction; Method Error; Neural Network Architecture.

1. Introduction

When considering matching methods in detail, it is worth noting that their key characteristics include the general level of computational complexity and the quality of matching formed on the basis of real data [1, 2]. Nowadays, the most promising methods are those that involve the use of deep machine learning [3–5]. Machine learning is one of the two main categories that are classified for image matching [6, 7]. Their operation involves conducting preliminary training, for which a large sample of training data is used. The main disadvantage of this approach is the extremely high level of computational complexity, which prevents its use within a certain list of tasks [8, 9]. In the second category of methods, everything comes down to solving optimization problems [10], where a displacement field appears as a result of minimizing the target functional.

When considering problems in the field of computer vision, great attention is paid to special algorithms and a set of actions that help implement image matching [11, 12]. The need to match images is currently a priority task for electronic vision. This problem is present in most practical applications, for example, in binocular stereo reconstructions, as described in Zimiao et al. [13]. Binocular stereo reconstructions, which are based on estimating the displacement of right and left images taken by stereo cameras, usually arise from a binocular phenomenon. Image matching made through such reconstructions opens up the possibility of quantitatively assessing various characteristics of observed objects in the matched images.

* Corresponding author: as.tatarkanov@yandex.ru

http://dx.doi.org/10.28991/HIJ-2023-04-03-011

It should be noted that the general case of the image matching problem implies the simultaneous presence of several formulations, and the use of each specific formulation is determined by the application. For example, if it is necessary to perform a parametric comparison, the task of searching for a special transformation within the context of existing parametric transformations [14], for example, affine transformations [15], is implemented here. They can be used to match images based on different prospective distortions. As part of this research, the problem associated with nonparametric image comparison is considered in more detail [16]. This is the most general case of such a problem, the essence of which is that each pixel of the original image receives some independent transformation, and at the same time, there is a proportional relationship between the total number of degrees of freedom and the number of pixels. Another significant factor is the specific way in which similar images are matched. At first, it is required to consider sparse matching, where the focus is on matching individual, specific image elements.

This study also carefully considers dense matching. This approach involves comparing all existing image pixels, after which a two-dimensional displacement field is formed, which is the solution to the problem. Using this field, the transformation of each existing pixel of the image is determined. This article provides a detailed description of the image-matching problem. It inherently involves conducting binocular stereo reconstruction based on the formation of estimates for the displacements of the left and right images obtained using a stereo camera, and the formation of these images is ensured by the binocular effect.

2. Literature Review and Analysis

From the perspective of creating computer vision involved in many applications, this study will be reduced to comparing several or a certain sequence of images, which is described in detail in Wang et al. [17]. Three-dimensional stereo reconstruction based on two images from a stereo camera illustrates tasks that involve image matching.

One of the key features of human visual perception is the difference between the images formed by the left and right eyes. In this case, one should consider the option of displacing the object in relation to each eye, while the value describing the displacement will be the reciprocal of the distance from the researcher's pupil to the object in question. Knowledge of such a feature can be used to implement the stereo reconstruction principle. As a result, all these factors make it possible to generate three-dimensional scene geometry by accurately assessing the depth of each pixel in the image [18, 19].

The specific value of the level of horizontal displacement of objects between the views of the stereo camera on the right and left sides is inversely proportional to the distance to the specific observed object. In this regard, there is a need to generate a description for the camera model. Within the framework of Scharstein et al. [20], an exact description is given for the point model of the camera. Simultaneously, the studies emphasize that the greatest difficulty in the process of implementing stereo reconstruction is the formation of an assessment for a one-dimensional displacement field. Therefore, the principles of photogrammetry and the specifics of reconstructing a height map using aerial photography serve, perhaps, as the main driver for the accelerated improvement of stereophonic comparison techniques, which is described in Bisson-Larrivée and LeMoine [21] and Yang et al. [22]. Robotics has also begun to use stereophonic reconstructions more actively, as described in Pu et al. [23], which also affects most cutting-edge systems that provide unmanned control of cars, as described in Guan et al. [24].

It is important to understand that a key step in the reconstruction procedure is to match a pair of images, although the specifics of this process are determined on the basis of the camera model specification (Figure 1). The one-dimensionality of the displacement field is achieved on the basis of the results of image rectification, which is implemented through affine transformations. The problem associated with stereo matching is considered in detail in Xu et al. [25].

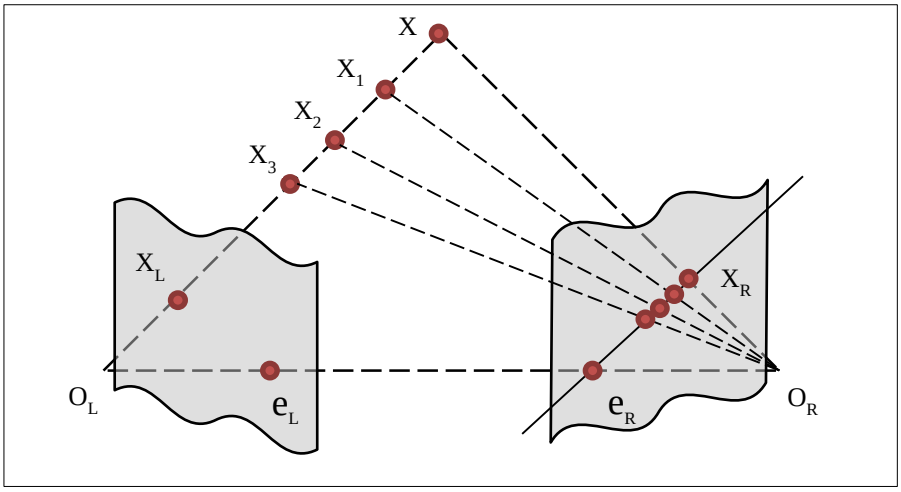

Figure 1. Main features of the geometry of two types of stereo cameras, which display the principles of image formation for the left and right views

3. Methodology and Research Results

3.1. Key Features of Neural Network Architecture

The task of stereo matching involves the formation of a specific assessment of the depth value of each available pixel, for which two images are analyzed simultaneously [26]. The reconstruction process will imply the determination of a specific disparity field, i.e., it is required to determine a horizontal displacement field. It is important to understand that, at its core, the disparity field will be one-dimensional. To calculate the disparity, provided that a point model of the camera is used, the following relationship must be applied:

$$z = \frac{fB}{d} \tag{1}$$

where f is a focal distance, and B is the distance between the centers of the cameras.

Next, we should consider specific problems that arise during the practical solution of stereo matching problems on real images. To begin with, it is necessary to consider that the images of the right and left stereo cameras always undergo certain random changes associated with the appearance of noise or various photometric changes. Moreover, it is crucial to consider the differences in lighting when observations are made from different points. In particular, a certain category of image sites will include extremely bright reflections of light sources and some objects that reflect light. It should also be considered that photographs of road scenes will contain a fairly large number of homogeneous areas without texture. Based on this feature, a conclusion can be made about a potentially large number of matches, where only one of them will correspond to some real three-dimensional object.

Another problem in this situation is the need to consider occluded objects. Obviously, solving such a problem requires reconstructing the depth of the objects presented in one of the views. It is important to understand that matching cannot be established on the basis of visual information alone; it will require the introduction of a certain list of additional restrictions in terms of the expected shape of objects. Difficulties also arise from perspective transformations because the appearance of objects will differ when they are photographed from different angles.

Next, it is proposed to generate a more accurate description for the convolutional recurrent neural network, which is required for fast stereo matching. This neural network was proposed by the author of this research.

This method involves a series of calculations that are similar in nature to the series of methods used for online stereo image matching, which is described in detail in the research literature. This is a dynamic programing technique used to aggregate the energy tensor, which is accomplished through a complex series of one-dimensional passes. In addition, this effect can be achieved through the use of special image filters that make it possible to consider specific boundaries of objects; this category includes controlled, recursive, and bilateral filters.

The essence of this technique is to use a special differentiable recursive filter, which was created on the principle of analogy between the calculation graph and the direct pass of the recurrent neural network. Noteworthy, for the first time, this approach was used to solve a problem in the field of semantic segmentation.

The proposed method uses specific convolutional neural networks. At the same time, it is essential to recall that machine learning in this case is used at the stages of combining power, which in the general case form a simulation that requires minimal calculations.

3.1.1. Key Features of the Energy Calculation Process during Stereo Matching

The study of this method showed that it involves storing multidimensional stereo matching tensors in memory cells in the form of multidimensional information arrays. To determine the energy, the sum of the two main terms should be found:

$$E(x,y,d) = \alpha E_{SAD}(x,y,d) + (1-\alpha)E_{census}(x,y,d) \tag{2}$$

where coefficient $\alpha \in (0,1)$ makes it possible to characterize the specific contribution of each term.

At its core, the first term acts as the absolute value of the difference in intensity of the available pixels. To determine this, the following formula should be used:

$$E_{SAD}(x,y,d) = \sum_{r,q,b} |I^L(x,y) - I^R(x-d,y)|, \tag{3}$$

Note that within this formula, all fragments have a dimensionality of 1x1, that is, the use of individual pixels is implied. This ensures the storage of information about the image texture features. The smoothing procedure is implemented only on basis of the results of energy tensor aggregation.

Now let us consider the features of obtaining the second term; this implies the process of matching local descriptors, which is described in detail in Zahiri-Azar and Salcudean [27]. The following algorithm of actions is required to determine this descriptor. If the image is black and white, it is necessary to define a function of the following form, considering pixels *p* and *q*:

$$\xi(p,q) = \begin{cases} 1, if\ I(q) < I(p) \\ otherwise\ \ 0 \end{cases} \quad (4)$$

It enables to define a census transformation, as a result of which each pixel will be matched with a multidimensional vector consisting of zeros and ones. This is described by the following formula:

$$R_r(p) = \bigotimes_{|i,j| \in D_w} (p, p + [i,j]), \quad (5)$$

where $\otimes$ is a concatenation operation and a D_w is a set of possible two-dimensional offsets.

Here, it is necessary to understand that each bit is determined by a careful matching of the activities of the main pixels of any gap with the intensity of the remaining pixels of the gap in question. As a result, the resulting descriptors specified by bit sequences, are compared using the Hamming distance. Figure 2 demonstrates the architecture of the convolutional edge detector in more detail.

Using the convolution operation, the neural network extracts the features from 5 image scales. The pooling operation is used to make the transition to a larger scale. The linear combination of neural network predictions is calculated using the last convolutional layer. Additionally, feature extraction can be achieved using direct connections, and the process is shown in more detail in Figure 2. Thus, this approach makes it possible to simultaneously consider data from five scales as part of the prediction process with *a* and *b* as binary vectors.

$$H(a,b) = \sum_i I(a_i \neq b_i). \quad (6)$$

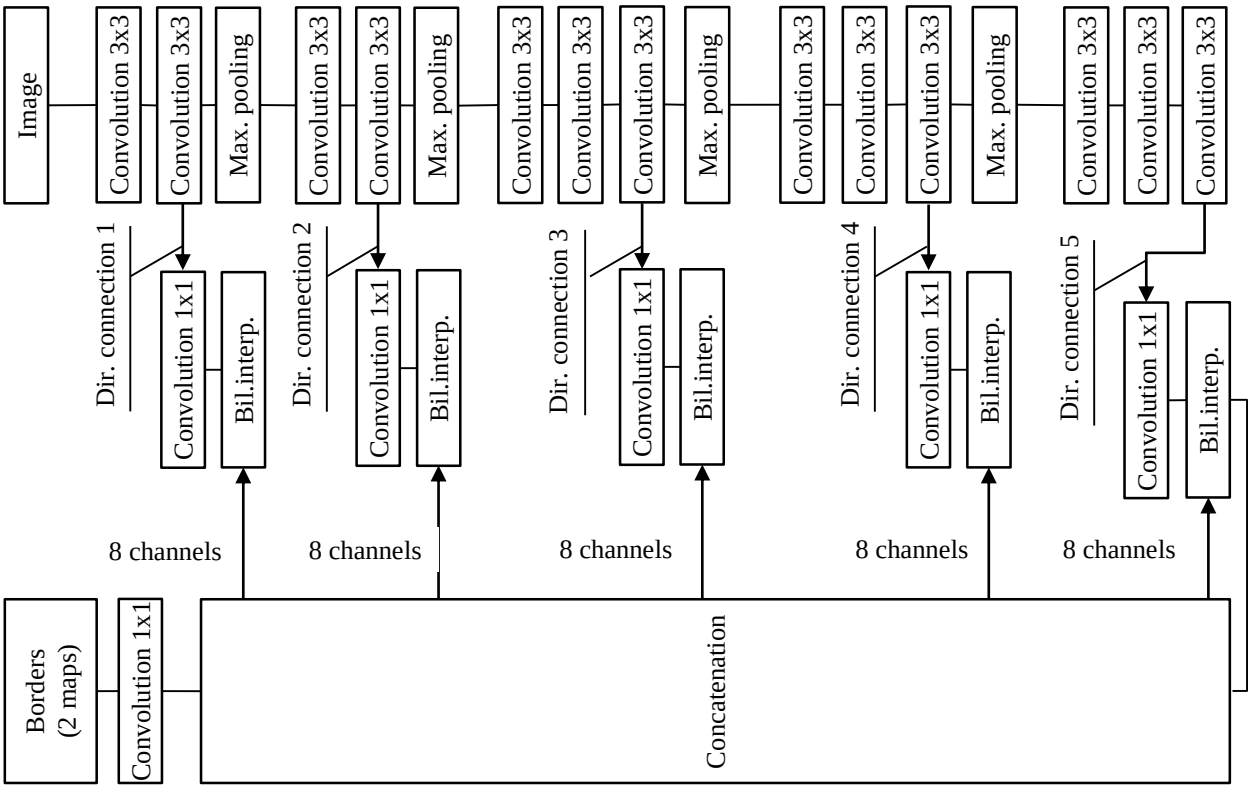

Figure 2. Main features of the architecture of the convolutional edge detector

This algorithm, which is used to determine a certain local descriptor, implies natural parallelization, which makes it possible to effectively implement this process within a graphics accelerator. This feature enables performing an operation in a specific constant time. The resulting descriptors given by the bit sequences are compared using the Hamming distance.

3.1.2. Characteristics of the Object Edge Detector within an Image

As in Scharstein et al. [20], the methodology under consideration implies the implementation of an energy aggregation process, which significantly depends on the characteristics of the input image. It is important to understand that a machine learning-based scheme will ensure that smoothing is done with regard to the specifics of the stereo matching task, and it will also consider all the key features of the training sample used. The essence of the technique is to determine the specific edges of objects that are relevant from the viewpoint of the difference in the disparity field. As a result, the method error can be significantly reduced using this approach.

3.1.3. Characteristics of the Recursive Filter-based Smoothing Process

A special recursive filter helps to most accurately consider the edges described in Wang et al. [26]. This filter is the basis for the energy aggregation method under consideration.

Such general recursive filtering cascades demonstrate better performance when applied to two-dimensional images in separate formats, which is determined by the fact that separation is established in a series covering each one-dimensional pass, which has its own directionality. When considering the principles described in section 3.1.4, it makes sense to use a number of separate weight cards to implement each pass. Below is a formula proposed to describe the functioning of a two-dimensional filter that receives an image and two weight maps and allows determination of the output image:

$$I_{filt} = F(I, W_h W_v). \tag{7}$$

Now, let us present a series of relations that are used for the algorithm calculating four recurrent passes:

$$I^L(x, y, d) = (1 - W_h(x, y))I(x, y) + W_h(x, y)I(x - 1, y), \tag{8}$$

$$I^R(x, y, d) = (1 - W_h(x, y))I^L(x, y) + W_h(x, y)I^L(x + 1, y), \tag{9}$$

$$I^T(x, y, d) = (1 - W_v(x, y))I^R(x, y) + W_v(x, y)I^R(x, y - 1), \tag{10}$$

$$I^B(x, y, d) = (1 - W_v(x, y))I^T(x, y) + W_v(x, y)I^T(x, y + 1), \tag{11}$$

To create trainable filtering simulations, it is necessary to ensure weight prediction using the input snapshots. More detailed descriptions of the functioning of an ordered chain of operations capable of performing backpropagation should be based on the hypothesis that the output values of the current operations will be fed to the input channels of some subsequent layers. Calculating the input filter gradient implies the use of the following relation:

$$\frac{\partial L}{\partial x_i} = (1 - w_i)\frac{\partial L}{\partial y_i}, \tag{12}$$

In turn, to obtain an exact value for the output gradient based on a certain set of weights, it is necessary to use a relation of the form:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial w_i} + (y_{i-1} - x_i)\frac{\partial L}{\partial y_i}, \tag{13}$$

The expression below makes it possible to define the calculation of the output gradient based on *y:*

$$\frac{\partial L}{\partial y_{i-1}} = \frac{\partial L}{\partial y_{i-1}} + w\frac{\partial L}{\partial y_i}, \tag{14}$$

It is important to understand that the four filter passes presented can be used in various combinations of the trained model. The thing is that the sequence of calculation of the presented equations will greatly determine the results, since the output of one expression is the input information for others.

3.1.4. Characteristics of the Main Features of the Energy Tensor Aggregation Process

The author of this research chose an approach that implies that the smoothing of the energy tensor is conducted with regard to the existing edges of objects within the image. This strategy is also used in Scharstein et al. [20]. Thus, this author's approach can be considered as a generalization of the research theses [28]. The use of a convolutional neural network makes it possible to predict filter parameters by the input image.

Figure 3 demonstrates in more detail the operation scheme of this aggregation algorithm. It should be understood that the filtration procedure itself involves the use of four directed passes.

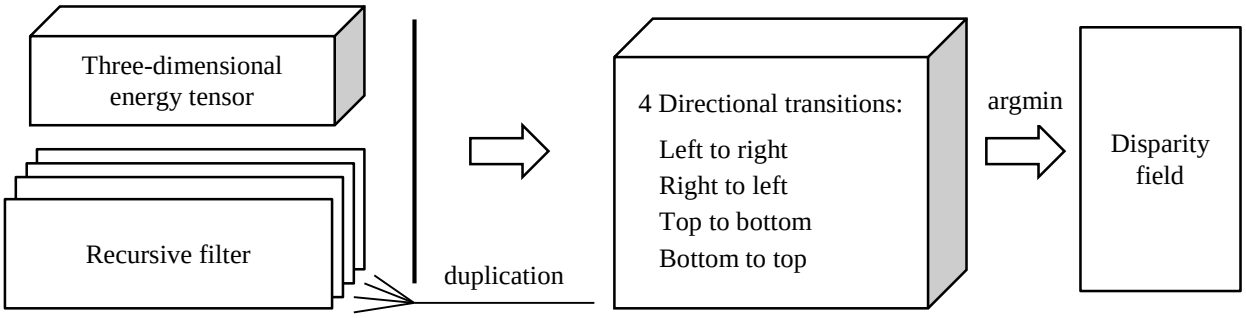

Figure 3. Structure of the energy tensor aggregation algorithm

Using the weight of the recurrent filter, all available two-dimensional slices will be filtered for three-dimensional energy tensors, which is also conducted using four directed passes. Note that slices are filtered in parallel.

$$E_d^{filt} F(E_d, W_h W_u), \tag{15}$$

$$d = \{0,1,\ldots,d_{max}\}\}$$

Deviation fields are calculated using established minimum elements relative to 3 dimensions.

The approach described in Hoskins and Svensson [28] involves the use of separate cards for vertical and horizontal passage. The essence of this choice is based on available practical observations, namely the fact that the frequency and magnitude of changes in disparity within the horizontal and vertical directions will differ considerably. In turn, by using two maps, an increase in the number of arithmetic operations during execution can be avoided, and the matching error can be significantly reduced. Reducing the computational complexity of the technique is achieved by using an edge detector in relation to images that are reduced to half the size of the existing original images. A huge effect is achieved through bilinear interpolation. The energy tensor is determined in the context of the original scale. When considering the process of using the outputs of a convolutional neural network as input data to a recurrent neural network, the following linear transformation is required:

$$W_h = exp(-\sigma E_h), \tag{16}$$

$$W_v = exp(-\sigma E_v), \tag{17}$$

where σ is the proportionality coefficient that provides adjustment, and the E_h and E_v are the convolutional neural network outputs.

3.1.5. Features of the Loss Function

Matching filtered tensors with ideal deviation fields assumes that each reference label will be represented by a delta function with a peak corresponding to the reference values. The effect is achieved through the softmax operation and the use of the cross-entropy function. The simulation training process is conducted using the backpropagation method. The loss dependence opens up the possibility of considering the difference between the calculated and ideal fields of a particular deviation. Figure 4 shows in more detail the diagram of a neural network that uses these operating principles. The inputs to the operation of calculating energy tensors will be represented by the right and left images coming from stereo cameras. The left one acts as a reference that is used by the convolutional neural network to make predictions for relevant edges. The peculiarity of the recursive filter is that it receives two weight maps and an energy tensor as input. Convolutional neural network filters are trained using the back propagation technique.

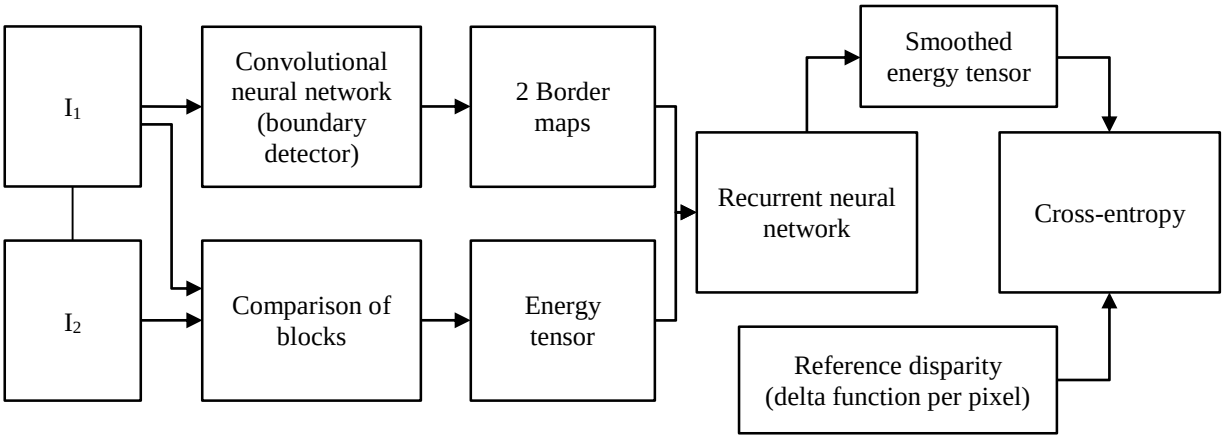

Figure 4. External view of the neural network diagram

3.2. Characteristics of the Conducted Numerical Experiments

3.2.1. Key Features of the Training Set

The algorithm proposed in this study was tested on an open collection of images. This collection includes pairs of photographs obtained from a stereo camera and undergone a rectification procedure. Each pair is characterized by certain reference disparity values, the calculation of which was based on depth data estimated using a laser scanner. The scanner was mounted on the roof of the motor vehicle, on which the camera was installed. Thus, within the framework of this approach, the reference disparity acts as a set of horizontal stripes, and therefore, the total amount of trainable data will decrease. Using heuristic approaches, it is possible to increase the amount of information; for example, the dilatation technique is effective. However, in practice, the use of additional factors is significantly complicated by the need to consider the huge number of distortions introduced by these changes.

3.2.2. Basic Principles of Methodology for Assessing Method Error

When considering stereo matching problems, we need to understand that several different methods can be used to estimate the error. The first approach involves determining a certain average absolute field error. In most cases, this strategy is used when an optical approach is applied. If we consider the second approach, it involves obtaining the exact number of pixels where the absolute error value exceeds a certain threshold. That is, the following formula is used:

$$e(D, D_{gt}) = \frac{1}{N}\sum_{(x,y)}(I[\left|D(x,y) - D_{gt}(x,y)\right| > t]), \quad (18)$$

where N characterizes the total number of pixels for which the displacement field is defined as D_{at}, and for this error estimation methodology, threshold value $t = 3$.

Many differences characterize the method in which errors are analyzed using deviation fields in relation to the distortions of the second image. Note that distorted images will be compared with ideal ones. Here, it is essential to exclude the number of pixels that will correspond to obscured objects from consideration.

3.2.3. Consideration of the Training Process

The set of images described earlier was further divided into a training set of 160 image pairs and a validation set of 40 image pairs. Three-channel color images were used as input data for the convolutional neural network. Figure 5 shows a quantitative comparison of the different techniques.

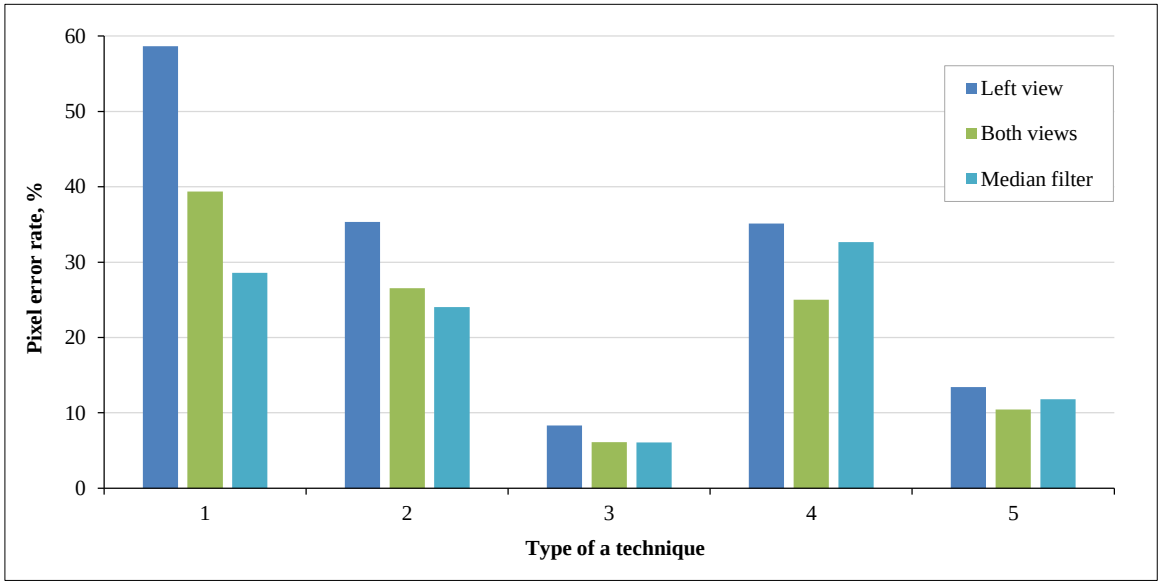

Figure 5. Quantitative comparison of various energy aggregation techniques: 1– without aggregation; 2– without training; 3– the proposed method; 4– the method presented in Scharstein et al. [20]; 5– the method of Scharstein et al. [20] + energy tensor

The energy tensor under consideration corresponds to a linear combination for the absolute value of the pixel-by-pixel difference. During the experimental evaluation of the linear search technique, the optimal value for the linear combination coefficient was determined to be 0.43.

In addition, we made a number of minor changes to the initial convolution architecture on which the edge detector relies. We could reduce the number of maps representing convolutional layers to reduce computational complexity. We also increased the total number of convolutional feature maps, the number of which increased from 1 to 8.

The cross-entropy error function was the basis for training a convolutional recurrent neural network; this function is described in more detail in Yang [29]. Preliminary training of the edge detector was conducted on a collection of images. Each image used has a size of 1242×375, while the range of acceptable disparity values fully corresponds to the value of 0-256 pixels. Figure 6 shows the contribution of combining neighboring pixel points when using trained general recursive filtering cascades.

Figure 6. Demonstration of smoothing process features using a recursive filter

Each fragment presented on the right side makes it possible to demonstrate some relative contribution of the central pixel to the overall intensity of the remaining pixels of the fragment. Fragments that are displayed in green and red exemplify marks or boundaries of the road surface.

3.2.4. Evaluation of the Total Number of Arithmetic Operations

During the algorithm operation, the aggregation of the energy tensor is the most labor-intensive stage. The following formula is used to calculate the complexity of this stage: $O(nd_{max})$, where n is the number of pixels in the image, and d_{max} is the largest disparity. The computational complexity of the fastest method is estimated as $O(nd_{max})$, where k is the dimensionality of the deep descriptors.

3.2.5. Duration of Execution on the Graphics Accelerator

This technique was implemented by combining the Theano framework and the effective use of procedures that allow the calculation of the energy tensor. The edges are calculated using a specialized dedicated library. The total training process takes approximately 4–5 hours. Figure 7 shows in more detail the time costs required to execute the different stages of the proposed method.

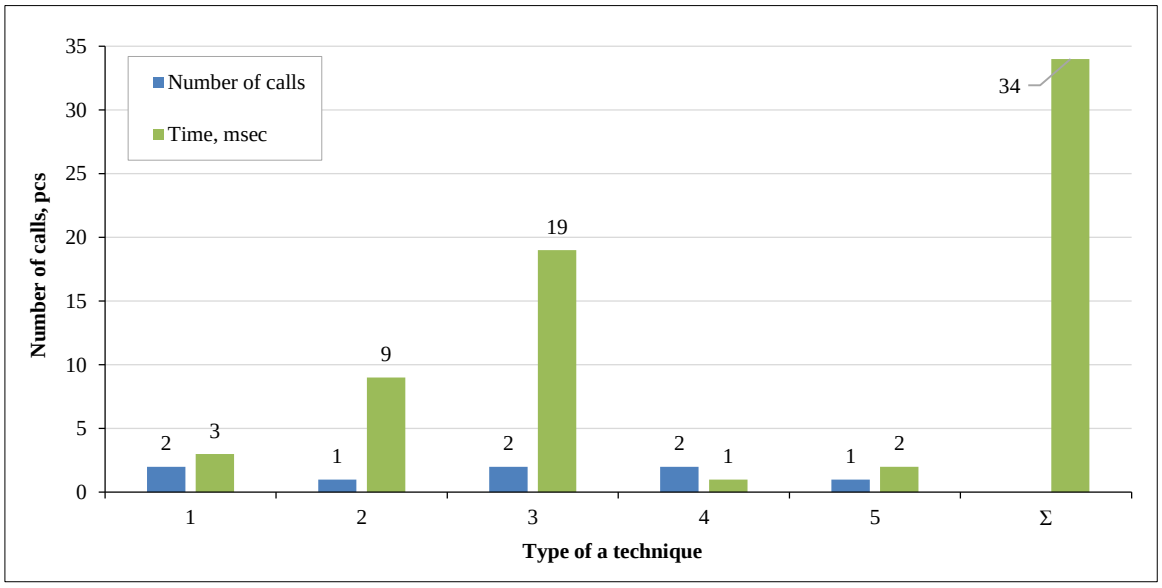

Figure 7. Time costs required to execute a parallel implementation of the graphics accelerator-based technique: 1– calculation of the energy tensor; 2– convolutional edge detector; 3– recursive filter; 4– argmin operation; 5– detection of layered objects; Σ– total time.

3.2.6. Main Features of Detecting Obscured Objects

It should be kept in mind that errors of certain deviations can only be perceived for a complete image, which presupposes the presence of a number of obscured objects. Therefore, it makes sense to determine the masks of each occluded object by calculating the deviation fields established for both types of stereo cameras. Any pixel will become the recipient of any current label based on the following requirements:

$$l(x,y) = \begin{cases} 0, \text{ if } |d - D^R(x-d,y)| \leq 1 \text{ for d} = D^L(x,y), \\ 1, \text{ if } |d - D^R(x-d,y)| \leq 1 \text{ for a certain d= } \in [0, d_{max} \\ \text{otherwise } 2 \end{cases} \quad (19)$$

where D^R and D^L is right and left assessment of disparities.

Interpolation of the disparity field is carried out using the following algorithm: in pixels with a zero label, the disparity remains unchanged. In pixels with label 2, the disparity value is determined on the basis of the nearest pixels with label 0, which are in the same row on the left. Pixels with label 1 get their values from the nearest pixels with label 0. Using this approach can significantly reduce the level of matching errors.

A flowchart from the workflow that briefly shows the process of the methodology is presented in Figure 8.

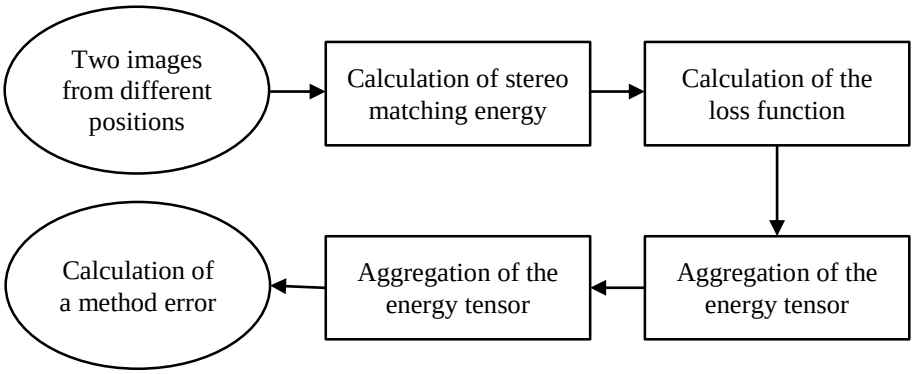

Figure 8. Flowchart of the methodology

4. Analysis and Discussion of the Results

A quantitative comparison of energy aggregation methods showed that even compared with the method proposed in the literature in combination with the energy tensor, the method developed in this study showed significantly better average performance. The proposed method proved to be 59% more effective in image matching, due to a reduction in the number of errors. An analysis of the proposed method for the number of errors showed that, compared to the existing method (when they are based on similar energy tensors), it has approximately 37% fewer errors during image matching. The histograms (Figure 9) present quantitative comparative characteristics for the energy aggregation methods in more detail. At the same time, the errors were determined for three main cases, namely, the error of the method, which is based only on the left view, the error caused by post-processing, and the error caused by the application of the median filter.

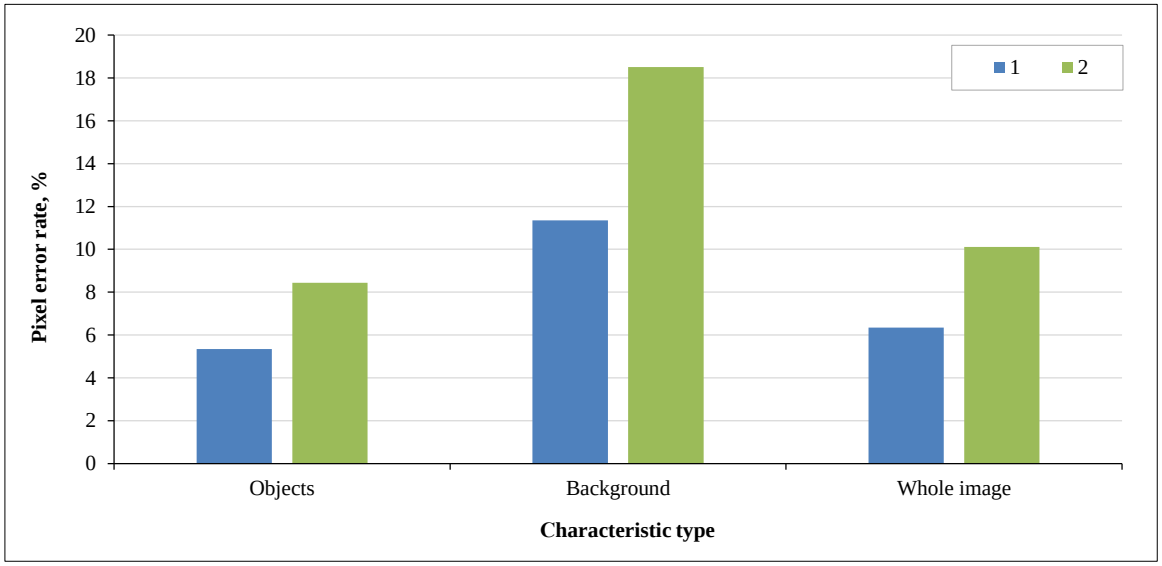

Figure 9. Comparison of characteristics: 1- the proposed methodology and 2- the methodology proposed in Wang et al. [26] based on the test set

The proposed methodology corresponds to the Pareto optimality criterion based on two criteria. Figure 10 shows the results of a comparison of different techniques for aggregating the energy tensor based on the test images. The proposed method is presented in the lowest part of Figure 10.

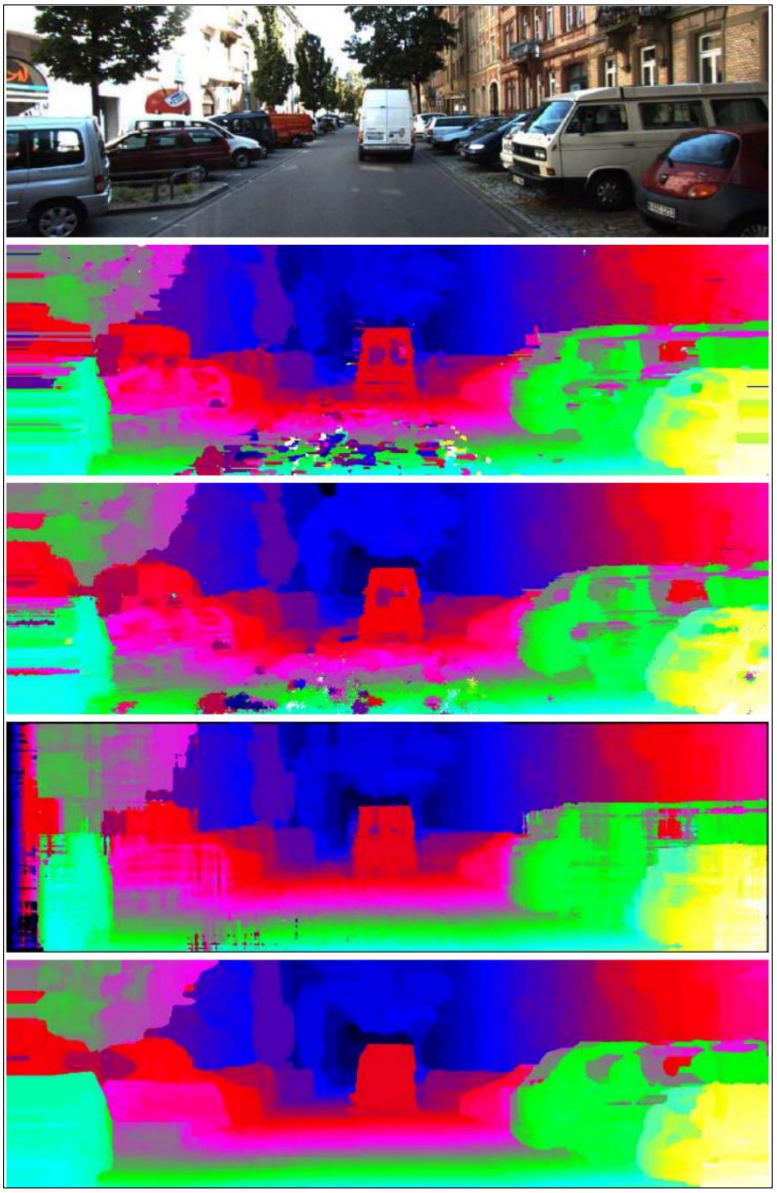

Figure 10. Results of comparison of various methods of energy tensor aggregation

The energy aggregation method is based on a recursive filter that considers the boundaries of objects in the image. In this case, the boundaries relevant to the task are predicted using a convolutional neural network. Thus, the proposed convolution-recursive model takes a pair of images as input and computes the displacement field without the need for post-processing. Unlike methods proposed in the literature, this model does not need to compare a large number of deep descriptors of high dimensionality, which significantly reduces the computational complexity and allows us to obtain an implementation that works in real time.

5. Conclusion

Most advanced methods imply the following stages of image matching: determination of the energy tensor, its aggregation, and subsequent optimization of the disparity field. Each of these stages can be used within a specific trained model. The disparity fields are calculated with low error through the comparative characterization of higher-dimensional descriptors, which are determined using modifications of a convolutional neural network with Siamese architecture. Simultaneously, the author confirmed that the use of trained aggregation of the energy tensor is characterized by significantly greater efficiency in terms of the computational resources used. High accuracy and efficiency are inherent in methods capable of assessing deviation fields based on advanced machine learning principles. At the same time, the simplicity of the training samples is the key factor in improving these methods.

The proposed technique is based on energy tensor aggregation, which allows the evaluation of the geometry of the scene. It should be understood that, at its core, the displacement field within the problem is one-dimensional; hence, the tensor itself will be three-dimensional, and therefore, it will occupy a relatively small storage space. A custom convolutional network is used to predict task-relevant edges. Thus, the proposed convolutional recurrent model uses a series of images obtained as input, after which the shift field is determined without post-processing. The main advantage of this model is that there is no need to compare a huge number of deep descriptors, which makes it possible to significantly reduce the overall computational complexity of the algorithm. In addition, it should be noted that the numerical experiments confirmed the importance of using data on specific edges of objects in the image. High accuracy and efficiency are inherent in methods capable of assessing deviation fields based on the principles of advanced machine learning. Simultaneously, the simplicity of the training samples is the key factor in improving these methods. As a direction for future research, it seems most perspective to develop a method that can combine deep descriptor learning and the proposed energy tensor aggregation method.

6. Declarations

6.1. Author Contributions

Conceptualization, V.Zh.K.; methodology, V.Zh.K.; software, A.A.T.; validation, V.Zh.K., A.A.T., and A.A.U.; formal analysis, A.A.U.; investigation, V.Zh.K.; resources, A.A.T.; data curation, V.Zh.K.; writing—original draft preparation, A.A.T.; writing—review and editing, V.Zh.K., A.A.T., and A.A.U.; visualization, A.A.U.; supervision, V.Zh.K.; project administration, V.Zh.K.; funding acquisition, A.A.T. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The presented results were carried out at the Federal State Autonomous Scientific Institution Institute for Design-Technological Informatics RAS with the financial support under project No. 075-11-2022-029 dated 04/08/2022 between STREAM LABS LLC and the Ministry of Science and Higher Education of the Russian Federation.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Lebedev, G. S., Linskaya, E. Y., Terekhov, V. Y., & Tatarkanov, A. A. bievich. (2023). Monitoring and Quality Control of Telemedical Services via the Identification of Artifacts in Video Footage. International Journal of Intelligent Systems and Applications in Engineering, 11(2), 82–92.

[2] Kuklin, V., Alexandrov, I., Polezhaev, D., & Tatarkanov, A. (2023). Prospects for developing digital telecommunication complexes for storing and analyzing media data. Bulletin of Electrical Engineering and Informatics, 12(3), 1536–1549. doi:10.11591/eei.v12i3.4840.

[3] Fang, L., Zhao, J., Pan, Z., & Li, Y. (2023). TPP: Deep learning based threshold post-processing multi-focus image fusion method. Computers and Electrical Engineering, 110, 108736. doi:10.1016/j.compeleceng.2023.108736.

[4] Aldao, E., Fernández-Pardo, L., González-deSantos, L. M., & González-Jorge, H. (2023). Comparison of deep learning and analytic image processing methods for autonomous inspection of railway bolts and clips. Construction and Building Materials, 384, 131472. doi:10.1016/j.conbuildmat.2023.131472.

[5] Pandey, B., Kumar Pandey, D., Pratap Mishra, B., & Rhmann, W. (2022). A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. Journal of King Saud University-Computer and Information Sciences, 34(8), 5083–5099. doi:10.1016/j.jksuci.2021.01.007.

[6] Nam, W., & Jang, B. (2024). A survey on multimodal bidirectional machine learning translation of image and natural language processing. Expert Systems with Applications, 235, 121168. doi:10.1016/j.eswa.2023.121168.

[7] Ziafati Bagherzadeh, S. H., & Toosizadeh, S. (2022). Eye Tracking Algorithm Based on Multi Model Kalman Filter. HighTech and Innovation Journal, 3(1), 15–27. doi:10.28991/hij-2022-03-01-02.

[8] Ma, J., Jiang, X., Fan, A., Jiang, J., & Yan, J. (2021). Image Matching from Handcrafted to Deep Features: A Survey. International Journal of Computer Vision, 129(1), 23–79. doi:10.1007/s11263-020-01359-2.

[9] Ualiyeva, R. M., Kukusheva, A. N., Insebayeva, M. K., Akhmetov, K. K., Zhangazin, S. B., & Krykbayeva, M. S. (2022). Agrotechnological methods of plant feeders applying for spring wheat agrocenoses – North-Eastern Kazakhstan varieties. Journal of Water and Land Development, 55, 28–40. doi:10.24425/jwld.2022.142301.

[10] Zhang, Y., & Hou, X. (2023). Application of video image processing in sports action recognition based on particle swarm optimization algorithm. Preventive Medicine, 173, 107592. doi:10.1016/j.ypmed.2023.107592.

[11] Chen, Q., & Yao, J. (2023). Outliers rejection in similar image matching. Virtual Reality and Intelligent Hardware, 5(2), 171–187. doi:10.1016/j.vrih.2023.02.004.

[12] Alsakka, F., Assaf, S., El-Chami, I., & Al-Hussein, M. (2023). Computer vision applications in offsite construction. Automation in Construction, 154, 104980. doi:10.1016/j.autcon.2023.104980.

[13] Zimiao, Z., Hao, Z., Kai, X., Yanan, W., & Fumin, Z. (2022). A non-iterative calibration method for the extrinsic parameters of binocular stereo vision considering the line constraints. Measurement, 205, 112151. doi:10.1016/j.measurement.2022.112151.

[14] Liu, Y., Li, Y., Dai, L., Yang, C., Wei, L., Lai, T., & Chen, R. (2021). Robust feature matching via advanced neighborhood topology consensus. Neurocomputing, 421, 273–284. doi:10.1016/j.neucom.2020.09.047.

[15] Ma, J., Zhao, J., Jiang, J., Zhou, H., & Guo, X. (2019). Locality Preserving Matching. International Journal of Computer Vision, 127(5), 512–531. doi:10.1007/s11263-018-1117-z.

[16] Wang, X. F., & Ye, D. (2010). On nonparametric comparison of images and regression surfaces. Journal of Statistical Planning and Inference, 140(10), 2875–2884. doi:10.1016/j.jspi.2010.03.011.

[17] Wang, T., Zhang, J., Zhang, S., Zhang, X., & Wang, J. (2023). A combined computer vision and image processing method for surface coverage measurement of shot peen forming. Journal of Manufacturing Processes, 91, 137–148. doi:10.1016/j.jmapro.2023.02.035.

[18] Krishnaveni, S., Subramani, K., Sharmila, L., Sathiya, V., Maheswari, M., & Priyaadarshan, B. (2023). Enhancing human sight perceptions to optimize machine vision: Untangling object recognition using deep learning techniques. Measurement: Sensors, 28, 100853. doi:10.1016/j.measen.2023.100853.

[19] Ualiyeva, R. M., Kaverina, M. M., Ivanko, L. N., & Zhangazin, S. B. (2023). Assessment of Spring Wheat Varieties for Pest Resistance. OnLine Journal of Biological Sciences, 23(4), 489–503. doi:10.3844/ojbsci.2023.489.503.

[20] Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., & Westling, P. (2014). High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. Pattern Recognition, 31–42, Springer, Cham, Switzerland. doi:10.1007/978-3-319-11752-2_3.

[21] Bisson-Larrivée, A., & LeMoine, J. B. (2022). Photogrammetry and the impact of camera placement and angular intervals between images on model reconstruction. Digital Applications in Archaeology and Cultural Heritage, 26, 224. doi:10.1016/j.daach.2022.e00224.

[22] Yang, B., Ali, F., Zhou, B., Li, S., Yu, Y., Yang, T., Liu, X., Liang, Z., & Zhang, K. (2022). A novel approach of efficient 3D reconstruction for real scene using unmanned aerial vehicle oblique photogrammetry with five cameras. Computers and Electrical Engineering, 99, 107804. doi:10.1016/j.compeleceng.2022.107804.

[23] Pu, C., Yang, C., Pu, J., Tylecek, R., & Fisher, R. B. (2023). A multi-modal garden dataset and hybrid 3D dense reconstruction framework based on panoramic stereo images for a trimming robot. ISPRS Journal of Photogrammetry and Remote Sensing, 202, 262–286. doi:10.1016/j.isprsjprs.2023.06.006.

[24] Guan, J., Yang, X., Lee, V. C. S., Liu, W., Li, Y., Ding, L., & Hui, B. (2022). Full field-of-view pavement stereo reconstruction under dynamic traffic conditions: Incorporating height-adaptive vehicle detection and multi-view occlusion optimization. Automation in Construction, 144, 104615. doi:10.1016/j.autcon.2022.104615.

[25] Xu, Y., Liu, X., Qin, L., & Zhu, S.-C. (2017). Cross-View People Tracking by Scene-Centered Spatio-Temporal Parsing. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). doi:10.1609/aaai.v31i1.11190.

[26] Wang, J., Zhang, S., Wang, Y., & Zhu, Z. (2021). Learning efficient multi-task stereo matching network with richer feature information. Neurocomputing, 421, 151–160. doi:10.1016/j.neucom.2020.08.010.

[27] Zahiri-Azar, R., & Salcudean, S. E. (2006). Motion estimation in ultrasound images using time domain cross correlation with prior estimates. IEEE Transactions on Biomedical Engineering, 53(10), 1990–2000. doi:10.1109/TBME.2006.881780.

[28] Hoskins, P. R., & Svensson, W. (2012). Current state of ultrasound elastography. Ultrasound, 20(1), 3–4. doi:10.1258/ult.2012.012e02.

[29] Yang, Q. (2012). A non-local cost aggregation method for stereo matching. 2012 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2012.6247827.