A Technique to Generate Depth Maps from Real Scenes without Manual Calibration

Carlos W. Carvalho, Ricardo S. Casado, Marcio M. Fernandes, and Emerson C. Pedrino

***Abstract*—This paper proposes a technique for the generation of a disparity map from a real scene, captured by a stereo vision system. The underlying motivation for this work is to develop a system not requiring the use of a calibration pattern, which usually involves manual intervention. This is a well-desired feature to allow its use in real-life environments, e.g., helping people with severe visual impairment or blindness to navigate through open spaces. Experimental results showed that the developed technique has a level of effectiveness similar to the other two well-established techniques found in the literature, making it a promising alternative to be employed in situations where the calibration step becomes a burden to the user.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8653

***Index Terms*—stereo vision; disparity map; calibration; visual impairment; blindness.**

I. Introduction

Among various types of existing disabilities, one of the most common is *visual impairment*, defined as partial or total loss of sight. According to the World Health Organization (WHO), several million of people worldwide suffer from some form of severe visual impairment or blindness.

Advances in technology have played a significant role in the task of making daily life easier for people with physical disabilities. In particular, computer vision is a research area showing concrete results, with various works seeking to help in tasks such as pattern recognition and scene mapping, among others.

A subject of great research interest for scene mapping is the *analysis of disparity between elements in a scene*. Disparity can be defined as a measure quantifying how far from the observer is an element in a given scene. Early studies about it were able to quantify the disparity between elements in a scene using some specific objects added to it. A well-known method was proposed by Zhang in [1] and [2], which was characterized by the use of a known object (a chessboard) to help the calculation of the disparity between elements in the scene. Since then, this has been an area of intense research interest, including some recent developments such as [3], [4], [5], [6] and [7].

Considering that various research efforts aim to reproduce some aspects of the human vision, including scene mapping, it is natural to think that a method able to map the disparity between elements in a scene can be used to develop a system to help visually impaired people in daily tasks, such as mapping distances between someone and surrounding objects [8].

In this context, the main objective of the work presented in this paper was to develop a method to generate disparity maps from real scenes, with minimal human intervention, a good degree of effectiveness, and relatively low computational complexity. The technique was developed through analysis, combination, and adjustments among a set of existing algorithms [9], [10]. The underlying motivation for this work is to use the technique as part of embedded systems employed by aid devices for spatial navigation, targeting people with severe visual impairment or even total blindness.

II. Development of the Proposed Technique

Many existing stereo vision systems use a calibration pattern to perform the calibration process, based on the technique presented in [2]. Considering that one of the objectives of this work is to be employed in aid devices for visually impaired people, the need of a manually performed calibration step should be avoided.

For the generation of a disparity map without human intervention, the analysis of the geometry of the scene can be done by searching the elements present in it, seeking to obtain a fundamental matrix, leading to the calculation of the disparity between elements.

However, to make possible the calculation of such a matrix, it is necessary to know at least a certain number of corresponding pairs of points in the images that were captured by the stereo vision system [11]. To tackle this issue, the set of adopted algorithms searches for initial pairs of corresponding points in a scene and calculates the *fundamental matrix* for them, followed by *rectification*. These steps can replace the conventional calibration process since such a matrix is shown to be sufficient for the rectification process, which is followed by the disparity map generation.

During the development of this project, the use of a stereo-vision system was necessary to test the algorithms. So, the *Minoru3D* webcam, from *Promotion and Display Technology*, was adopted, as it is capable of capturing images in three dimensions by using two distinct lenses. All algorithms were tested using *MatLab* tools (R2014b, 64 bits), with support of its specific toolbox for stereo vision (*The MathWorks, Inc.*).

The next subsections show the algorithmic steps that constitute the proposed technique, which are summarized by the block diagram shown in Fig. 1.

Carlos W. Carvalho, Ricardo S. Casado, Marcio M. Fernandes, and Emerson C. Pedrino are with Federal University of São Carlos, São Carlos, Brazil (e-mails: carloswdecarvalho@outlook.com, rscasado@ifsp.edu.br, marcio@dc.ufscar.br and emerson@dc.ufscar.br).

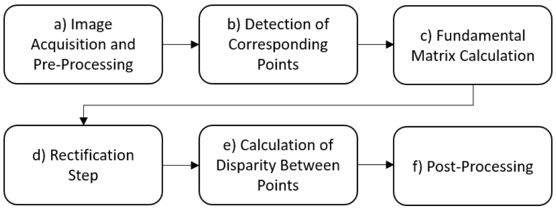

a) Image Acquisition and Pre-Processing
b) Detection of Corresponding Points
c) Fundamental Matrix Calculation
d) Rectification Step
e) Calculation of Disparity Between Points
f) Post-Processing

Fig. 1. Technique overview.

A. Image Acquisition and Pre-processing

The whole process starts by capturing a pair of images using the webcam, followed by a pre-processing step. Although issues such as brightness and image quality may influence the effectiveness of the method, due to the type of camera used, these issues rarely showed to be a serious problem in our experiments.

B. Detection of Initial Corresponding Points

Once the pre-processed images are obtained, the next step consists of obtaining the initial pairs of corresponding points from the scene elements. In this step, it is essential to maximize the number of obtained points, to increase the precision of the matrix generated by them, as the goal of the fundamental matrix is to approximate the relationship between points from a pair of images by a curve.

To estimate the fundamental matrix, random corresponding points need to be chosen from the images being processed. However, getting random points from an image and finding the match between them in another image can be highly complex. Thus, one way to find initial corresponding points is to focus on the *corners of the elements* from both images.

Many algorithms for the detection of corners can be found in the literature, so we have analyzed some of them to select the one that would produce the largest number of corners that could be matched. The algorithms considered in this work were BRISK [12], SURF [13], Harris [14], Minimum Eigeen Value (Min8Val) [15], FAST [16], [17] and MSER [18].

Each of those algorithms for corner detection was applied to the scenes, to find the number of points obtained. Then, a correspondence metric was applied to obtain the number of pairs of corresponding points from those corners that were found. The adopted algorithm to compute the correspondence metric in this analysis was the *Sum of Squared Differences* [19]. This is a metric that, starting from the image luminosity channel, uses a finite window around an analyzed point to define its disparity.

So, let $R(u, v)$ be a pixel of a reference window of length *len* and width *wid*, and $S(l, w)$ be a pixel of the search window. The difference of all the points of the defined window is calculated, and the square of all these differences is added, as shown by Equation 1:

$$\sum_{v=0}^{R_{len}} \sum_{u=0}^{R_{wid}} [R(u,v) - S(l+u, w+v)]^2 \qquad (1)$$

Once obtained the number of corners that can generate pairs, a *match rate* measure was developed, to analyse the ability that each algorithm for corner detection has to produce corresponding points. This rate is represented by T_c, as shown in Equation 2, where C is the number of obtained corresponding points, and I_l, I_r are the number of corners in the left and right images, respectively:

$$T_c = \frac{C}{\left\lceil \frac{l_i + l_r}{2} \right\rceil} .100 \qquad (2)$$

To select the most suitable algorithm, the match rate, which represents the percentage of corners that create correspondences, was calculated for all the algorithms for corner detection that were considered in this work, which was applied to 10 scenes S. The obtained results shown in Table I indicate that the *SURF algorithm* produces the largest match rate of corresponding points concerning the others. Thus, SURF has been adopted as the algorithm to find corners from stereo images in the developed technique described in this paper.

TABLE I
FINDING CORNERS: MATCH RATE FOR EACH ALGORITHM (%)

S	BRISK	SURF	Harris	Min8Val	FAST	MSER
1	4.5	**46.4**	19.9	12.2	14.3	20.2
2	7.0	**53.6**	16.5	8.9	13.8	29.4
3	4	**51.6**	10.1	3.3	0	9.1
4	4.1	**41.3**	7.1	9.6	8.7	15.4
5	6.5	**53**	14	6	13.2	25.1
6	0.6	**13**	5.3	2.2	7.9	4.2
7	5.6	**51.3**	18.8	9.6	11.8	30.7
8	0	**18.8**	4.5	13.6	0	2.7
9	0	**22.4**	3.8	5.9	0.7	2
10	3.2	**71.5**	17.4	27.5	7.9	6.1

C. Fundamental Matrix Calculation

Once the initial corresponding points using SURF are obtained, the next step is to calculate the fundamental matrix. As said before, the fundamental matrix is used as a linear estimation between coordinates of corresponding points in a pair of stereo images to map a point from a given image to its corresponding point in another image. Usually, the fundamental matrix is estimated using a *linear approximation algorithm* applied to a random set of elements chosen from the set of corresponding points. Then, the obtained matrix is submitted to a metric, which represents its efficiency rate for the intended use. Once the matrix efficiency is attested, it can be used as a possible fundamental matrix.

Different metrics to express how good a fundamental matrix is can be found in the scientific literature. Experimental analysis was performed on some of them to find the metric which can *best approximate the number of corresponding points*, while still maintaining the *highest number of correct correspondences*. The algorithms considered for this analysis were Least Median of Squares (LMedS) [20], Random Sample Consensus (RANSAC) [21], M-estimator Sample Consensus (MSAC) [22], and Least Trimmed of Squares (LTS) [20].

Experiments were carried out to obtain metrics for the fundamental matrices based on the corresponding points obtained using the SURF algorithm. The *real correspondence rate* T_s for each scene was calculated based on the expression, shown by the Equation 3, where C_{SURF} is the original number of corresponding points obtained by the SURF algorithm, and C_F is the number of real corresponding points obtained after the use of a fundamental matrix metric:

$$T_s = \frac{C_F}{C_{SURF}}.100 \tag{3}$$

As a result of this analysis, the data in Table II show the real correspondence rates obtained for each metric. It should be noticed that scene 8 was omitted from the results due to the fact this scene did not generate any correspondences able to be mapped by the fundamental matrix.

TABLE II
REAL CORRESPONDENCE RATE FOR SELECTED METRICS (%)

SCENE	LMedS	RANSAC	MSAC	LTS
1	**50.3**	25.1	27.4	49.7
2	**50.2**	28.4	24.4	49.8
3	50	**54.2**	**54.2**	50
4	**50.3**	29.1	28.5	49.6
5	**50.9**	20.7	20.4	49.8
6	**50**	15.3	19.2	**50**
7	**50.1**	24	26.6	49.8
8	0	**0**	**0**	**0**
9	**50.6**	29.6	23.5	49.4
10	**50.4**	27.7	23.5	49.6

Based on this analysis, it can be concluded that the most suitable metric for the proposed technique is the *Least Median of Squares (LMedS).* The analysis also showed that the MSAC metric can be effective using only 8 correspondences (against 16 for the others). So, when the number of corresponding points used to estimate a fundamental matrix is less than sixteen, the MSAC metric can be used instead of the LMedS metric.

Once the fundamental matrix is estimated, any point from an image can be mapped to a point in the other image. Using this information, it is possible to execute a process of *rectification* for the obtained pair of images, by determining the coordinates of the points that need to stay at the same height (*y* coordinate) in both images. By doing so, *it is no longer necessary to perform a manual calibration step* to proceed with the mapping of the scene elements.

D. Rectification Step

Once the fundamental matrix was obtained from the geometry of the analyzed scene, the next step consisted of the rectification of the obtained pair of images. The goal of the rectification process is to guarantee that each pair of corresponding points has the same height (i.e., the *y* coordinate) in both images.

Usually, a rectification process uses the fundamental matrix to compute two new matrices (one for each image), which correspond to the transformations that the planes of each image should undergo. Then, such matrices are converted employing projective transformations that are applied to the pair of images, and so achieving the required rectification [23]. The applied rectification adopted by our method does not use parameters related to calibration. Instead, the fundamental matrix originally obtained from the analyzed scene produces a flat rectification of the original images [11].

Considering that the fundamental matrix is capable of mapping any pair of corresponding points from the images, it is possible to estimate the dislocation degree that should be applied to each plane of the captured image to ensure that corresponding points possess the same height when the images are projected on the common plane created by the rectification process.

To summarize, the transformations that each image plane should undergo are obtained using the following steps :

- The left camera of the system is rotated (using the fundamental matrix) so that the epipoles of its image go to infinity along the *x* axis;
- The same previous rotation is applied to the right camera to recover the scene geometry;
- The right camera is rotated, this time based on the rotation matrix;
- Finally, the scale of both images is adjusted; each homography between the original positions and the new positions of each camera is a projective transformation.

Since planar rectification is a slightly easier process than cylindrical rectification, this is the rectification type adopted by this work. As an example, Fig. 2 shows the result of the rectification process, applied to a pair of images obtained by the stereo vision system.

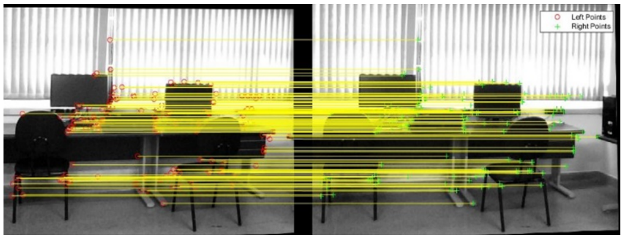


Fig. 2. Pair of stereo images after the flat rectification process, with yellow lines representing correspondences between corners.

Finally, as the fundamental matrix is highly dependent on the scene geometry, it may be possible in some cases that the projective transformations calculated during rectification are so distinct that they are capable of greatly distorting the plane of the pair of images. Usually, this condition occurs when epipolar lines cross themselves in one of the image planes – more specifically when the image plane intersects the baseline of the scene geometry. This potential issue is tackled in Section III-B.

E. Calculation of Disparity between Points

Once the rectified pair of images is obtained, the next step consists in calculating the disparity between them. In this work, we have adopted a semi-global method to analyze the correspondence between points, which works like an adapted local method.

A local method is generally based on a measurement of contrast obtained with the use of the *Sobel* filter [24], followed by the application of the *Sum of Absolute Differences*. By doing so, it is possible to compare blocks of points (pixels) present in the images, obtaining a disparity measure between them. The Sum of Absolute Differences is similar to the method of the *Sum of Squared Differences* [19].

So, let $R(u,v)$ be a pixel of a reference window, which possesses length *len* and width *wid*. Also, let $S(l,w)$ be a pixel of the search window. The Sum of Absolute Differences is represented by Equation 4:

$$\sum_{v=0}^{R_{len}} \sum_{u=0}^{R_{wid}} |R(u,v) - S(l+u, w+v)| \tag{4}$$

The difference between the local and the semi-global methods is that, once the disparity for all the windows is calculated, the local method ends. On the other hand, in the semi-global method, the disparity for each window is adjusted to force similarity between the neighboring windows. By doing so, the semi-global method guarantees the consistency of the disparity calculated for the whole map [25].

F. Post-processing

Following the disparity map generation, a step of post-processing is executed to improve the quality of results. A detailed analysis of the disparity map initially obtained showed that the process of rectification previously applied produced areas not belonging to the original scene. This occurs because the process of rectification distorts the planes of the original images, creating an area of black pixels around each rectified image (as can be seen in Fig. 2), which constitutes data of no interest.

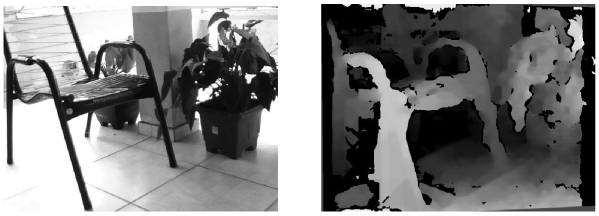
In order to fix this issue and achieve a representation as close as possible to the original scene, it is possible to apply the original transformations used in the rectification, creating a new transformation, and applying it to the disparity map.

So, starting from the projective transformations produced in the rectification step, a new transformation is calculated in two steps: first, the arithmetic mean of the two original projective transformations (used in the rectification step) is calculated. Then, this result is inverted, obtaining a new transformation, which is applied to the original disparity map.

Mathematically, being M_L and M_R the correspondence matrices of the projective transformations from the right and left images, respectively, the new transformation M_C is given by Equation 5:

$$M_c = \left(\frac{M_E + M_D}{2} \right)^{-1} \tag{5}$$

After applying the new projective transformation to the disparity map, the new map has a closer perspective to the original images. However, when the new transformation is applied, an unknown grayscale block appears on one side of the map. Since this block is not a result of the calculation of the disparity, it should be removed. Therefore, the vertical blocks containing only unknown shades of grey are removed, generating the final disparity map calculated for the original scene image. An example of the output obtained by the technique is shown in Figs. 3a and 3b.



(a) (b)

Fig. 3. An original scene and the corresponding disparity map, obtained by the technique. Original scene (a) Disparity map (b).

III. Technique Adjustments

Once the development of the first version of the system to generate disparity maps from a given scene (Section II) was finished, some initial experimental results could be obtained. Thus, the effectiveness of the process could be questioned, and possible adjustments applied to the system, seeking to guarantee the required performance for its intended use, as presented in the next subsections.

A. Interference of External Agents in the Map

Since the developed method is intended to be used in real (not controlled) environments, it is not uncommon for external agents to create inconsistencies in the obtained disparity map. To better understand these inconsistencies, we sought to understand how each external agent can potentially affect the disparity map and tried to find alternatives to avoid these situations. By doing so, it was possible to highlight the following issues:

- Excess or lack of *luminosity* in the scene can affect the generation of a disparity map, as it becomes harder to find corresponding corners in a pair of images.
- *Noise pixels* in the original image can be mapped as a corner by the SURF algorithm, possibly affecting the precision of the calculation of corresponding points.
- *Pattern repetitions* in a scene may result in areas with high disparity being mapped into areas with low disparity, and vice-versa.
- Scene elements *overlapped* by other elements during image capturing may not be mapped in the disparity map, since it may happen that they only appear in one image of the stereo pair.
- Image *reflections* or *transparencies* may induce the generation of disparity maps showing objects that actually do not exist in the original scene.

Problems due to luminosity, pattern repetitions, and reflections/transparencies can be addressed by *generating the disparity map more than once*, or even recapturing the scene image, forcing the corners to be recalculated.

As for problems due to elements overlapping, it can be argued that the mapping of those in the foreground is enough to meet the system requirements. Finally, applying digital noise filters during the pre-processing step can significantly reduce the observed some adverse effects.

B. Possible Problems to Generate the Disparity Map

During the development process, we came across two situations where the disparity map could not be generated:

1) When the number of corresponding points is *less than eight*, it is impossible to generate the fundamental matrix, even using the MSAC metric.
2) When *epipolar lines cross each other* across the image planes, it implies a high distortion of the respective original image. This way, it is impossible to guarantee that corresponding points will have the same y coordinate after rectification.

In general, when one of these problems occurs, it is initially infeasible to generate the disparity map. However, some factors can be observed in those cases:

In the first one, the low number of corresponding points is generally caused by the low quality of the captured images of the scene. Consequently, this condition results in the SURF algorithm detecting a low number of corners for each image. This problem can be solved by recapturing the images and restarting the generation process of the disparity map.

In the second case, the epipolar lines appear due to the way the fundamental matrix is calculated, probably using incorrect correspondences. Since the nature of the fundamental matrix calculation is *random* (as it is based on a random choice of corresponding points), the process does not need to be restarted when this issue occurs. It is just necessary to recalculate the fundamental matrix until a suitable one is generated (i.e., one that does not generate epipolar lines). This strategy has been successfully adopted by the system presented in this paper.

C. Reducing Distortions in the Disparity Map

Also due to the random nature of the calculation of the fundamental matrix, it is possible that the generated disparity map is not suitable for a proper stereo analysis, even after applying the aforementioned corrections. This problem may arise due to the adopted process to generate the fundamental matrix. In this context, the rectification step might produce a pair of projective transformations that can cause a great distortion in the original images of the stereo vision system, producing areas of unknown shades of gray.

To overcome this issue, a disparity map is considered valid if the existing distortion is minimal. So, a simple algorithm was created, which calculates the area percentage not belonging to the true disparity map. If this area is less than a threshold, the map is considered valid. Otherwise, it is discarded, and a new fundamental matrix is calculated, with the whole process being restarted from the rectification step onwards. This procedure is repeated until a valid disparity map is found. Some experimental analyses have shown that, in general, a valid disparity map contains less than 5% of the non-disparity area.

IV. Validation of the Proposed Technique

A. Analysis of Effectiveness

To verify the effectiveness of the proposed technique, the disparity maps obtained using it were compared with disparity maps produced by a well-established method.

A possible way to conduct this analysis is by using a *similarity metric*. A metric frequently used in stereo vision works is based on the calculation of the *mean error* between two maps, or between a given map and a reference disparity map (called *groundtruth*). The mean error is given by the mean of the values obtained from an error map E, which consists of a bi-dimensional matrix with the same dimensions as the disparity maps under analysis. So, the mean error of each point (x, y) of E can be calculated by Equation 6, where $d(x, y)$ is the adopted comparative metric, and t is a threshold reflecting the highest acceptable difference between points of equal coordinates present in both, the disparity map M and the ground-truth map G.:

$$E(x,y) = \begin{cases} 0 & if\ d(x,y) < t \\ 1 & otherwise \end{cases} \tag{6}$$

The comparative metric $d(x, y)$ can be calculated as the absolute differences between the intensities of points in position (x, y) of each map, as shown in Equation 7:

$$d(x,y) = |M(x,y) - G(x,y)| \tag{7}$$

Alternatively, it can be calculated as the square differences of those points, as shown in Equation 8:

$$d(x,y) = [M(x,y) - G(x,y)]^2 \tag{8}$$

For both of those metrics, the closer to zero the value $d(x, y)$ is, the better the map is. For validation purposes, we have compared results obtained using the proposed methodology against two other methods.

The first one is based on results quoted by the *The Middlebury Stereo Datasets*, which usually takes the absolute difference as a metric. The technique employed by that work uses the semi-global algorithm [25] to generate a disparity map. The efficiency of the disparity maps produced by it is already present in the Middlebury Datasets, with the mean error value (using absolute differences) already computed and standing at around **25%** [26]. That implies in a success rate of around 75%.

The second comparison was made with *Zhang's technique* [2], as it is considered canonical for works with stereo vision. Thus, the data presented in Table III show the similarity between maps obtained using Zhang's methodology and the technique proposed in this paper, with figures for both, the absolute and squared differences. The value of ten grey tones was the adopted threshold between points.

According to those figures, the obtained mean value stands around 71% and 75%, depending on the adopted metric for calculation. Those values can be considered very close to the results quoted by the Middlebury Datasets [26]. Furthermore, the disparity maps obtained by the proposed technique were shown to be similar to the ones obtained using Zhang's method.

TABLE III
SIMILARITY RATE BETWEEN THE PROPOSED TECHNIQUE AND ZHANG'S METHOD [2] (%)

SCENE	**Similarity Rate based on Absolute Differences**	**Similarity Rate based on Squared Differences**
1	72.83	70.86
2	80.45	75.28
3	75.44	71.76
4	83.80	82.51
5	76.07	74.23
6	72.87	69.03
7	71.45	67.91
8	65.10	60.10
9	71.54	67.73
10	77.60	74.23
$\bar{x}$	**74.71**	**71.36**
σ	**5.23**	**5.91**

B. Analysis of Efficiency

To evaluate the efficiency of the developed technique in terms of processing time to generate a disparity map from a scene, some experiments were carried out. Thus, a given image set was selected from the *Middlebury Stereo Datasets* repository from [27] and [28], and it was verified the required time to create a disparity map for each image of size 640x480 pixels.

The hardware setting used in the experiments consisted of a standard personal computer running the Win10 operating system, configured with an Intel® Core™ i5 processor (1.70GHz), and 4 GB RAM, without using GPU processing. The technique under evaluation is implemented using Matlab Tools, as already mentioned in Section II.

For each image, the data presented in Table IV show the *processing time* required to generate a disparity map. Five runs were performed for each image, resulting in the calculation of the corresponding mean ($\bar{x}$) and standard deviation (σ). All results are quoted in seconds.

TABLE IV
TIME SPENT TO GENERATE A DISPARITY MAP (SECONDS)

	Time in Seconds for tests 1-5, $\bar{x}$, σ						
SCENE	**1**	**2**	**3**	**4**	**5**	$\bar{x}$	σ
1	0.27	0.28	0.27	0.27	0.27	**0.27**	**0.00**
2	0.31	0.29	0.30	0.31	0.32	**0.31**	**0.01**
3	0.29	0.29	0.29	0.28	0.28	**0.29**	**0.01**
4	0.28	0.30	0.27	0.28	0.29	**0.29**	**0.01**
5	0.28	0.28	0.28	0.28	0.28	**0.28**	**0.00**
6	0.30	0.29	0.28	0.30	0.28	**0.29**	**0.01**
7	0.27	0.27	0.28	0.27	0.27	**0.27**	**0.01**
8	0.28	0.28	0.27	0.27	0.28	**0.28**	**0.00**
9	0.26	0.28	0.26	0.26	0.26	**0.26**	**0.01**
10	0.27	0.27	0.28	0.28	0.26	**0.27**	**0.01**

Those figures show that, for the image scenes considered, the time spent to generate a disparity map is *less than half a second*, using a hardware platform that can be considered modest at the time of writing this paper. It should be noted that, for those experiments, it was never necessary to recreate the maps due to the aforementioned issues.

V. CONCLUSIONS

The objective of this work was to analyze, combine, and adjust state-of-art algorithms to create a technique to generate a disparity map from a real scene, representing the depth of the elements in it, without manual calibration. The developed technique was shown to be effective in generating disparity maps employing successive processing steps, composed by stereo image acquisition, detection of corresponding points, fundamental matrix calculation, rectification step, and calculation of disparity between points, finalized by post-processing operations.

The analysis of experimental results using the proposed technique showed that it can produce maps with a similarity rate of around 75% when compared with two other well-established methods. In terms of processing efficiency, the proposed technique was able to produce maps in less than half a second, for a given set of images and using a standard computing platform. It should be observed that, if necessary, the algorithm to calculate disparity can be replaced by another one (possibly more effective), with minimal changes among other modules of the whole system.

Finally, as a result of the underlying motivation for this work, it should be emphasized that the proposed technique has the advantage of *not requiring the use of a calibration pattern*. That is an important feature for a system intended to be used as part of the design of aid devices for people with severe visual impairment or blindness.

REFERENCES

[1] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artificial intelligence*, vol. 78, no. 1-2, pp. 87–119, 1995. https://doi.org/10.1016/0004-3702(95)00022-4.

[2] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000. https://doi.org/10.1109/34.888718.

[3] S. Trejo, K. Martinez, and G. Flores, "Depth map estimation methodology for detecting free-obstacle navigation areas," in *2019 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 916–922, IEEE, 2019. https://doi.org/10.1109/ICUAS.2019.8798272.

[4] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8645–8654, 2022.

[5] M. Cui, Y. Zhu, Y. Liu, Y. Liu, G. Chen, and K. Huang, "Dense depth-map estimation based on fusion of event camera and sparse lidar," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022. https://doi.org/10.1109/TIM.2022.3144229.

[6] M. Beshley, P. Volodymyr, H. Beshley, and M. Gregus Jr, "A smartphone-based computer vision assistance system with neural network depth estimation for the visually impaired," in *International Conference on Artificial Intelligence and Soft Computing*, pp. 26–36, Springer, 2023. DOI: https://doi.org/10.1007/978-3-031-42508-0_3.

[7] X. Gui and X. Zhang, "An efficient dense depth map estimation algorithm using direct stereo matching for ultra-wide-angle images," in *Computer Graphics International Conference, CGI 2022, Virtual Event, September 12–16, 2022, Proceedings*, pp. 117–128, Springer, 2023. https://doi.org/10.1007/978-3-031-23473-6_10.

[8] B. Sae-jia, R. L. Paderon, and T. Srimuninnimit, "A head-mounted assistive device for visually impaired people with warning system from object detection and depth estimation," in *Journal of Physics: Conference Series*, vol. 2550, p. 012034, IOP Publishing, 2023. https://doi.org/10.1088/1742-6596/2550/1/012034.

[9] C. W. de Carvalho, "Uma metodologia automática para geração de mapas de disparidades de ambientes reais," Master's Thesis (in portuguese), Universidade Federal de São Carlos, Brazil, 2017. https://repositorio.ufscar.br/handle/ufscar/9692.

[10] M. M. Valipoor and A. De Antonio, "Recent trends in computer vision-driven scene understanding for vi/blind users: a systematic mapping," *Universal Access in the Information Society*, vol. 22, no. 3, pp. 983–1005, 2023. https://doi.org/10.1007/s10209-022-00868-w.

[11] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*, pp. 2548–2555, Ieee, 2011. https://doi.org/10.1109/ICCV.2011.6126542.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, pp. 404–417, Springer, 2006. https://doi.org/10.1007/11744023_32.

[14] C. Harris, M. Stephens, *et al.*, "A combined corner and edge detector," in *Processding of the 4th Alvey vision conference*, pp. 147–151, 1988.

[15] J. Shi *et al.*, "Good features to track," in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pp. 593–600, IEEE, 1994. https://doi.org/10.1109/CVPR.1994.323794.

[16] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, pp. 1508–1515, Ieee, 2005. https://doi.org/10.1109/ICCV.2005.104.

[17] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*, pp. 430–443, Springer, 2006. https://doi.org/10.1007/11744023_34.

[18] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (mser) tracking," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 1, pp. 553–560, Ieee, 2006. https://doi.org/10.1109/CVPR.2006.107.

[19] N. Roma, J. Santos-Victor, and J. Tomé, "A comparative analysis of cross-correlation matching algorithms using a pyramidal resolution approach," in *Empirical Evaluation Methods in Computer Vision*, pp. 117–142, World Scientific, 2002. https://doi.org/10.1142/9789812777423_0006.

[20] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984. https://doi.org/10.1080/01621459.1984.10477105.

[21] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981. https://doi.org/10.1145/358669.358692.

[22] P. H. Torr and D. W. Murray, "The development and comparison of robust methods for estimating the fundamental matrix," *International journal of computer vision*, vol. 24, no. 3, pp. 271–300, 1997. https://doi.org/10.1023/A:1007927408552.

[23] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Machine vision and applications*, vol. 12, no. 1, pp. 16–22, 2000. https://doi.org/10.1007/s001380050120.

[24] R. C. Gonzalez and R. E. Woods, "Image processing," *Digital image processing*, vol. 2, no. 1, 2007.

[25] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 807–814, IEEE, 2005. https://doi.org/10.1109/CVPR.2005.56.

[26] M. College, "Middlebury stereo evaluation." Available in: https://vision.middlebury.edu/stereo/eval3/, 2016. Access date: Jan 2023.

[27] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002. https://doi.org/10.1023/A:1014573219977.

[28] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*, pp. 31–42, Springer, 2014. https://doi.org/10.1007/978-3-319-11752-2_3.

Carlos W. de Carvalho holds a Bachelor's degree in Computer Science from Universidade Estadual Paulista - UNESP (2013), and a Master's degree in Computer Science from Federal University of São Carlos - UFSCar (2017), with emphasis on Digital Image Processing, Stereo Vision, and Computer Graphics. (e-mail: carloswilldecarvalho@outlook.com)

Ricardo Salvino Casado obtained his master's degree from the University of São Paulo, São Carlos campus (EESC/USP). He is an RDE professor at the Federal Institute of São Paulo. Currently, he is pursuing a Ph.D. in Computer Science at the Federal University of São Carlos. His areas of interest include computer vision, deep learning, monodepth estimation, and genetic programming. (e-mail: rscasado@ifsp.edu.br).

Marcio M. Fernandes has a full degree in Computer Science: University of São Paulo (undergraduate, 1989), Federal University of São Carlos (master's, 1993), and The University of Edinburgh, UK (PhD, 1999). Has been working as a professor for undergraduate and postgraduate courses at UFSCar since 2008, participating in research projects that have resulted in the publication of several scientific articles. (e-mail: marcio@dc.ufscar.br)

Emerson Carlos Pedrino is an Associate Professor (Ph.D.) in the Department of Computer Science at the Federal University of São Carlos. He holds a degree in Electrical Engineering from the University of São Paulo and a Bachelor's degree in Computational Physics, both from the University of São Paulo - EESC (2016) and IFSC (2000). He earned a Master's degree in Electrical Engineering from the University of São Paulo - EESC (2003) and a Ph.D. in Electrical Engineering from the University of São Paulo - EESC (2008). Additionally, he completed a Post-doctorate in Electronic Engineering (as a Visiting Professor) at the Department of Electronic Engineering, University of York, England, with research funding provided by FAPESP (2018-2019). He continues to collaborate in the development of hardware applications involving intelligent systems. (e-mail: emerson@dc.ufscar.br).