A New Stereo Matching Function by a Hybrid Convolutional Neural Network

Nazia Amreen¹*, Tahera Abid² ,Zahoora Abid³

¹,²Department of Information Technology, Nawab Shah Alam Khan College of Engineering & Technology, Hyderabad, Telangana, India.
³Department of Computer Science & Engineering, Nawab Shah Alam Khan College of Engineering & Technology, Hyderabad, Telangana, India.

Received: 14.07.2023 Accepted: 22.07.2023 Published Online: 31.07.2023

Abstract: This research introduces a novel method for creating stereo matching algorithms, combining elements of learned and custom methods. The purpose of this strategy is to provide more reliable outcomes than conventional approaches. The primary goal of the technique for stereo corresponding remains to harvest a difference map. Three-dimensional (3D) reconstruction is only one of many uses for this map. Convolutional neural network (CNN) produced raw disparity maps may still contain mistakes in low-resolution textures. The programme will utilise a hybrid CNN-based approach and a faster directional intensity computation to enhance the accuracy of the matching cost calculation phase. Use of the modified truncated value of directional intensity has the potential to greatly reduce radiometric errors. To further enhance the accuracy attained by the approach, a cost aggregation phase is used in which the bilateral filter (BF) is applied to the raw matching cost. Building a gap map with the whole cost of costs is the first step in the WTA optimisation process. The final disparity map is an improvement above the basic map thanks to several rounds of refinement. By means of the Middlebury Online Stereophonic Benchmarking Scheme, the article verified the algorithm efficacy.

Key words: Stereo vision, Convolutional Neural Network, Directional intensity, Stereo matching algorithm.

*Correspondence: Assistant Professor, Department of Information Technology, Nawab Shah Alam Khan College of Engineering & Technology, Hyderabad, Telangana, India. Email:naziaamreen594@gmail.com
https://doi.org/10.58599/IJSMIEN.2023.1704

This work is licensed under a Creative Commons Attribution 4.0 International License CC BY-NC-ND 4.0.
1. INTRODUCTION

The sum of major developments in stereo idea in recent years has increased. In the subject of stereo vision, the issue of stereo matching seems to be one of the most highly disputed problems at the time [1]. The chart recalls the dissimilar pixel values between the left and right pictures for features that are spatially similar. The distance amongst the camera and a certain item may be calculated using a map. Numerous applications exist for this information, such as navigation, obstacle avoidance, robotics, and 3D picture reconstruction. The 3D face picture utilised in facial recognition was created with the help of stereo vision technology [2].

Both the stereo-based approach and the single opinion before monocular-based methodology may be used to get depth information, as shown in [3]'s analysis of the literature. With this method, a single image is all that's needed to create a disparity map. The stereo-based technique, however, has been demonstrated to provide more precise findings than the monocular depth method [3, 4]. The standard single-view strategy lacks several crucial pieces of information that the multi-view technique provides. One advantage of the stereo-based technique is that it may provide better results than the monocular approach in terms of object retrieval, as stated by. The stereotypical method has this benefit. LiDAR, which stands for "light detection and ranging," is another technique for gathering depth information. LiDAR is a method that can provide precise coordinates in three dimensions [4]. In contrast, this approach is not very efficient in terms of either money or time [5].

The main operations of the stereo idea algorithm are shown in Figure 1. The steps of the algorithm have been given a more formal structure in light of the prior research [6, 7]. Finding out how much a match costs is the first order of business [8]. Several studies use a method that is perpendicular to the direction of the intensity difference when calculating the price of a match [9]. The cheap computational complexity of absolute and squared difference calculations also makes them popular. Next, we have cost aggregation, which may reduce the amount of mismatches that happened in the first place. Several filters are able to carry out the job because they preserve the edges. These filters soften the input while keeping some of its useful sharpness. Therefore, its aggregation phase output is superior than that of low pass filters (such as the Gaussian and box filters). The final step is to give the disparity map a numerical value. The winner-take-all (WTA) optimisation strategy is often utilised at this step of the local procedure. When making the first map showing the discrepancy, WTA chooses the minimum cost value difference for each pixel point [10]. Inaccuracies in the initial disparity computation might occur during the third phase due to occlusions, poor texturing, or improper matching [11]. Therefore, in the end stage, it may be necessary to use a number of post-processing techniques to make any necessary tweaks to the chart. Localization is another prominent use [12] of the median filter. This non-linear filter is another variation on the theme of eliminating unnecessary hiss. Due to the intricacy of this last phase, the whole procedure might end up taking longer than expected.
The number of major developments in stereo vision in recent years has increased. One of the most passionately contested topics in the field of stereo vision at the present [1] is the question of stereo matching. When comparing characteristics between the left also right photos, the chart stores the pixel value differences between the two images. Using a map, one may pinpoint the precise location of a camera in relation to an item of interest. This data may be used for several purposes, including as navigation, obstacle avoidance, robotics, and the reconstruction of three-dimensional images. The three-dimensional face picture utilised in facial recognition was created using stereo vision[2].

Reviewing the relevant literature reveals that [3] finds that depth information may be obtained using either the stereo-based approach or the single opinion or monocular-based methodology. With this technique, you just need a single image to crop a disparity chart. In contrast to the monocular depth approach, the stereo-based methodology has been shown to provide more reliable results [3, 4]. In comparison to the multi-view strategy, the traditional single-view method fails to provide many crucial pieces of data. In terms of object retrieval, utilising a technique based on stereo vision may be preferable than using an approach that just employs one eye, as stated in. The commonly used method has this benefit.

Light detection and ranging, or "LiDAR" for short, is another technique for measuring depth. Laser-based light detection and ranging, or LiDAR for short, is one technique used to measure depth. The method known as LiDAR can reliably generate three-dimensional coordinates [4]. However, this approach is not efficient in terms of time savings or cost savings [5]. Because of the limitations of the monocular also LiDAR approaches, we decided towards create our own algorithm and refine the stereo-based method.

Figure 1 depicts the main operations of the stereo idea algorithm. The different steps of the method have been given a more formal structure in light of the prior literature [6]. Establishing the price of a match is the first order of business. Some studies use a method that is perpendicular to the direction in which the intensity difference runs when calculating the price of a match [10]. Low processing complexity is another factor in the widespread usage of absolute and squared difference calculations.
The following process, termed cost aggregation, may reduce the amount of inconsistencies that existed to begin with [11]. Several filters are able to do this goal because they preserve the edges, such as the bilateral filter also the guided filter. These filters reduce the input’s overall sharpness while keeping its utility intact. That’s why its aggregation phase produces better results than typical low pass filters like the box and Gaussian filters. Last but not least, the disparity map has to be assigned a numerical value [12, 13]. The winner-take-all (WTA) optimisation strategy is often used at this step of the local process. WTA chooses the least costly cost value difference for each pixel point when it generates the first map representing the disparity. Inaccuracies in the first disparity computation might occur in the third phase due to occlusions, poor texturing, or poor matching. The last step of the process, making any necessary alterations to the map, may thus need the use of many post-processing techniques. Left-to-right consistency (or LRC for short) was used by the paper’s authors [14] to spot the misaligned pixels. When it comes to localisation, the median filter is often used. Another way that hiss may be filtered out of an audio stream using a non-linear filter. The intricacy of this last phase raises the possibility that the whole procedure may take much longer than expected.

2. THE PROPOSED METHOD

It should be emphasised that our proposed method is provided in the context of the numerous categories outlined in [6]. The big picture of our method is shown in Figure 2. Our projected algorithm comprises of the following 4 parts:

![Figure 2. Procedures of the proposed stereo vision algorithm](image)

Our CNN-based typical is built from the ground up using the MC-CNN-acrt framework. Though, as shown in [15], we were successful in refining the layout in a variety of ways. By doing this, we help push the field forward by uncovering fresh information. As we discussed in our prior work, we adopt a convergent classification strategy to calculate matching costs in our Siamese-based CNN perfect. Following the precedent we set, we kept the number of CNN layers at eight. The network’s similarity score may be used to quickly and easily determine whether a match is favourable or unfavourable. Each
disparity $d$ at respectively pixel location $p$ is assigned a cost value, which is considered in Equation 1.

$$CNNN(p, d) = -s(PL(p), PR(p - d))$$  \hspace{1cm} (1)

In the additional stage, we improve upon the raw corresponding costs to construct a additional precise disparity map. This remains since there is inherent variability in the first-stage cost volume. Given the novelty of the situation, we will utilise a bilateral filter (BF) to sum the costs across the board at this first stage. The BF is responsible for all of the procedures that ensure the matching expenses are kept to a minimum and their edges are kept sharp. In accordance with BF [11], the entire amount of raw matching expenditures shall be calculated.

The process of refining passes through a number of phases, each of which is distinct. In [15], an explanation is given on the use of the left-right consistency checkered and interpolation for the aim of correcting occlusion and mismatches. We will start by using LRC to locate any pixels that have issues so that we can fix them. After that, the process that is known as hole filling is used in order to replace the pixel that was discovered to be erroneous. Afterward that, we used a guided image filter because of the superior edge preservation that it provided [16, 17]. In the course of this inquiry, we make use of the GIF filter kernel that is covered in the previous section.

$$d_{GF}(p) = G_{p,q}(I)d(p)$$  \hspace{1cm} (2)

Wherever $I$ represents the guiding image also $p$ represents the picture element that is located at those particular $(x, y)$ coordinates. The location $q$ of another pixel may be used to determine which pixel is located closest to you in the wk support area. The standards of the standard aberration and the mean intensity are each represented by the corresponding symbols and. The smoothness term may have its value changed by adjusting the parameter. In the equation (8), the enhanced GIF disparity map is shown by the notation $d_{GF}$.

3. RESULTS AND DISCUSSION

The effectiveness of the suggested method was assessed on the PC platform using the v3 datasets. The fundamental encryption is built in Python also the Keras also Tensorflow libraries are used towards infer the structure of the CNN. A personal computer with a 3.0 GHz and an Nvidia GTX1060 graphics processing unit was used for this investigation. Our CNN model’s hyperparameters have not been modified from prior work. Test datasets consisting of photographs were made available via Middlebury’s online benchmarking system . We compare the proposed technique to others, including the pyramid stereo matching network and CNN variants and MC-CNN-fst. The errors that occurred as a result of incorrect disparity values in each pixel are listed in Table 1. In Table 2, we also
Table 1. Middlebury College Benchmarking Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adirondack</td>
<td>5.47</td>
<td>2.27</td>
<td>6.60</td>
<td>4.34</td>
<td>5.42</td>
<td>8.93</td>
<td>5.83</td>
<td>9.41</td>
</tr>
<tr>
<td>Piano</td>
<td>6.74</td>
<td>6.24</td>
<td>4.87</td>
<td>6.10</td>
<td>8.22</td>
<td>5.99</td>
<td>8.23</td>
<td>6.35</td>
</tr>
<tr>
<td>ArtL</td>
<td>8.46</td>
<td>19.20</td>
<td>9.43</td>
<td>19.80</td>
<td>18.30</td>
<td>14.91</td>
<td>21.60</td>
<td>5.80</td>
</tr>
<tr>
<td>Motorcycle</td>
<td>7.39</td>
<td>6.89</td>
<td>5.53</td>
<td>7.32</td>
<td>8.93</td>
<td>9.26</td>
<td>9.47</td>
<td>7.75</td>
</tr>
<tr>
<td>MotorcycleE</td>
<td>7.34</td>
<td>5.88</td>
<td>5.14</td>
<td>7.31</td>
<td>8.85</td>
<td>8.36</td>
<td>9.49</td>
<td>7.34</td>
</tr>
<tr>
<td>Jadeplant</td>
<td>38.70</td>
<td>23.71</td>
<td>57.81</td>
<td>35.11</td>
<td>33.61</td>
<td>69.50</td>
<td>37.31</td>
<td>65.51</td>
</tr>
</tbody>
</table>

see the NonOcc error%, which is a valuable indicator. This error manifests itself in non-occluded pixels displaying erroneous disparity values. In comparison to the innovative MC-CNN-acrt also other available approaches, Table 1 shows that the suggested algorithm framework may get better outcomes. This is true for all kinds of mistakes. Using the method outlined, it is conceivable that standard errors may be reduced by 9.29%. It outperforms other deep learning algorithms like SGBMP , JMR , and others that have been released in the last several years. The photos of the jadeplant, pianoL, vintage, and playtable highlight the significant discrepancy in the results of the two methods, which are otherwise equivalent.

Table 2 displays the results of a statistical analysis that compares the suggested technique to others based on the proportion of NonOcc errors produced by each method. Even the suggested technique outperformed The provided method has a much lower error rate of 7.06% compared to PSMNet’s output (9.70%). This means that, with the exception of the Vintage also Piano photos, our method outperforms PSMNet in each case.

The two methodologies are compared and contrasted in detail in Figure 3, which can be found here. A real-life example of a PlayTable is shown in the first part of Figure 3. There is an area on the floor of the PlayTable photographs that has a low-texture patch, and this can be seen in both the left and right images of Figure 3 and Figure 3(c). The suggested technique products a horizontal difference map with variable depths in the region of low texture also plain colour, as can be seen in Figure 3(d), which illustrates this result. The quality of the map produced by this technique is superior to that of the maps produced by other methods, as can be seen in Figures 3. Respectively. The findings of this study uncovered a method that accomplishes the primary objective of the research, which was to discover a technique to reduce the number of errors that occurred in the region with poor texture.
According to Table 1, the PlayTable image has a much lower level of All mistake as compared to the other methods.

![Figure 3](image)

**Figure 3.** Middlebury image assessment, (a) ground truth, (b) left, (c) right, (d) proposed method, (e) SGBMP, (f) JMR (g) MCWS, (h) MCacrt, (i) LSELAS

![Figure 4](image)

**Figure 4.** Middlebury- image (a) left image, (b) without projected method, and (c) projected method

The first step towards integrating CDI and CCNN into CIM is supported by evidence. By applying CIM to the whole scope of costs error rate may be lowered from 16.2% to 9.29% (for All faults) and from 14.4% to 6.05% (for NonOcc error). The All Error and NonOcc Error rates are represented by these percentages. Figure 4a shows the Adirondack scene’s left side, which may be examined qualitatively. Figure 4(b) displays the Adirondack image’s disparity map after being processed with CDI without CCNN. The overall impact of CIM expenses on budget is seen in Figure 4c. As a result,
the proposed matching cost calculation stages may take use of the improved precision afforded by this amalgamation.

4. CONCLUSION

Finally, we show how to compute the price of matching by combining a learning-based approach with more traditional methods. A hybrid approach is used to describe this strategy. Due to the potential for noise in the newly formed quantity of cost data, an additional step for aggregating costs was implemented. The input may be smoothed down yet the edge’s sharpness is preserved thanks to BF’s use as a ”edge-preserving filter.” The original disparity map was made using the WTA programme. Post-processing techniques such as LRC, hole substantial, GIF, and WMF may be used to create a final disparity map. The suggested method improves the disparity map’s correctness in the low texture area while preserving the thing edge, as seen by the final disparity map. In addition, when measured against other well-established algorithms.

References


