Fast and Robust Stereo Matching Algorithms for Mining Automation

Jasmine Banks\(^1\), Mohammed Bennamoun\(^1\) and Peter Corke\(^2,3\)

1: Space Centre for Satellite Navigation, Queensland University of Technology
2: Centre for Mining Technology and Equipment
3: CSIRO Manufacturing Science and Technology

The mining environment, being complex, irregular and time varying, presents a challenging prospect for stereo vision. For this application, speed, reliability, and the ability to produce a dense depth map are of foremost importance. This paper evaluates a number of matching techniques for possible use in a stereo vision sensor for mining automation applications. Area-based techniques have been investigated because they have the potential to yield dense maps, are amenable to fast hardware implementation, and are suited to textured scenes. In addition, two non-parametric transforms, namely, the rank and census, have been investigated. Matching algorithms using these transforms were found to have a number of clear advantages, including reliability in the presence of radiometric distortion, low computational complexity, and amenability to hardware implementation.

1 Introduction

Perception of the three-dimensional environment is a pre-requisite for mine equipment automation, since autonomous vehicles and robot devices need to be aware of the surrounding environment in order to plan their actions and carry out tasks. Stereo vision is a technique used to discern depth information from a scene, in which two (or more) images of a scene are taken from different perspectives, and depth is computed from stereo disparity. Applications for stereo vision include aerial photogrammetry\(^1\), autonomous vehicle guidance\(^3\), robotics and industrial automation. The mining environment, being complex, irregular and time varying, presents a challenging prospect for stereo vision. For this application, speed, reliability, and the ability to produce a dense depth map are of foremost importance\(^4\).

A fundamental issue is to establish correspondence or matching of points in two images, in order to compute the disparity and subsequently the 3-D world coordinates. A wide variety of solutions exist for the matching problem, as attested to in literature\(^5\)–\(^10\). This paper considers two types of matching algorithms: area-based and feature-based\(^11\). These matching techniques are distinguished by the type of spatial domain primitives they employ. In area-based techniques, neighbourhoods of grey-level pixel values are compared to locate the optimum match. Matching metrics\(^12\) are then used to mathematically compare two pixel regions. In contrast, feature-based techniques extract interesting features such as spots or edges, then compare the symbolic representations of these features to find the best match.

Feature-based matchers tend to be faster than area-based matchers, since they only use a small subset of pixels for matching. However, they only yield very sparse depth maps, since matching only takes place at locations where features occur. Depth information for intermediate points must be discerned from interpolation, which relies on assumptions about scene geometry between features. Feature-based matchers are also highly accurate since features can be located with sub-pixel precision. They are best suited to images where features are relatively sparse, such as scenes containing planar surfaces delineated by edges. Scenes of this nature would typically arise from a man-made environment. Area-based matchers are usually unsuitable for these images, since smooth surfaces lack sufficient pixel variation for an area-based matcher to establish a match.

Area-based techniques, on the other hand, are best suited to highly textured scenes, in contrast to feature-based techniques which tend to be confused by a large amount of surface texture\(^13\). Area-based matchers can also potentially yield matching results for every image pixel and hence yield a dense depth map. However, their accuracy is not as high as feature-based methods, due to the averaging effect introduced by using a neighbourhood of pixels for matching.

An extension of area-based techniques involves applying a local transform to the grey level images
prior to matching. Two non-parametric transforms which have been recently proposed for the stereo matching problem are the rank and the census[14]. The motivation behind their application to matching is twofold. Firstly, they result in improved reliability in the presence of radiometric distortion and small amounts of random noise. This has been supported by experimental work[15]. Secondly, they are amenable to fast hardware implementation [16, 17], making them suitable for real-time applications.

This paper focuses in particular on area-based techniques and non-parametric transforms, as they have the potential to yield dense depth maps, and are amenable to fast hardware implementation. Furthermore, scenes comprised of rocks contain a large amount of surface texture, and area-based techniques are well suited to textured scenes. Previous studies[12] have compared area-based metrics for a range of image types. This paper differs from this previous work in that in addition to area-based metrics, two non-parametric transforms, namely the rank and census, have been compared. In this study, these techniques have been applied to a particular scene domain, i.e., close range scenes of rocks, which would typically be encountered in the mining environment.

2 Area-Based Matching

In area-based matching, a point to be matched essentially becomes the centre of a small window of pixels, and this window is compared with similarly sized regions in the other image. Matching metrics are used to provide a numerical measure of the similarity between a template window in the first image and a candidate window in the second image, and hence are used to determine the optimum match.

Epipolar geometry[6] is used to improve the efficiency of the matching process by constraining the search to one dimension. Stereo images may be rectified such that the epipolar lines correspond to the horizontal scan lines[18]. A simple approach used in area based matching is to compute the value of the matching metric using a fixed window in the first image and a shifting window in the second image, as illustrated in Figure 1. The shifting window is moved in integer increments along the epipolar line, where the amount of shift is the test disparity. The disparity having the optimum value for the matching metric is then selected.

2.1 Matching Problems

All area-based matching algorithms must deal with at least the following problems:

- **Occlusions** caused by portions of a scene being visible in only one image.
- **Repetitive patterns** which can potentially result in invalid matches.
- **Bland regions** which do not contain enough information for matching, e.g., a featureless wall.
- **Perspective distortion** which occurs because the shape of objects will change when they are viewed from different vantage points.
- **Radiometric distortion** which may result in a constant offset between pixel values in the two images, and/or pixel intensities in one image being multiplied by a gain factor with respect to the other image. These effects are caused by differences in camera parameters, such as gain, bias and gamma factor.
- **Specular reflection** caused by the reflectance properties of the object. Matching algorithms usually assume Lambertian reflection model, in which an object reflects light equally in all directions. It is therefore assumed that a particular point will have the same intensity regardless of the direction from which it is viewed. However, this is often not the case, with specular (mirror-like) reflection being the most dramatic departure from the Lambertian case.
- **Noise** which is introduced by the image acquisition and digitisation process.

2.2 Matching Metrics

A number of classical matching metrics are listed in Table 1. All these metrics use a square window of pixels as the basis for comparison, and are based on the intuitive assumption that image regions should match well if they “look the same”.

![Figure 1: Epipolar constrained area based matching.](image-url)
Table 1: Area based matching measures[12]. In all cases, \( I_1 \) denotes the template window, \( I_2 \) is the candidate window, and \( \sum_{(u,v) \in W} \) indicates summation over the window.

| Sum of Absolute Differences | SAD | \( \sum_{(u,v) \in W} |I_1(u,v) - I_2(x+u,y+v)| \) |
|----------------------------|-----|------------------------------------------------------------------|
| Zero Mean Sum of Absolute Differences | ZSAD | \( \sum_{(u,v) \in W} \left( |I_1(u,v) - \bar{I}_1| - (I_2(x+u,y+v) - \bar{I}_2) \right) \) |
| Sum of Squared Differences | SSD | \( \sum_{(u,v) \in W} (I_1(u,v) - I_2(x+u,y+v))^2 \) |
| Zero Mean Sum of Squared Differences | ZSSD | \( \sum_{(u,v) \in W} \left( (I_1(u,v) - \bar{I}_1) - (I_2(x+u,y+v) - \bar{I}_2) \right)^2 \) |
| Normalised Cross Correlation | NCC | \( \frac{\sum_{(u,v) \in W} I_1(u,v) \cdot I_2(x+u,y+v)}{\sqrt{\sum_{(u,v) \in W} I_1^2(u,v) \cdot \sum_{(u,v) \in W} I_2^2(x+u,y+v)}} \) |
| Zero Mean Normalised Cross Correlation | ZNCC | \( \frac{\sum_{(u,v) \in W} (I_1(u,v) - \bar{I}_1) \cdot (I_2(x+u,y+v) - \bar{I}_2)}{\sqrt{\sum_{(u,v) \in W} (I_1(u,v) - \bar{I}_1)^2 \cdot \sum_{(u,v) \in W} (I_2(x+u,y+v) - \bar{I}_2)^2}} \) |

The SAD and the SSD are intuitively the simplest, and computationally the least expensive of all the matching measures[19], and are based on subtraction of pixel values. Two areas which consist of exactly the same pixel values would yield a score of zero. However, these metrics will no longer yield the correct results in the case of radiometric distortion. The ZSAD and the ZSSD have been devised to deal with this problem, by subtracting the mean of the match area from each intensity value. However, the improved performance of the ZSAD and ZSSD over the SAD and SSD is offset by substantially increased computational complexity.

The NCC measure deals with a possible gain factor by dividing by the variances of each window, while the ZNCC measure additionally deals with the offset problem by first subtracting the mean from each pixel value. For grey level images, these metrics will have a value ranging from -1 to 1, where 1 represents the best match.

2.3 Validation of Matches

Once the optimum match is selected using a matching metric, a number of simple validation techniques may be applied in order to identify incorrect matches.

One such technique is left-right consistency checking[11, 13], which involves reversing the roles of the two images and performing matching a second time, as illustrated by Figure 2. Firstly, epipolar constrained matching is carried out using a template window centred on \( I_1 \), and the point \( I_2 \), which is the best match for \( I_1 \), is found. Matching is then performed again, this time using a template window centred on \( I_2 \). If this match leads back to the original point \( I_1 \), then the match is consistent, otherwise, it is flagged as inconsistent. This validity test is likely to detect invalid matches which may result from bland areas, and also from occlusions. The pixels which comprise an occluded area are likely to match, more or less at random, with locations in the other image. However, these locations are unlikely to match back
to the pixels in the occlusion area, rather, they are more likely to match with their own corresponding points. This validation technique can be fooled by repetitive patterns.

The number of correct matches can be further increased by removing locally anomalous matches from the matches which remain after left-right consistency checking. This heuristic is based on the assumption that locally anomalous matches are more likely to be incorrect[11, 20].

3 Non-Parametric Techniques

Non-parametric techniques are based on the relative ordering of pixel intensities within a window, rather than the intensity values themselves. Consequently, these techniques are robust with respect to radiometric distortion, since differences in gain and bias between two images will not affect the ordering of pixels within a window. In addition, these transforms are tolerant to a small number of outliers within a window, and are therefore robust with respect to small amounts of random noise[21]. Two non-parametric transforms which are suited to fast implementation are the rank and census[14].

The rank transform is defined as the number of pixels in the window whose value is less than the centre pixel. The images will therefore be transformed into an array of integers, whose value ranges from 0 to $N - 1$, where $N$ is the number of pixels in the window. A pair of rank transformed images are then matched using one of the matching metrics of Table 1. For hardware implementation, it is advantageous to use a matching metric based on integer arithmetic, such as the SAD or the SSD. An example of the rank transform, for a 3 $\times$ 3 pixel window, is shown as follows:

$$
\begin{array}{ccc}
113 & 87 & 42 \\
96 & 74 & 51 \\
23 & 18 & 77
\end{array}
$$

In this example, there are 4 pixels in the transform window which are less than the centre pixel, therefore, the rank transform at this location is 4.

The census transform maps the window surrounding the centre pixel to a bit string. If a particular pixel's value is less than the centre pixel then the corresponding position in the bit string will be set to 1, otherwise it is set to zero. Two census transformed images are compared using a similarity metric based on the Hamming distance, ie, the number of bits that differ in the two bit strings. The Hamming distance

$$
\sum_{(u,v) \in W} \text{Hamming}(I_1(u,v), I_2(x + u, y + v))
$$

where $I_1$ and $I_2$ represent the census transforms of $I_1$ and $I_2$. Two hardware implementations of this scheme are discussed in [16, 17]. Using the same 3 $\times$ 3 window as Eq. (1), an example of the census transform is as follows:

$$
\begin{array}{ccc}
113 & 87 & 42 \\
96 & 74 & 51 \\
23 & 18 & 77
\end{array}
$$

In this example, the centre pixel is compared with all pixels in the census window, starting from the top left hand corner, and going from left to right along each row in turn.

4 Experimental Results

A matching scheme used to compare various area-based metrics is shown in Figure 3. In each case, the epipolar resampled images are input to the matching stage, which uses one of the metrics from Table 1 to determine the initial disparity maps with respect to each image. The left-right consistency criterion, in addition to filtering to remove locally anomalous matches, are then applied, in order to remove invalid matches. The resulting output consists of a disparity map with respect to the right image, from which invalid matches have been removed.

The steps involved in matching using the rank and census transforms are shown in Figure 4 and Figure 5 respectively. The rank transformed stereo images are matched using the SAD metric, while the census transformed images are matched using the Hamming measure of Eq. (2). In each case, the disparity maps output from the matcher may then be input to the validity checking stage of Figure 3.

The algorithms of Figures 3, 4 and 5 were tested using a number of stereo pairs of rock scenes[15]. Figure 6 shows one test stereo pair, IROCKS1. This

![Figure 3: Overall matching process using area-based matching metrics.](image-url)
Table 2: Proportion of matched pixels for each matching metric.

<table>
<thead>
<tr>
<th>Metric</th>
<th>SAD</th>
<th>ZSAD</th>
<th>SSD</th>
<th>ZSSD</th>
<th>NCC</th>
<th>ZNCC</th>
<th>RANK+SAD</th>
<th>CENSUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROCKS1</td>
<td>0.22</td>
<td>0.71</td>
<td>0.22</td>
<td>0.71</td>
<td>0.69</td>
<td>0.55</td>
<td>0.71</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Figure 4: Overall matching process using rank transform.

Figure 5: Overall matching process using census transform.

test pair was used in the JISCT stereo evaluation [22], and is affected by radiometric distortion, the left image being approximately 28% brighter than the right. The disparity maps obtained for this stereo pair using the area based metrics of Table 1 are shown in Figure 7. Lighter regions in the result disparity maps correspond to larger disparities, while black regions consist of invalid matches which were removed. A matching window size of $11 \times 11$ was used for each metric. The disparity results using the rank and census transforms are shown in Figure 8. The census transform was performed using windows of size $5 \times 5$, however, the matching process used windows of size $11 \times 11$. The proportion of matches remaining after validity checking for each metric are shown in Table 2.

5 Discussion

The SAD and SSD metrics are clearly not robust with respect to radiometric distortion, as shown in Figure 7(a) and (b). The disparity maps of Figure 7(c)–(f), show that the ZSAD, ZSSD, NCC and ZNCC metrics result in improved robustness to radiometric distortion. These metrics consequently yield a higher proportion of valid matches, as shown in Table 2. However, these metrics also result in increased computational complexity, since they consist of floating point operations. The NCC and ZNCC are particularly computationally expensive due to the presence of floating point multiplication, division and square root operations. In [23], the computational complexity of different area-based metrics and non-parametric transforms is quantified in terms of the amount of chip area required for the necessary operations in a hardware implementation. It is shown that the area requirement and subsequently computational complexity of matching using the SAD, SSD, rank and census is considerably less than that of the NCC metric.

Figure 8 shows that the rank and census transforms are robust to radiometric distortion. In Table 2, these algorithms resulted in a similar proportion of valid matches to the ZSAD, ZSSD and NCC metrics. These results were consistent not only for the stereo pair of Figure 6, but for all test imagery used [15].

It should also be noted that the proportion of matched pixels as shown in Table 2 is highly dependent on the content of the images. For example, stereo pairs containing large occluded regions would lead to a lower proportion of matched pixels for this pair. Also, the presence of large bland regions, for example, a background wall, can also decrease the proportion of matched pixels. Despite these perturbations, results for all test stereo pairs show that the SAD and the SSD are consistently out-performed by all the other matching metrics tested, as well as the rank and census transforms.

6 Conclusion

This paper has compared both area-based techniques and non-parametric transforms for possible use in a stereo vision sensor for mining automation applications. The requirements of this sensor are speed, reliability and the ability to produce a dense depth map.
Figure 6: IROCKS1 stereo pair. Note the radiometric distortion, the left image being approximately 28% brighter than the right.

Figure 7: Disparity of IROCKS1 stereo pair, produced using the (a) SAD, (b) SSD, (c) NCC, (d) ZSAD, (e) ZSSD, and (f) ZNCC metrics. Note the poor performance of the SAD and the SSD, due to radiometric distortion. While the ZSAD, ZSSD, NCC and ZNCC result in improved robustness, they introduce additional computational complexity.

Figure 8: Disparity of IROCKS1 stereo pair, produced using (a) Rank transform followed by SAD and (b) Census transform followed by the Hamming metric. The improvement of the rank and census results over the SAD and SSD clearly illustrates the robustness of the rank and census transforms to radiometric distortion. An additional advantage of the rank and census methods is that they do not introduce the computational overhead of the ZSAD, ZSSD, NCC and ZNCC.
Area-based techniques were considered because they have the potential to yield dense maps, are amenable to fast hardware implementation, and are suited to textured scenes. In addition, two matching algorithms based on non-parametric transforms have been tested — the rank transform followed by matching with the SAD metric, and the census transform followed by matching with the Hamming metric. Both were found to be robust with respect to radiometric distortion, thus improving matching reliability. This is significant because mining automation applications would typically employ low cost cameras, which would be prone to radiometric distortion.

Matching algorithms based on the rank and census transforms are also suitable for fast hardware implementation. In fact, census based implementations implemented in hardware are among the fastest of all matching implementations to date, taking the order of tens of us for a single similarity measure[24]. As a consequence of these two main advantages, the rank and census are prime candidates for use in a real-time, robust stereo matching system for mining automation applications.

For applications of stereo vision such as mining automation, it would be useful to not only discern disparity information, but in addition, to be able to identify invalid matches and/or assign a level of confidence to each match. In this work, two methods of identifying invalid matches: left–right consistency checking; and removal of locally anomalous disparities, have been used. It would be useful to develop a scheme which could compute an estimate of match confidence at each location of the disparity map. This scheme could be integrated into a real-time stereo vision sensor. However, the investigation of such a scheme has been left for future work.

Finally, it should be mentioned that the matching techniques discussed in this paper are not limited to mining automation applications. These techniques could be applied to any application involving stereo imagery, however, they would be best suited to textured scenes, and to applications where speed and robustness to radiometric distortion are an issue.

References


Jasmine Banks completed her B.Eng (Electronics) / B. Inf Tech at Queensland University of Technology in 1993. She then worked as a software engineer at BHA Computer, Brisbane, then as a research engineer at the Space Centre for Satellite Navigation at QUT. She commenced her PhD studies at QUT in 1997, and her research is being undertaken in collaboration with the Centre for Mining Technology and Equipment (CMTE). Her research interests are stereo vision and investigation of image matching techniques suitable for mining automation.

Mohammed Bennamoun finished his Masters of Science (M.Sc.) degree by research in “Control Theory” at Queen’s University, Canada, and PhD from Queensland University of Technology. He attended the International Space University (ISU) summer program in 1994 in Barcelona, Spain. He is currently the Director of the Space Centre for Satellite Navigation in the School of Electrical and Electronics Systems Engineering at Queensland University of Technology. He has been invited to organise invited sessions at international conferences. He participated in the organisation of many conferences, such as the International Symposium on Signal Processing and its Applications, and IEEE TENCON’97. His areas of interest are: control theory, robotics, obstacle avoidance, object recognition, artificial neural networks, and signal processing.

Peter Corke graduated in Electrical Engineering at the University of Melbourne in 1981, where he subsequently taught and completed a Masters degree. He joined CSIRO in 1984 and worked on a variety of projects involving robotics, high-speed machine vision, and machine vision applications. He spent one year as a CSIRO Overseas Fellow at the University of Pennsylvania GRASP laboratory, and in 1994 completed his PhD at the University of Melbourne in the area of high-performance visual control of machines. He is currently a Principal Research scientist with the CSIRO Division of Manufacturing Science and Technology. His research interests include visual servoing, robot dynamics and control, 3D vision, and high-speed machine vision architectures.