Fast and Robust Stereo Matching Algorithms for Mining Automation

Jasmine Banks
Space Centre for Satellite Navigation,
Queensland University of Technology /
Centre for Mining Technology and Equipment
Email: j.banks@qut.edu.au

Mohammed Bennamoun
Space Centre for Satellite Navigation,
Queensland University of Technology
Email: m.bennamoun@qut.edu.au

Peter Corke
CSIRO Manufacturing Science and Technology/
Centre for Mining Technology and Equipment
Email: pic@cat.csiro.au

KEY WORDS: stereo vision, image matching, rank transform, census transform

ABSTRACT

The mining environment, being complex, irregular and time varying, presents a challenging prospect for stereo vision. The objective is to produce a stereo vision sensor suited to close-range scenes consisting primarily of rocks. This sensor should be able to produce a dense depth map within real-time constraints. Speed and robustness are of foremost importance for this investigation. A number of area based matching metrics have been implemented, including the SAD, SSD, NCC, and their zero-meaned versions. The NCC and the zero meaned SAD and SSD were found to produce the disparity maps with the highest proportion of valid matches. The plain SAD and SSD were the least computationally expensive, due to all their operations taking place in integer arithmetic, however, they were extremely sensitive to radiometric distortion. Non-parametric techniques for matching, in particular, the rank and the census transform, have also been investigated. The rank and census transforms were found to be robust with respect to radiometric distortion, as well as being able to produce disparity maps with a high proportion of valid matches. An additional advantage of both the rank and the census transform is their amenability to fast hardware implementation.

1 INTRODUCTION

Perception of the three-dimensional environment is a prerequisite 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. A number of techniques exist for discerning three-dimensional information from scenes[Jarvis, 1983], including:

Active range finding methods in which a controlled energy beam is applied to the scene and reflected energy detected. These techniques, although well suited to static scenes and underground situations, are too slow for real-time applications[Corke et al, 1997].

Structured lighting approaches which include striped, grid and patterned lighting. These tend to be useful in tightly controlled domains such as industrial automation, and usually are not useful for more general environments[Barnard and Fischler, 1987].

Monocular image based techniques which use information from a single image, and include depth from focus, shape from shading, depth from occlusion cues and depth from texture gradient[Sonka et al, 1993]. Depth and object orientation are typically inferred from statistical assumptions.

Stereo vision which involves taking two (or more) images from different perspectives, and computing depth from stereo disparity.

Both range finding and structured lighting methods fall into the category of *active* methods, which require an external light or energy source to be applied to the scene. In contrast, both stereo and monocular techniques are known as *passive* techniques, which require only the presence of ambient light. Our work is primarily concerned with stereo vision, since it has the potential to produce results within real-time constraints and is suited for irregular environments.

The mining environment, being complex, irregular and time varying, presents a challenging prospect for stereo vision. Our objective is to produce a stereo vision sensor suited to close-range scenes consisting primarily of rocks. This sensor should produce a dense depth map within real-time constraints. It is also desired that the system be robust, ie, in the case of an invalid match, an answer of NO MATCH should be returned, rather than an erroneous result. However, a high level of accuracy is not of foremost importance for this application, since it is desired to obtain information concerning the presence and overall extent of obstacles, for example, rather than the precise location of their boundaries.

A fundamental issue is to establish *correspondence* or *matching* of points in two images, such as the rock stereo pair of **Figure 1**, in order to compute the disparity and subsequently the 3-D world coordinates. Image matching techniques are divided into two main categories[Fua, 1993]:

Area-based which are distinguished by the fact that actual grey-level pixel values in the images are compared to find the best match. The information contained in a single pixel is not sufficient for unambiguous matching, therefore regularly sized pixel neighbourhoods are compared.

Feature-based which are characterised by the use of image features such as edges, vertices and contours as the matching primitives. These methods rely on feature extraction. The representations of these features are then compared to find the best match.

Feature-based matchers tend to be faster than area-based methods, since only a small subset of pixels are used. How-

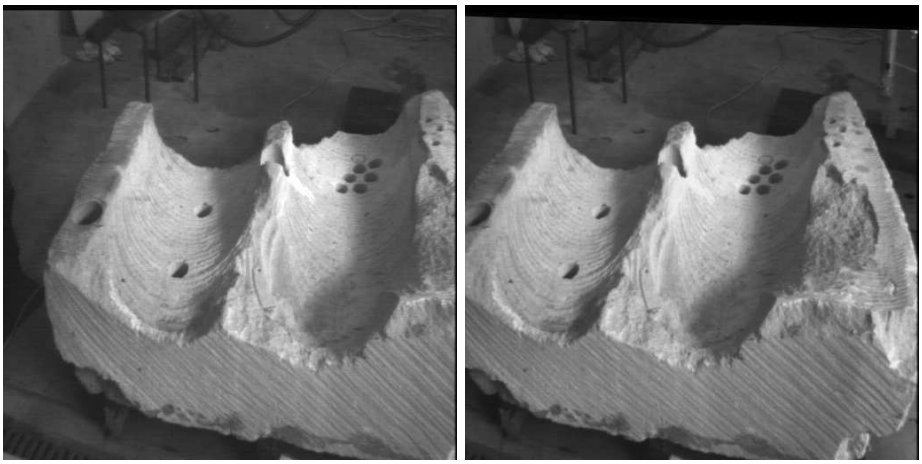


Figure 1: Stereo pair ROCK.

ever, they typically yield very sparse depth maps, since matching only takes place at image locations where features occur, and results for intermediate points must be obtained by interpolation[Grimson, 1981; Li, 1995]. This interpolation process relies on assumptions about the scene geometry between features. Feature-based matchers are also highly accurate since features may 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. Such scenes would typically be comprised of man-made objects. Area-based matchers are usually unsuitable to use on these images, since their smooth surfaces lack sufficient texture for an area-based matcher to match on.

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[Hannah, 1989]. Area-based matchers can also potentially yield matching results for every image pixel and hence yield a dense depth map. The advantages of area-based algorithms include their simplicity and straightforward implementation, as well as their amenability to hardware realisation. However, their accuracy is not as high as the feature-based methods. This is due to the "smoothing" effect introduced by using a square window of pixels for matching[Cochran and Medioni, 1992].

Our work to date has investigated area-based matching techniques, for the following reasons:

1. Scenes comprised of rocks usually have a large amount of surface texture, and are therefore well suited to area based matching.
2. They have the potential to yield a dense depth map.
3. They are amenable to fast hardware implementation.

Section 2 discusses area-based matching metrics, and also describes some of the problems which can make matching more difficult. Validation techniques to identify incorrect matches are also discussed. Section 3 outlines the theory of non-parametric transforms, in particular, the rank and the census transform. An implementation of area-based matching metrics, the rank transform, and the census transform are described in section 4. The disparity results obtained from this implementation are included in Section 5. Section 6 discusses the robustness of each approach to radiometric distortion,

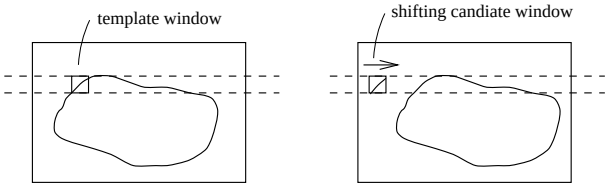


Figure 2: Epipolar constrained area based matching.

and recommends an algorithm as being most suitable for fast hardware implementation.

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 window of pixels in one image and a window in another image, and hence are used to determine the optimum match.

Epipolar geometry [Barnard and Fischler, 1987] 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[Ayache, 1991]. 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 2**. 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 chosen.

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, eg, 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.

As discussed in the remainder of this section, matching metrics such as the ZSAD, ZSSD, NCC and ZNCC are designed to be robust with respect to radiometric distortion, while validation techniques such as left-right consistency checking are able to identify most of the invalid matches due to occlusions and bland regions. Non-parametric techniques, discussed in Section 3, are invariant to radiometric distortion and small amounts of random noise.

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. The SAD and the SSD are intuitively the simplest, and computationally the least expensive of all the matching measures[Hannah, 1974]. Two areas which consist of exactly the same pixel values would yield a score of zero. However, these measures 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 "best" 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 [Hannah, 1989; Fua, 1993], which involves reversing the roles of the two images and performing matching a second time, as illustrated by **Figure 3**. Firstly, epipolar constrained matching is carried out using a template window centred on I_1,

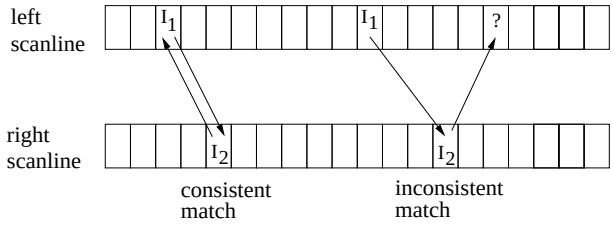


Figure 3: Consistent and inconsistent matches. The match on the left is consistent, while the match on the right is inconsistent[Fua, 1993].

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 isolated matches from the matches which remain after left-right consistency checking. This heuristic is based on the assumption that isolated matches are more likely to be incorrect[Fua, 1993; Faugeras et al, 1993].

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[Bhat and Nayar, 1996].

Two non-parametric transforms which are suited to fast implementation are[Zabih and Woodfill, 1994]:

Rank Transform This 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 section 2.2. For hardware implementation, it is advantageous to use a matching metric based on integer arithmetic, such as the SAD or the SSD.

Census Transform This 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 is

Sum of Absolute Differences	SAD	$\sum_{(u,v)\in W} \lvert I_1(u,v) - I_2(x+u,y+v)\rvert$
Zero mean Sum of Absolute Differences	ZSAD	$\sum_{(u,v)\in W} \left\lvert (I_1(u,v) - \overline{I_1}) - (I_2(x+u,y+v) - \overline{I_2})\right\rvert$
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	$\sum_{(u,v)\in W} ((I_1(u,v) - \overline{I_1}) - (I_2(x+u,y+v) - \overline{I_2}))^2$
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	$\frac{\sum_{(u,v)\in W} (I_1(u,v) - \overline{I_1})\cdot (I_2(x+u,y+v) - \overline{I_2})}{\sqrt{\sum_{(u,v)\in W} (I_1(u,v) - \overline{I_1})^2 \cdot \sum_{(u,v)\in W} (I_2(x+u,y+v) - \overline{I_2})^2}}$

Table 1: Area based matching measures[Aschwanden and Güggenbuhl, 1993]. 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.

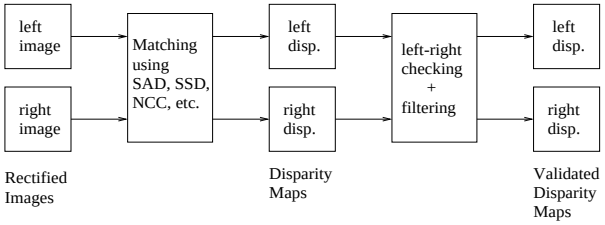


Figure 4: Overall matching process using area-based matching metrics.

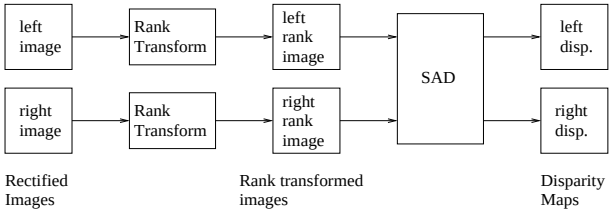


Figure 5: Overall matching process using rank transform.

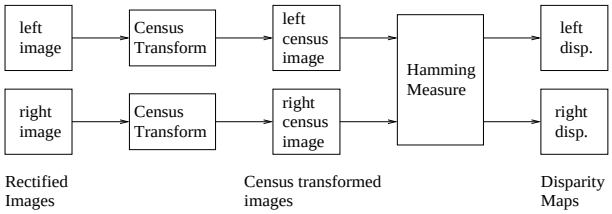


Figure 6: Overall matching process using census transform.

summed over the window, ie,

$$\sum_{(u,v)\in W} \text{Hamming}(I_1'(u,v), I_2'(x+u,y+v)) \quad (1)$$

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 [Woodfill and Herzen, 1993; Dunn and Corke, 1997].

4 IMPLEMENTATION

A number of area-based matching metrics and non-parametric transforms were implemented in software, in order to investigate their suitability for scenes containing rocks. All implementations accept a rectified stereo pair as input and output disparity maps with respect to each image. Either of the result disparity maps may then be used for 3-D reconstruction.

4.1 Area-Based Matching Metrics

The overall algorithm used to test various area-based matching metrics is shown in the block diagram of **Figure 4**. The left and right 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 isolated matches, are then applied, in order to remove invalid matches.

4.2 Non-Parametric Transforms

The steps involved in matching using the rank and census transforms are shown in **Figure 5** and **Figure 6** respectively. The rank transformed stereo images are matched using the SAD metric, while the census transformed images are matched using the Hamming measure of **Equation 1**. In each case, the disparity maps output from the matcher may then be input to the validity checking stage of **Figure 4**.

5 RESULTS

The disparity results obtained for the stereo pair of **Figure 1** using a number of matching metrics, the rank transform, and

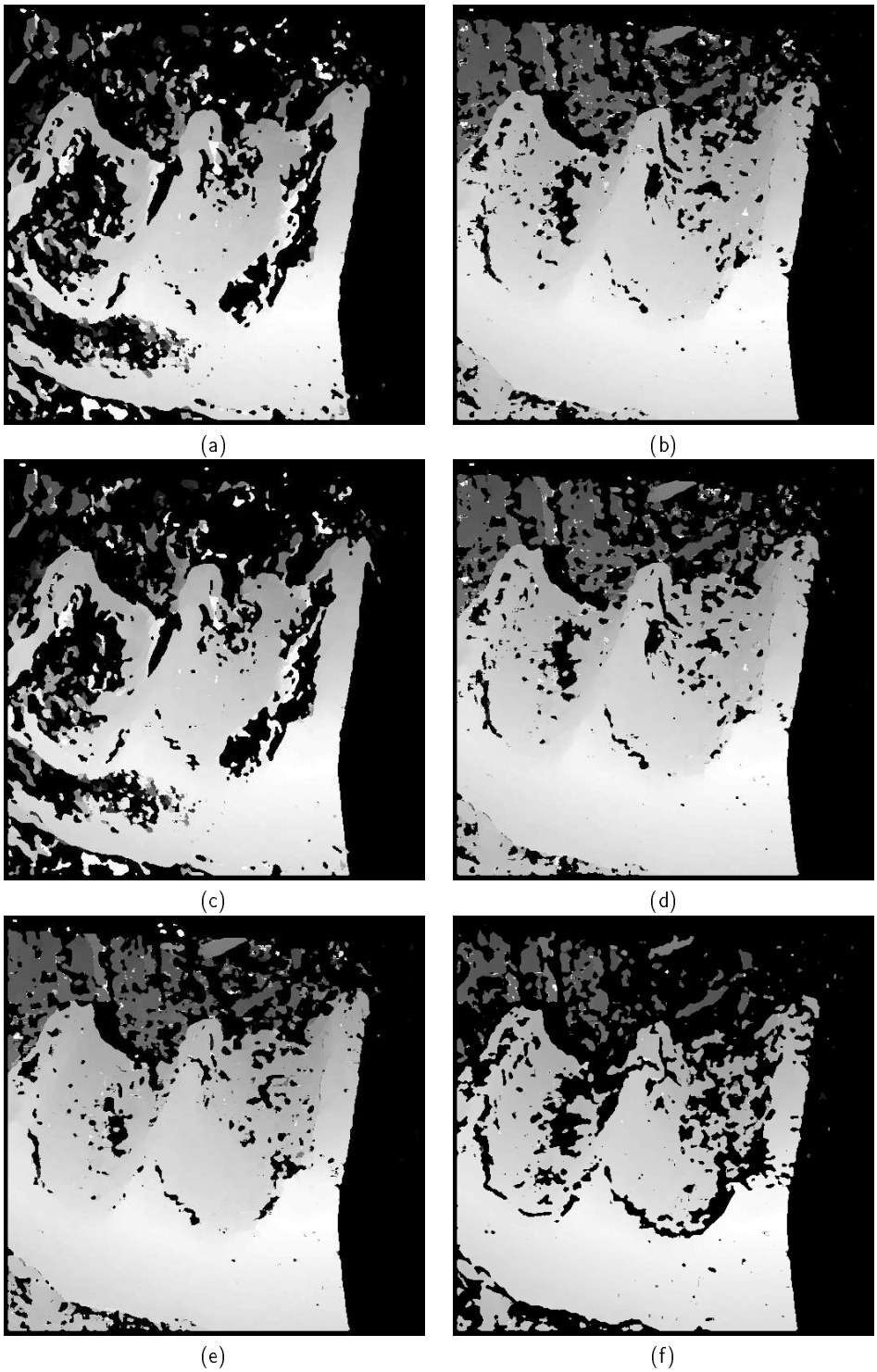
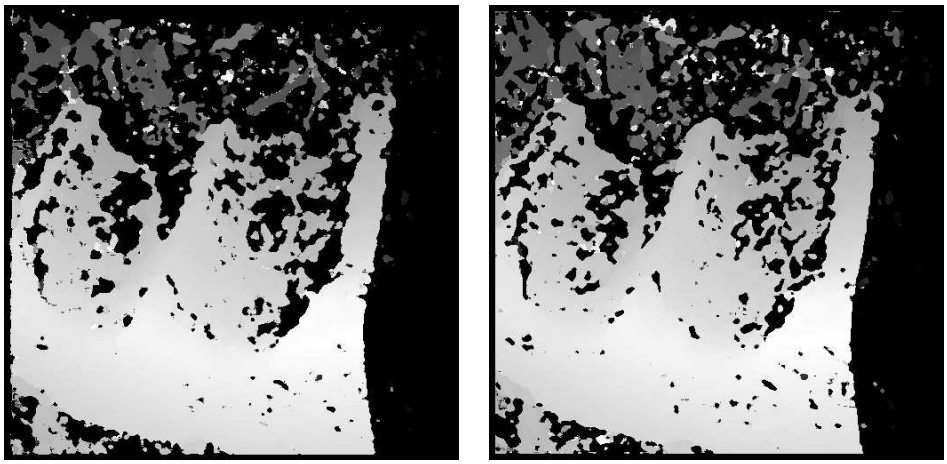


Figure 7: Disparity of ROCK stereo pair, produced using (a) SAD, (b) ZSAD, (c) SSD, (d) ZSSD, (e) NCC and (f) ZNCC metric. The ZSAD, ZSSD, NCC and ZNCC metrics result in the highest proportion of valid matches, however, these metrics have a significantly higher computational overhead than the SAD and SSD.



(a) (b)

Figure 8: Disparity of ROCK stereo pair, produced using (a) Rank transform followed by SAD and (b) Census transform followed by the Hamming metric. The rank and census methods result in a higher proportion of valid matches than the SAD and SSD, and in addition, they do not introduce the computational overhead of the ZSAD, ZSSD, NCC and ZNCC.

the census transform, are shown in **Figure 7** and **Figure 8**. In each case, the disparity map with respect to the right image is shown. Lighter regions in the result disparity maps correspond to larger disparities. A matching window size of 11×11 was used for each metric. The census transform was performed using windows of size 5×5, however, the matching process used windows of size 11×11.

The test stereo pairs, IROCKS1, J1 and K1, are shown in **Figure 9, 12** and **15** respectively. These test pairs were used in the JISCT stereo evaluation[Bolles et al, 1993], and are all affected by radiometric distortion, one image being brighter than the other in each case. The disparity maps obtained for these stereo pairs are shown in **Figure 10, 13** and **16**, while the results obtained using the rank and census transforms are shown in **Figure 11, 14** and **17**.

For each of the test images, the proportion of matches remaining after validity checking for each metric are shown in **Table 2**.

6 DISCUSSION

It can be seen from **Figure 10, 13** and **16** that the SAD and the SSD are clearly not robust with respect to radiometric distortion. These metrics have performed particularly poorly in the case of the IROCKS1 pair, in which the left image is approximately 28% brighter than the right. These metrics also performed quite poorly in the case of the J1 and K1 pairs, in which the right image is approximately 13% and 14% brighter than the left, respectively. Use of the ZSAD, ZSSD, NCC and ZNCC resulted in improved robustness and consequently a higher proportion of valid matches, as shown in **Table 2**. However, these metrics 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.

The proportion of matched pixels as shown in **Table 2** is highly dependent on the content of the images. For example, the ROCK pair contains a large area of pixels which are only visible in one image. This results in an unmatched area on the right hand side of each disparity map in **Figure 7 and 8**, which in turn leads to a lower proportion of matched pixels for this pair. However, **Table 2** shows that the SAD and the SSD are consistently out-performed by all the other matching metrics tested, as well as the rank and census transform techniques.

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, as shown by **Figure 8, 11, 14** and **17**. As shown in **Table 2**, both algorithms produced disparity maps with a high proportion of valid matches. An additional advantage of both these algorithms is their amenability to fast hardware implementation. Consequently, they are prime candidates for a real-time, robust stereo matching system for mining automation applications.

ACKNOWLEDGMENTS

This work was conducted as part of the "Automation" research program of the Cooperative Research Centre for Mining Technology and Equipment (CMTE).

REFERENCES

- Aschwanden, P. and Guggenbühl, W. (1993): Experimental Results from a Comparative Study on Correlation-Type Registration Algorithms. Robust Computer Vision, 268–289, Wickmann, 1993.
- Ayache, N. (1991): Artificial Vision for Mobile Robots, MIT Press, 1991.
- Barnard, S. and Fischler, M. (1987): Stereo Vision. Encyclopedia of Artificial Intelligence, 1083–1090, John Wiley & Sons, 1987.

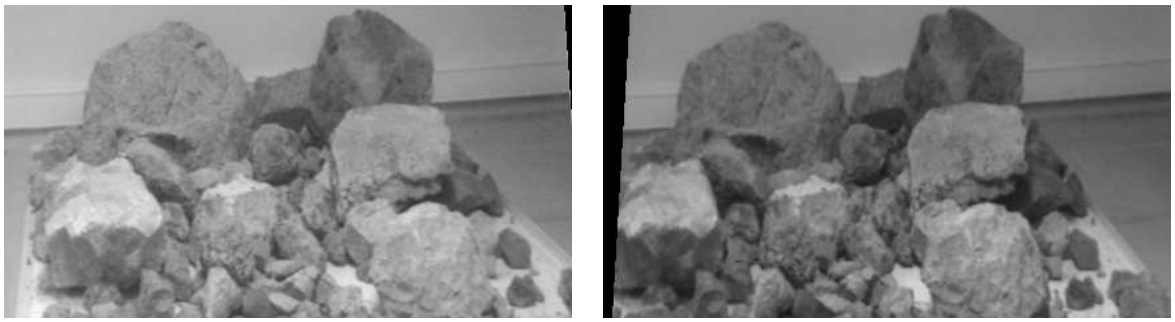


Figure 9: IROCKS1 stereo pair. Note the radiometric distortion, ie, the left image is approximately 28% brighter than the right.

	SAD	ZSAD	SSD	ZSSD	NCC	ZNCC	RANK+SAD	CENSUS
ROCK	0.48	0.65	0.49	0.65	0.65	0.65	0.52	0.61
IROCKS1	0.22	0.71	0.22	0.71	0.69	0.55	0.71	0.77
J1	0.40	0.75	0.46	0.75	0.76	0.72	0.77	0.81
K1	0.47	0.72	0.51	0.71	0.73	0.68	0.79	0.83

Table 2: Proportion of matched pixels for each matching metric, for each test stereo pair.

- Bhat, D. and Nayar, S. (1996): Ordinal Measures for Visual Correspondence. Proceedings of Computer Vision and Pattern Recognition, San-Fransisco, 351–357, 1996.
- Bolles, R., Baker, H. and Hannah, M. (1993): The JISCT Stereo Evaluation. Image Understanding Workshop, DARPA, 263–274, 1993.
- Cochran, S. and Medioni, G. (1992): 3-D Surface Description from Binocular Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 14, Number 10, 981–994, October 1992.
- Corke, P., Winstanley, G. and Roberts, J. (1997): Sensors and Control for Mining Robotics. Fourth International Symposium on Mine Mechanisation and Automation, Brisbane, Australia, B1-11—B1-21 , 6-9 July 1997.
- Dunn, P. and Corke, P. (1997): Real-time Stereopsis using FPGAs. FPGA97, Imperial College London, September 1997.
- Faugeras, O., et al (1993): Real-Time Correlation-Based Stereo: Algorithm, Implementations and Applications. Technical Report 2013, INRIA, 1993.
- Fua, P. (1993): A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features. Machine Vision and Applications, Volume 6, 35–49, 1993.
- Grimson, W. (1981): From Images to Surfaces, MIT Press, 1981.
- Hannah, M. (1974): Computer Matching of Areas in Stereo Images. PhD thesis, Stanford University, 1974.
- Hannah, M. (1989): A System for Digital Stereo Image Matching. Photogrammetric Engineering and Remote Sensing, Volume 55, Number 12, 1765–1770, December 1989.
- Jarvis, R. (1983): A Perspective on Range Finding Techniques for Computer Vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 5, Number 2, 569–586, March 1883.
- Li, S. (1995): MRF Modeling in Computer Vision, Springer-Verlag, 1995.
- Sonka, M., Hlavac, V. and Boyle, R. (1993): Image Processing, Analysis and Machine Vision. Chapman and Hall, 1993.
- Woodfill, J. and Herzen, B. (1993): Real-Time Stereo Vision on the PARTS Reconfigurable Computer. IEEE Workshop in FPGAs for Custom Computing Machines, 242–250, April 1997.
- Zabih, R. and Woodfill, J. (1994): Non-parametric Local Transforms for Computing Visual Correspondence. 3rd European Conference on Computer Vision, Stockholm, 1994.

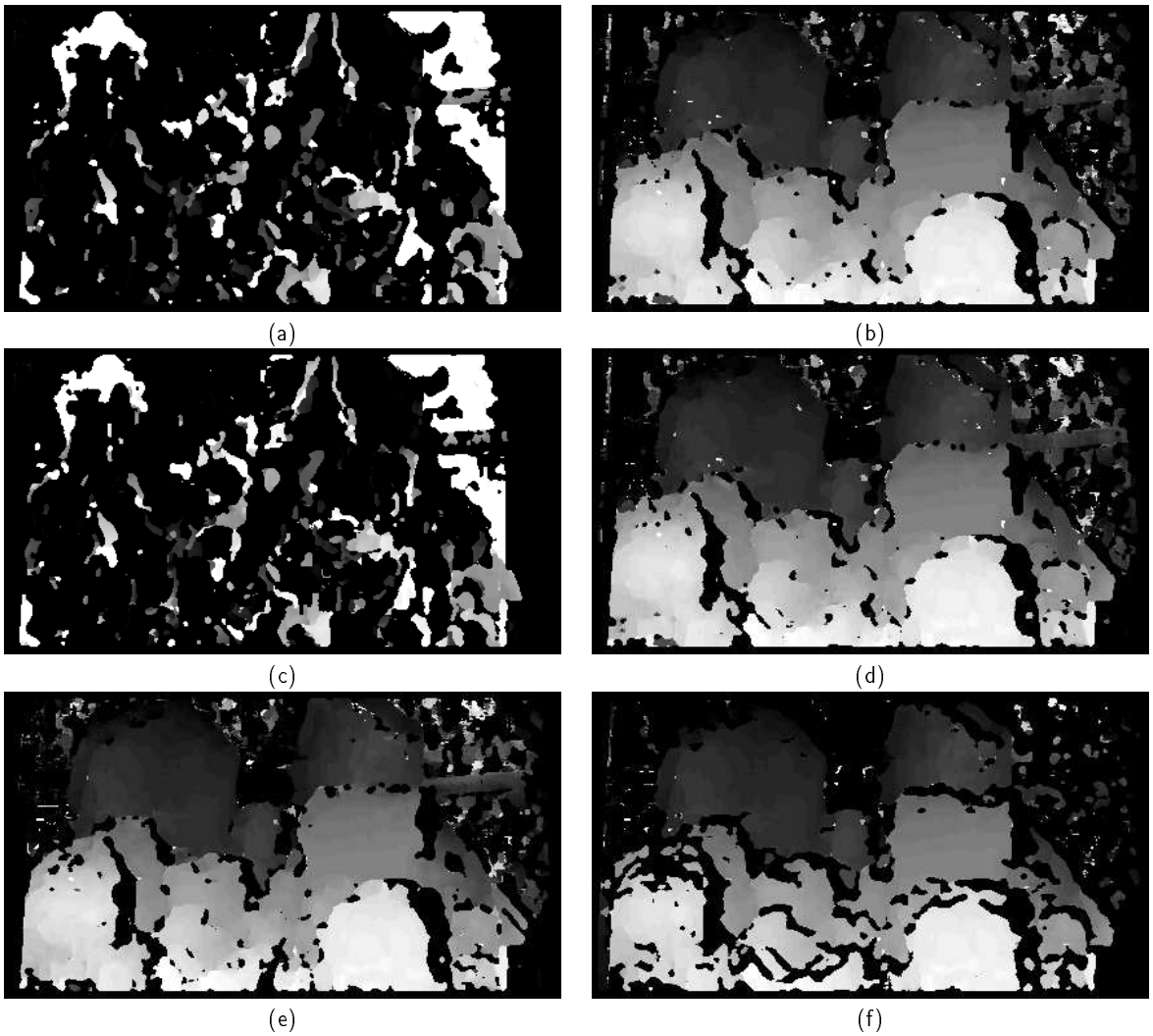


Figure 10: Disparity of IROCKS1 stereo pair, produced using (a) SAD, (b) ZSAD, (c) SSD, (d) ZSSD, (e) NCC and (f) ZNCC metric. Note the poor performance of the SAD and the SSD, due to radiometric distortion. The ZSAD, ZSSD, NCC and ZNCC result in improved robustness, however, they introduce additional computational complexity.

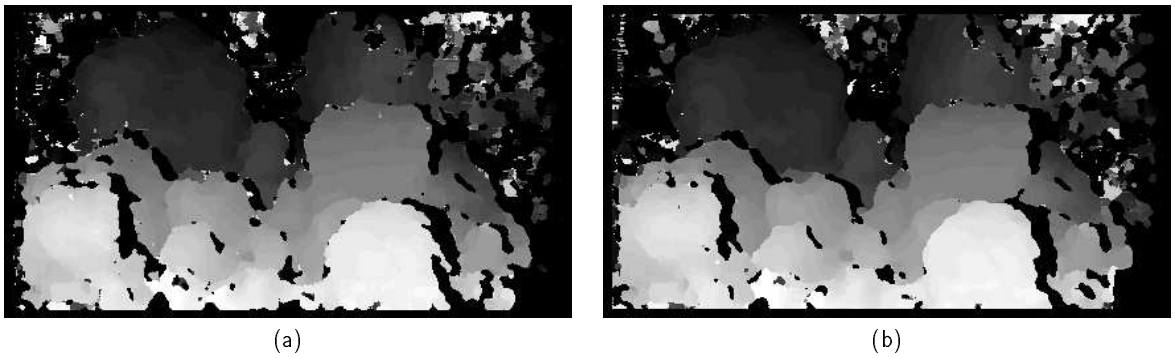


Figure 11: Disparity of IROCKS1 stereo pair, produced using (a) Rank transform followed by SAD and (b) Census transform followed by the Hamming metric. The improvement of these results over the SAD and SSD results of Figure 10 clearly illustrates the robustness of the rank and census transforms to radiometric 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.

Figure 12: J1 stereo pair. The right image is approximately 13% brighter than the left.

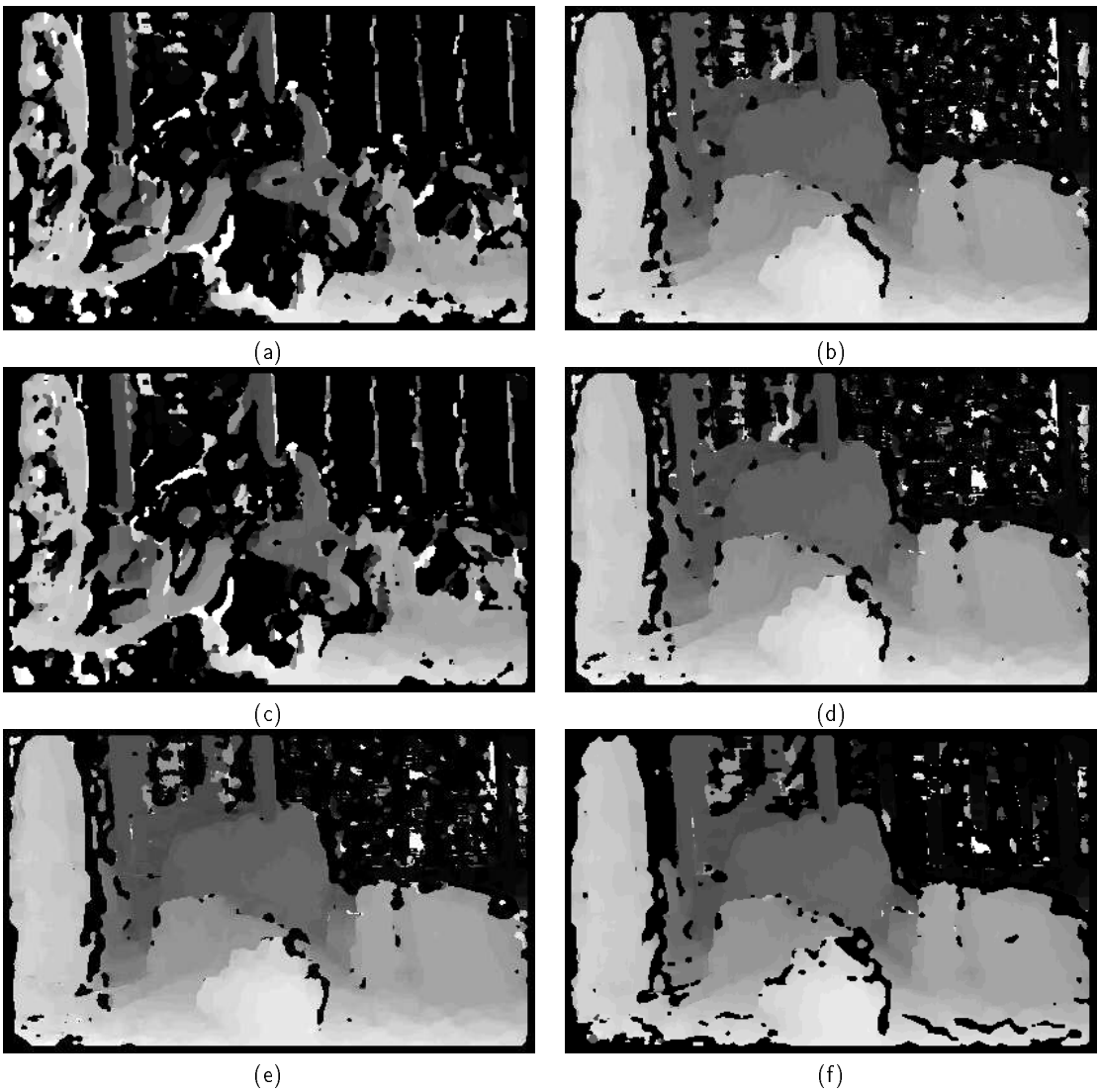
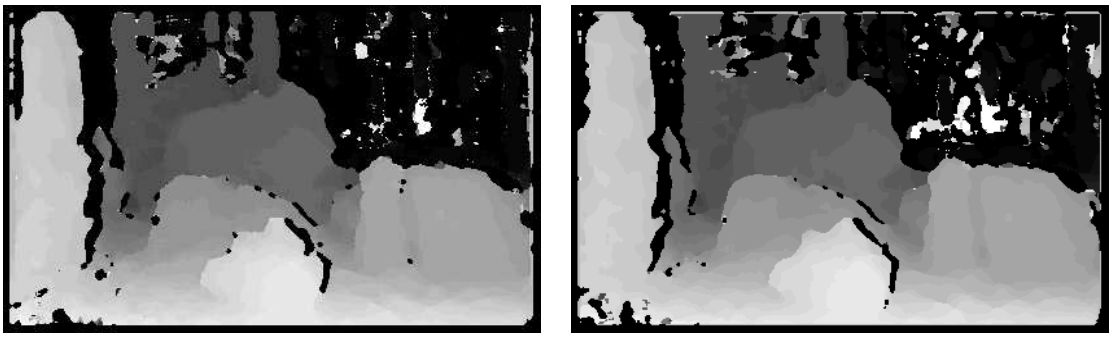


Figure 13: Disparity of J1 stereo pair, produced using (a) SAD, (b) ZSAD, (c) SSD, (d) ZSSD, (e) NCC and (f) ZNCC metric. As with Figure 10, the poor performance of the SAD and the SSD is due to radiometric distortion. The ZSAD, ZSSD, NCC and ZNCC result in improved robustness, however, they introduce additional computational complexity.



(a) (b)

Figure 14: Disparity of J1 stereo pair, produced using (a) Rank transform followed by SAD and (h) Census transform followed by the Hamming metric. As with Figure 11, the rank and census transforms result in improved robustness in the case of radiometric distortion, without introducing the computational complexity of the ZSAD, ZSSD, NCC and ZNCC.

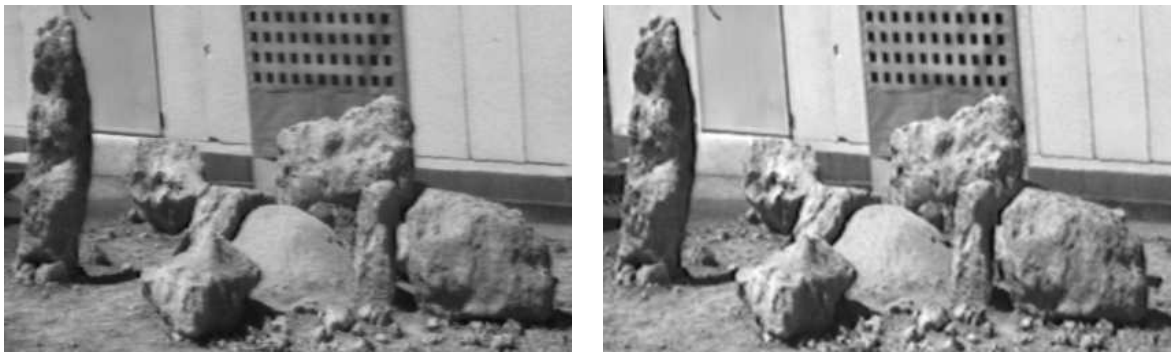


Figure 15: K1 stereo pair. The right image is approximately 14% brighter than the left.

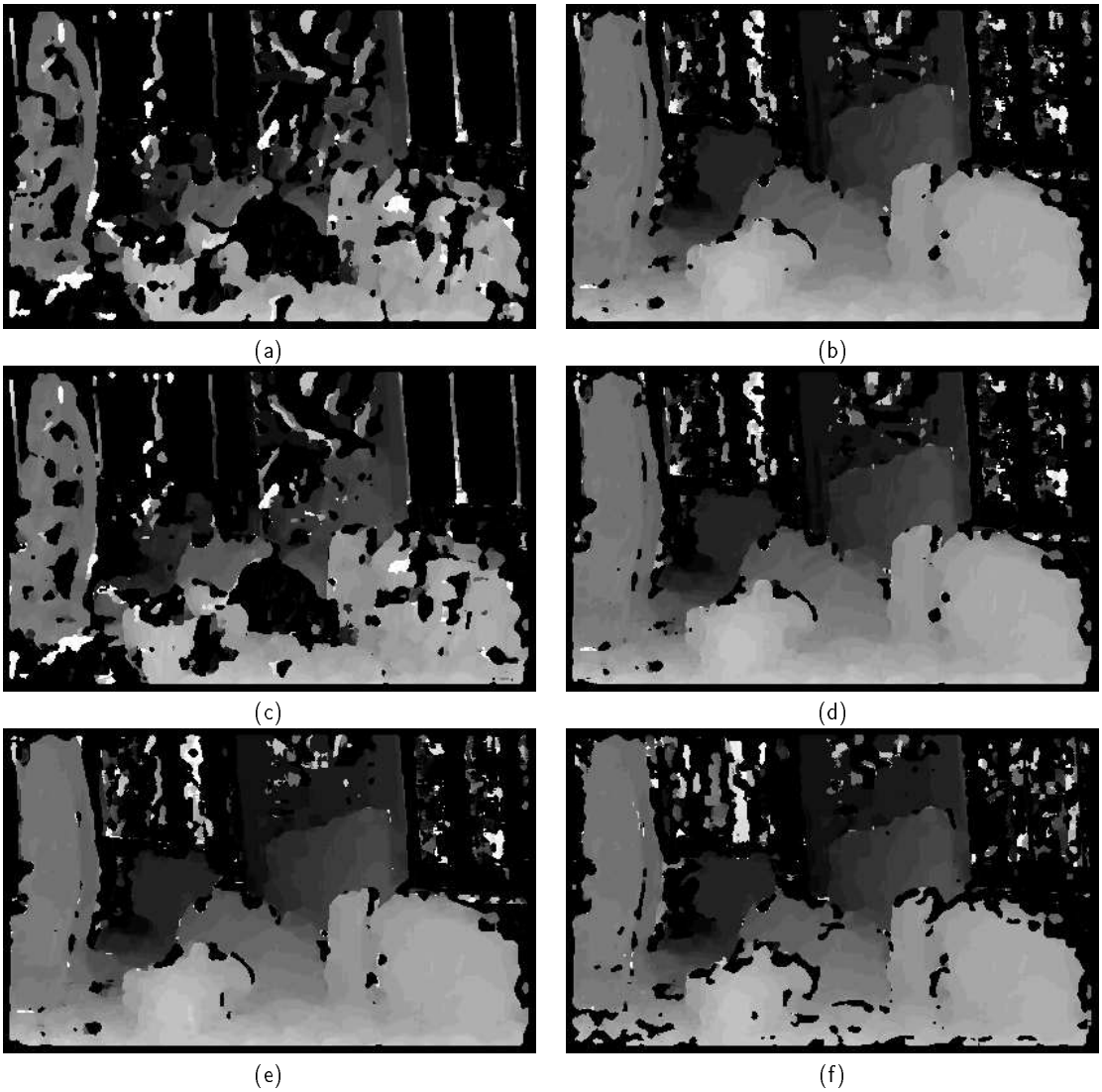


Figure 16: Disparity of K1 stereo pair, produced using (a) SAD, (b) ZSAD, (c) SSD, (d) ZSSD, (e) NCC and (f) ZNCC metric. As with Figure 10 and 13, the poor performance of the SAD and the SSD is due to radiometric distortion. The ZSAD, ZSSD, NCC and ZNCC result in improved robustness, however, they introduce additional computational complexity.

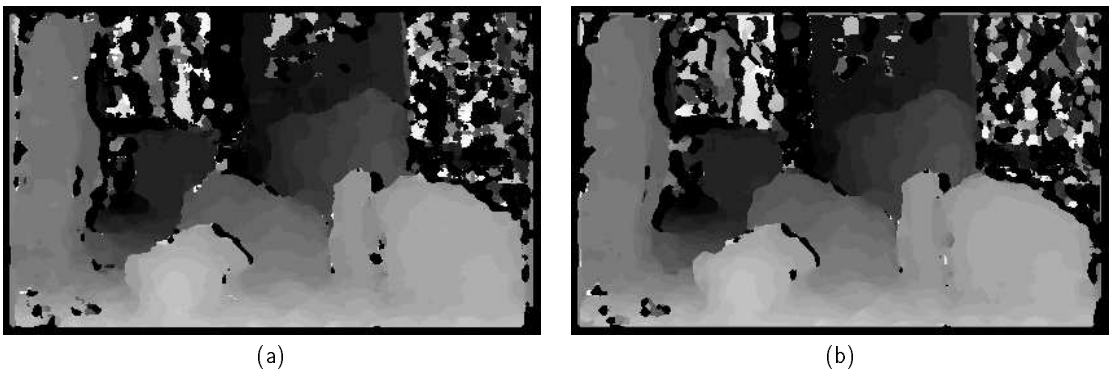


Figure 17: Disparity of K1 stereo pair, produced using (a) Rank transform followed by SAD and (b) Census transform followed by the Hamming metric. As with Figure 11 and 14, the rank and census transforms result in improved robustness in the case of radiometric distortion, without introducing the computational complexity of the ZSAD, ZSSD, NCC and ZNCC.