

COMPARATIVE STUDY OF DISPARITY ESTIMATIONS WITH MULTI-CAMERA CONFIGURATIONS IN RELATION TO DESCRIPTIVE PARAMETERS OF COMPLEX BIOLOGICAL OBJECTS

Michael Nielsen*, Hans J. Andersen, Erik Granum

Aalborg University
Laboratory of Computer Vision and Media Technology
mnielsen@cvmt.dk

KEY WORDS: 3D reconstruction, Multi-Baseline, Trinocular Stereo, Graph Cuts, Performance Evaluation, Biological Structures, Remote Sensing.

ABSTRACT

This paper aims at evaluating multi-camera configurations as a function of the descriptive parameters of complex biological objects. Multi-baseline Stereo has potential to handle projective distortion at large baselines. Being close to the observed object and the orientation of object surfaces pointing toward the camera increase the *projection distortion*. An example is 3D reconstruction of plants where the leaves can be pointing steeply toward the cameras, while, sub-leaf reconstruction needs high depth resolution, because the leaves overlap closely to each other. The paper presents a new dissimilarity measure, called Sums of Individual Sums of Squared Differences (SISSD). It takes projection distortion and changing specular highlights into account by learning the gradual changing of the feature window. The method was included in the comparative study that used realistic ray traced plant models, where the descriptive parameters of the objects could be controlled. Other configurations in the experiment were the commonly used Multi-baseline Sum of the Sums of Squared Differences (SSSD), the popular binocular graph cuts, and two trinocular correlation techniques. Comparison is in regard to leaf type, texture and orientation, proportion of occlusion and proportion of changing highlights by computing the overall-, occluded-, and highlighted- percentage of bad matching pixels ($pbmp$, $pbmp_{occ}$, and $pbmp_{high}$). The results showed a complicated relationship of trade-offs that points toward further development combining the strengths of the individual configurations.

1 INTRODUCTION

Computer vision based 3D reconstruction of close-up complex biological structures is a difficult discipline. There are various multi-camera configurations to choose from. It would be useful to learn about the performance related to descriptive parameters of the objects at hand, in order to choose the best configuration. The Descriptive parameters of the objects are *surface shape*, *surface orientation*, *presence of texture*, *proportion of changing specular highlight* and *proportion of occlusion*. The specular highlights in concern are those that changes gradually from one image to the next across the baseline. Multi-baseline Stereo has been described and tested in literature as a method for improving the handling of occlusion and ambiguity across the scan lines (Okutomi and Kanade, 1993)(Jeon et al., 2001) by using the sum of the energy measures across the camera array; e.g. Sum of Sums of Squared Difference (SSSD). Attempts have also been made at dealing with specular highlights by actively detecting specular highlights within the algorithm (Li et al., 2002) and treating them as occlusions. However, the problems related to nearby objects are overlooked as the algorithms assume that the area looks the same in all cameras. This paper presents an alternative measure that utilizes the fact that a multi baseline array consists of subsets of smaller baselines. A large baseline improves depth resolution but it also makes the correspondence more difficult (Okutomi and Kanade, 1993). Three factors increase this effect: Being close to the observed object, window correlation size, and orientation of object surfaces.

Precision agriculture is a field with rising interest in 3D computer vision, which is becoming tangible as new high dynamic range cameras and precalibrated multi-view cameras are being developed. These cameras satisfy the epipolar geometry constraints and the intrinsic- and extrinsic calibration can be skipped. Close-up 3D reconstruction of plants is an excellent example where the

leaves can be pointing steeply toward the cameras and it needs high depth resolution because the leaves overlap closely to each other. Excellent depth maps has potential to aid the segmentation of individual leaves (Lee et al., 1996), if the disparity maps have trustworthy discontinuity edges. This is useful in precision agriculture for segmenting individual leaves for autonomous weed identification, fruit picking, branch thinning, and for finding sampling points on specific locations of a plant (Christensen and Jørgensen, 2003,)(Nielsen et al., 2004). The image acquisition is expected to be done from a moving platform in an outdoor environment, so reconstruction must be done from a single time slice.

In general terms plants belong to the class of objects that are: semitransparent, biological, non-rigid structures. Disparities are often non-planar and can get very *steep* toward the cameras. Textures are non-existent or highly detailed, and having more or less specular highlights. Fortunately, they are segments of smooth surfaces, but intertwining and overlapping. It is very difficult to get dense ground truth. The Vision based depth map reconstruction is usually confined to fronto-planar depth scenes, where the depth maps can be described as regions of near-equal disparities. These scenes are viewed from a distance and have small finite disparity spaces, where it is reasonable to manually acquire ground truth. As an alternative, structured light can be used. It uses multiple images so that the objects must be rigid in time (Scharstein and Szeliski, 2003).

2 METHODS AND MATERIAL

The stereo correspondence algorithms were all based on a basic Sum of Squared Difference (SSD) dissimilarity (energy) function (eq. 1). *The presented methods assumes precalibrated images satisfying epipolar geometry constraints, equal baseline, and zero rotation.*

$$E_{i,j}(x, y, d) = \sum_{(u,v) \in W(x,y)} (I_i(u, v) - I_j(u + d, v))^2 \quad (1)$$

d is the tested disparity, W is the window around (x, y) , I_i is the i th image. The windows can be placed in various ways around the pixel and question, but we limited this experiment to centered windows. Adding multiple windows can improve the correspondence near disparity borders (Fusiello et al., 2000), but we wanted to keep this factor out of the experiment this time. It was shown in another experiment that five symmetric windows were optimal, ie. the center and the four diagonals (Nielsen et al., 2005).

In the classical multi baseline SSSD the Sum of Squared Difference between the reference camera and the i th camera is computed for N cameras. See equation 2.

$$S(x, y, d) = \arg \min_d \sum_{c=2}^N (E_{1,c}(x, y, \frac{d(c-1)}{N-1})) \quad (2)$$

We see that the binocular case ($N = 2$) is a special case of this equation.

2.1 Introducing SISSD

A new measure Sum of Individual Sums of Squared Differences is defined as SISSD (see equation 3). This measure was supposed to learn the graduate change in the feature window across the baseline. This could be a problem with occlusions as it would learn the feature of the occluding object, which was countered by including the weighted dissimilarity in regard to the reference camera. In the new measure we computed the Sum of Squared Difference between the $i - 1$ th and the i th camera, and between the 1st and the i th camera to ensure that it does not adapt to a completely different object.

$$\begin{aligned} S(x, y, d) &= \arg \min_d \sum_{c=2}^N [\alpha(E_{c-1,c}(x, y, \frac{d(c-1)}{N-1})) \\ &\quad + (1-\alpha)(E_{1,c}(x, y, \frac{d(c-1)}{N-1}))] \end{aligned} \quad (3)$$

We see that SSSD is a special case of SISSD, where $\alpha = 0.0$. Figure 1 shows an example of the case with steep object where the projection distorts the orientation of the leaf. The top shows parts of images of a five camera array. The middle plot the development of the dissimilarity (energy) across increasing baseline. It is obvious that SSSD increases exponentially, while SISSD is even less than SSD. The bottom plot shows the dissimilarity for the three measures across the scan line and prints the best match for SSD, SSSD, SISSD and Ground Truth (GT). This trait should also be an advantage in the presence of specular highlights that travel across the baseline. An example is shown in figure 2. Based on these preliminary results, a benchmark experiment was performed. The goal was to validate that SISSD performed better than SSSD on steep-leaved objects and in areas where the specular highlight state changes, and whether the reference similarity constraint could counter the occlusion problem.

2.2 Comparative Methods

The other common multi-camera alternative to the multi baseline camera array is called the right-angled trinocular L-setup (Mulligan and Daniilidis, 2002). Two different trinocular algorithms are used for comparison, trinocular minimum (T_m eq. 4) and trinocular sum (T_s eq. 5). In principle, they use two image pairs, where

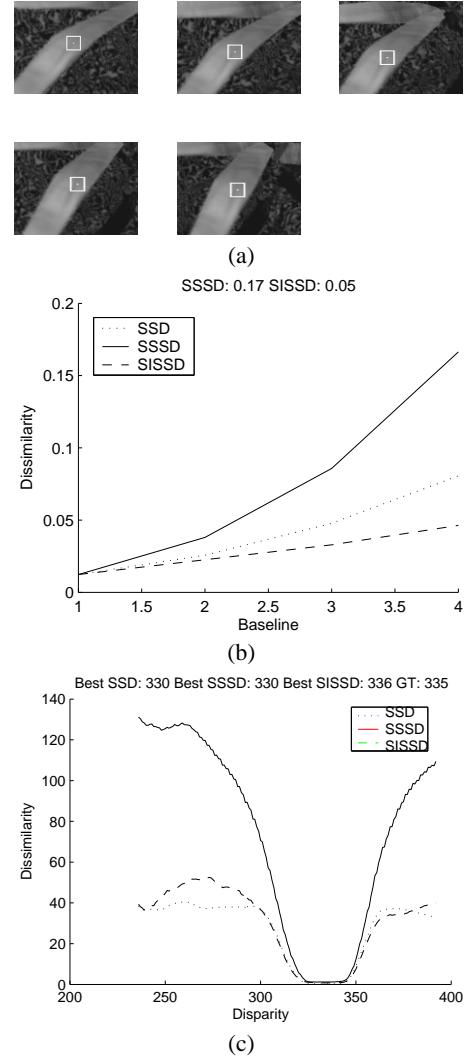


Figure 1: The case of steep leaves where projection changes orientation across the baseline. (a) five views of the location on the steep leaf. (b) The development of the dissimilarity across the baseline. (c) The dissimilarity/energy function across the scan line in the image. The best match for SSD, SSSD, SISSD ($\alpha = 1$), and ground truth (GT) is given over the graph.

the second switches the disparity to the y-axis. Their baselines are equal to the largest multi baseline (Image N).

$$T_m(x, y, d) = \arg \min_d \min(E_{1,N_x}(x, y, d), E_{1,N_y}(y, x, d)) \quad (4)$$

$$T_s(x, y, d) = \arg \min_d (E_{1,N_x}(x, y, d) + E_{1,N_y}(y, x, d)) \quad (5)$$

In theory the T_m should comparably be more robust to occlusions by choosing the best match in a single image pair. T_s should comparably be more certain of a match if the point is visible in all cameras by choosing the best match where both image pairs are good matches.

One of the best 3D reconstruction algorithms available uses a graph cut energy minimization, which yields similar results to the slower simulated annealing. The difference is that graph cuts preserves depth discontinuity (Kolmogorov and Zabih, 2002). It does not rely on window sizes which tend to dilate the depth regions and are sensitive to perspective distortion. The main adjustable parameter is the impact of the smoothness constraint, λ .

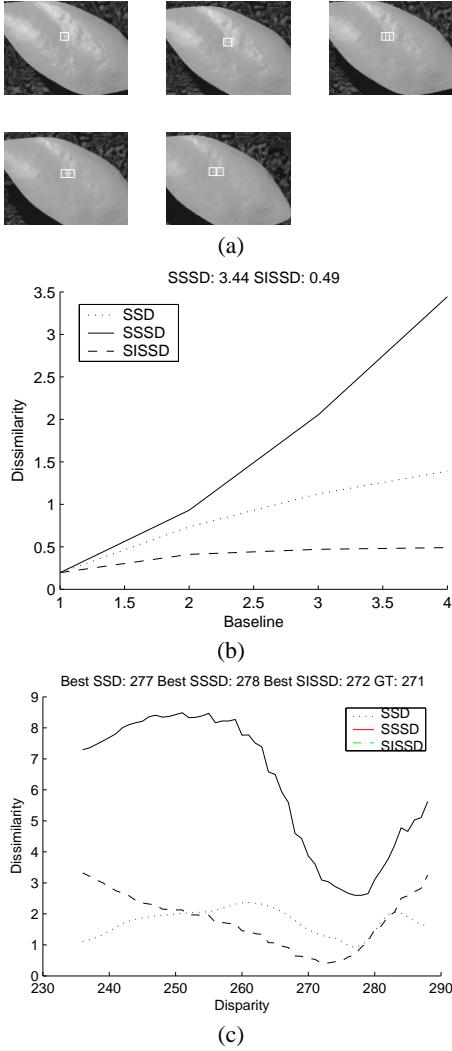


Figure 2: The case of flat leaves where the highlight changing across the baseline. The potential weakness of SISSD is that the dissimilarity difference between the correct match and its surroundings is not very pronounced. This makes the global minimum sensitive to jitter.

Since it assumes regions of equal depth, it excels at fronto-planar scenes, but may have trouble when it comes to steep leaves on plant structures. It was interesting to see how it performed in this new context. We used Kolmogorov's implementation of the graph cut algorithm (Kolmogorov and Zabih, 2002) that is referred to as *kz1*. This is only a binocular algorithm which used the 1st and the *N*th camera. λ was given a small value (half of the automatic setting).

There are three common quality metrics root-mean-square, reprojection/prediction of a novel view(Szeliski and Zabih, 1999), and percentage of bad matching pixels. The latter is chosen because the focus is to generate correct disparity maps. Root-mean-square error does not ensure that the structure and discontinuities are preserved. Reprojection error does not measure the actual disparity error, but whether the reprojeciton of one green pixel happen to hit a matching green pixel in the novel view. However, in a scene full of green plants that is very likely even if the disparity is very wrong.

The estimated disparity maps d_E were compared to ground truth (d_{GT}) using the Percentage of Bad Matching Pixels metrics as in

(Scharstein and Szeliski, 2002):

$$PBMP = \frac{1}{N} \sum_{(x,y)} |d_E(x,y) - d_{GT}(x,y)| > \delta \quad (6)$$

2.3 Experimental Setup

The experimental tests were conducted in order to learn more about the algorithms in the complex context of close-up reconstruction of complex structures. Hence, near-photo realistic ray traced scenes of plants were used in order to control the scene parameters and get valid ground truth disparity maps, occlusion masks, and highlight masks. The scenes had natural outdoor lighting and focal blur, which is a natural problem with plants with steep leaves. Blur is unavoidable, because the aperture cannot be very small and the shutter must be fast when capturing images from a moving platform and the plants are waving in the wind.

Two main classes of plants, long leaf (grass-like, e.g. cereal) and broad leaf (e.g. beet and tomato) were generated. This relates to *surface shape*. For each of these there were plants with steep leaves and flat leaves, respectively. This relates to *surface orientation*. Steep leaves compared to flat leaves have less highlight, more occlusion, and vice versa. A natural case with two grassy plants with flat and steep leaves and a lot of occlusion were used, too. Each scene was generated with textured (spotted) and no texture (glossy), both having bump maps. This relates to *presence of texture*. Finally, all images were generated with and without specularity. This served two purposes; 1. it was required to find the highlight masks (where highlights exist in one frame and not the other), and 2. in order to test overall performance of the algorithms and the same geometrical structure with and without the presence of highlights. There were 18 image sets in total. See figure 3 for an example with ground truth.



Figure 3: A natural case, where two grass-like plants are close together and leaves are occluded. The proportion of occluded pixels is 5% and the proportion of changing highlights are 5%.

3 RESULTS AND DISCUSSION

The overall results are shown in table 1. It is the mean and spread of performance over all plant types. Note that the ground truth maps were calculated in floating points as to represent the (scaled) inverse of the real height. The disparity maps were integer pixels. If the ground truth had been rounded, the values would have been 10-20% lower. *Multi_{3cam}* used the same cameras as *Multi_{5cam}*, but skipped camera 2 and 4.

The table shows that having those two extra cameras in between the three cameras did improve the result by 11% in average for all pixels, 8% for highlighted pixels, and 8% for occluded pixels. Meanwhile, their spread was approximately equal or slightly narrower (for occluded pixels). The significance of 8.9% versus 8.2% is up to the application to decide. The development within

Table 1: Comparison of Stereo setups. Mean PBMP (%) and their standard deviations calculated from all pixels (all), pixels with different specularity state (high), and occluded pixels (occ).

Stereo Setup	All	High	Occ
<i>Multi₃SSSD</i>	8.9(5.9)	22.1(14.6)	50.3(30.9)
<i>Multi₃α0.25</i>	8.9(5.6)	20.9(13.7)	55.4(28.7)
<i>Multi₃α0.50</i>	9.9(5.6)	20.6(12.1)	64.6(23.8)
<i>Multi₃α0.75</i>	13.5(6.6)	23.0(12.2)	69.1(24.0)
<i>Multi₅SSSD</i>	8.3(5.5)	20.3(13.9)	46.1(28.3)
<i>Multi₅α0.25</i>	8.2(5.3)	19.4(13.3)	49.9(24.5)
<i>Multi₅α0.50</i>	8.8(5.4)	19.1(12.7)	55.3(22.5)
<i>Multi₅α0.75</i>	11.6(6.0)	21.0(12.5)	69.1(20.4)
<i>GraphCut</i>	14.6(8.7)	19.6(16.3)	73.9(24.3)
<i>TrinoMin</i>	10.2(6.5)	23.1(12.5)	30.6(22.3)
<i>TrinoSum</i>	9.8(6.9)	23.6(15.8)	40.3(25.0)

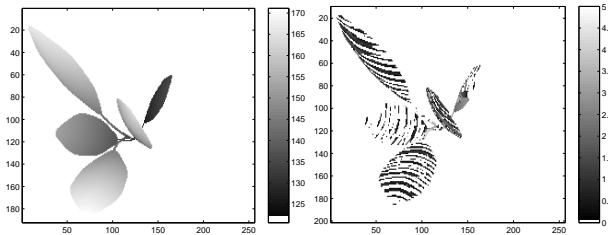


Figure 4: [Left] Ground truth and [Right] Graph Cuts Log(disparity error) for steep spotted broad leaf without highlights. The banding characteristics were caused by the attempt to impose fronto-planar regions on the steep leaves.

multi₅ by increasing α was devastating for occluded pixels by 50%, while overall and highlight pixels reach a local minima between $\alpha = 0.25$ and $\alpha = 0.5$. The benefit was rather small, though; 1% for all pixels and 5% for highlight pixels. The SISSD measure may be a improvement when using larger window sizes, which tend to be the case when using real images. The trinocular measures did well and they excel at occluded pixels, especially T_m . Graph cuts did the worst, except at correcting highlight pixels by smoothing those areas. Figure 4 shows why graph cuts did not do very well. The disparity map was banded, i.e. staircase shaped, instead of smooth.

Figure 5 shows the errors from the multi-baseline reconstruction of the same plant. The errors were more recognizable as noisy jitter, which could be removed by an energy minimizing sloped smooth surface technique.

Figure 6 shows the errors from trinocular results for the same plant. The very steep leaf in the middle and the one to the right of it are difficult for all the algorithms except trinocular minimum (T_m). It is so steep that it is almost a self-occlusion. In the second camera the leaf would be extended along orientation of the baseline, thus occluding the other leaf. T_m simply reconstructed it from the Y direction. The lesson is that it is not only the orientation toward the camera that affects the result, but if the orientation of a leaf aligns with the baseline it can be difficult to reconstruct it. This is especially a problem with textureless grass-like leaves that aligns with the baseline (Nielsen et al., 2004). In comparison, SISSD was able to reconstruct the steep leaf nearly as good, but the leaf to the right of it was as bad as Trinocular sum (T_s).

Figure 7 plots the all-pixel results grouped by descriptive object parameters, i.e. leaf shape, leaf orientation (flat or steep leaves), texture, and highlights and occlusion. Horizontal axis is the setup: M 0.0 (SSSD), M 0.25 (SISSD $\alpha = 0.25$), M 0.5, M 0.75, Binocular Graph Cut, Trinocular Minimum T_m , and Trinocular sum T_s . The vertical axis is the mean pbmp for window sizes

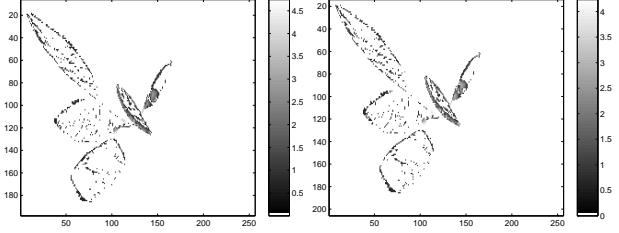


Figure 5: [Left] Log(disparity error) Multi-baseline SSSD and [Right] SISSD $\alpha = 0.5$. These results did not have any banding, but the difference between the SSSD and SISSD was very small. The result would be excellent if it were combined with a slope-and discontinuity preserving graph cut minimization.

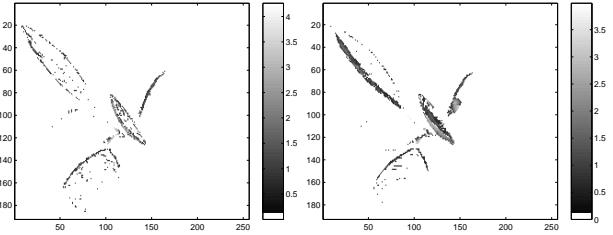


Figure 6: [Left] Log(disparity error) trinocular minimum (T_m) and [Right] trinocular sum (T_s).

ranging from 4-12. The same goes for figures 8 and 9 that show the pbmp of highlight pixels and occlusion pixels, respectively.

Figure 7 plot (a)(plants without specular highlights) clearly pins down the sources of error for reconstruction in general. The flat-leaved plants (since they had no specular highlights on this plot) all score very well. The errors were large when the leaves were steep or occluding (the model called *two grassy* is 5% occluded in comparison to the steep broad leaf which is only 1%).

The interesting aspect on plot (a) on figure 7 is that it was the steep leaves that best improved slightly from SISSD, while the flat leaves are reconstructed best through SSSD. However, taking a look at plot (b) reveals that when there were highlight on those flat leaves, SISSD was an improvement, too, especially for broad leaf plants.

Note also the fact that the steep leaves were troublesome for graph cuts on plot (a) and (c), especially the glossy steep broad leaf, which was easier for the others compared to grassy plants. Plot (a) to (d) shows consistently that T_s reconstructed grass-like plants better than T_m , but T_m reconstructed broad leaf plant best. This trend is revisited in figure 8.

Figure 7 Plot (d) shows that in the more natural case, SSSD and T_s were best, even though T_m was best in most occluded parts (figure 9 plot (a) and (b)). Maybe the algorithm could dynamically choose T_m by detecting occlusion with left-right consistency (Fusiello et al., 2000).

Figure 8 plot (a) and (b) shows the subtle strength of SISSD in the highlighted areas. The flat glossy broad leaf was the most difficult to reconstruct. Note that this is the plant type that was 50% highlighted, and there were no texture other than shading and bumps to correlate. The graph cut algorithm were particularly bad in this case, because it created non existant surfaces in over the plant from the errors of the highlights.

4 CONCLUSIONS

The relationship between the performances of the algorithms and the descriptive parameters of the plant objects were investigated. A new multi-baseline Sum of Squared Difference based correlation was defined (SISSD) in order to minimize the effect of perspective distortion within the windows. The results showed that there was a relationship between the performance and the descriptive parameters of the objects. However, SISSD was only a marginal improvement on images with steep leaves (slopes), but more so in the presence of highlights. It was mainly an improvement at the actual highlight areas, especially on shiny broad leaf plants. On the other hand SSSD was better at matching the occluded areas. The best algorithm for occluded areas was the trinocular T_m algorithm. Binocular Graph cuts were not able to reconstruct the slopes in steep leaves, but the smoothness optimization seemed to smoothen over the errors from highlights, when the highlight areas were not too large. The results showed a complicated relationship of trade-offs that points toward further development combining the strengths of the individual configurations.

4.1 Perspectives on future work

An improvement to the SISSD measure could be to have α depend on the distance from reference image. Another interesting aspect would be to place the 5 cameras in a trinocular setup. The five cameras would then complete two systems of three-camera multi-baseline systems in each direction.

Furthermore, a multi-baseline or trinocular algorithm in combination with graph cuts would be interesting to pursue, and to improve its ability to reconstruct steep slopes. There are other works on these aspects to pay special attention to (Buehler et al., 2002)(Lin and Tomasi, 2004). Buehler's trinocular algorithm does not handle the situation where occlusion only exist in one camera pair. This was the strength of the trinocular minimum algorithm in this paper. Lin and Tomasi's algorithm for sloped surfaces relies too strongly on large smooth surfaces. This may be a problem for natural leaves that can be curled and there might only be small segments showing of each leaf, while the surface boundaries are only vaguely defined by intensity edges (sometimes not at all).

The final step is to create a mesh that is able to treat intertwining and overlapping leaves as individual surfaces.

4.2 Acknowledgments

The authors would like to thank the Danish Technical Research Council, the Danish Agricultural and Veterinary Research Council and the Danish Ministry of Food, Agriculture and Fisheries for funding.

REFERENCES

- Buehler, C., Gortler, S. J., Cohen, M. F. and McMillan, L., 2002. Minimal surfaces for stereo. In: ECCV (3), pp. 885–899.
- Christensen, L. K. and Jørgensen, R. N., 2003. Spatial reflectance at sub-leaf scale discriminating NPK stress characteristics in barley using multiway regression (N-PLS). In: 2003 ASAE Annual International Meeting, Las Vegas, Nevada, USA, July 27-30, Paper no. 031138.
- Fusiello, A., Roberto, V. and Verri, A., 2000. Symmetric stereo with multiple windowing. International Journal of Pattern recognition and Artificial Intelligence 14(8), pp. 1053–1066.
- Jeon, J., Kim, K., Kim, C. and Ho, Y., 2001. Robust stereo matching algorithm using multiple-baseline cameras. In: IEEE Pacific Rim Conference on Communications, Computers and signal Processing, Vol. 1, pp. 263–266.
- Kolmogorov, V. and Zabih, R., 2002. Multi-camera scene reconstruction via graph cuts. In: IEEE European Conference on Computer Vision.
- Lee, W. S., Slaughter, D. C. and Giles., D. K., 1996. Development of a machine vision system for weed control using precision chemical application. In: International Conference on Agricultural Machinery Engineering '96, Seoul, Korea., pp. 802–811.
- Li, Y., Lin, S., Lu, H., Kang, S. and Shum, H.-Y., 2002. Multi-baseline stereo in the presence of specular reflections. In: IEEE Intl Conf. on Pattern Recognition, Vol. 3, pp. 573–576.
- Lin, M. and Tomasi, C., 2004. Surfaces with occlusions from layered stereo. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 26(8), pp. 1073–1078.
- Mulligan, J. and Daniilidis, K., 2002. Trinocular stereo: A real-time algorithm and its evaluation. International Journal for Computer Vision 47, pp. 51–61.
- Nielsen, M., Andersen, H. J., Slaughter, D. C. and Granum, E., 2005. Ground truth evaluation of 3d computer vision on non-rigid biological structures. In: J. Stafford (ed.), Precision Agriculture 05, Wageningen Academic Publishers, The Netherlands, pp. 549–556.
- Nielsen, M., Christensen, L. K. and Andersen, H. J., 2004. Sub-leaf scale remote sensor for npk discrimination using stereo vision. In: Engineering the Future, International Conference on Agricultural Engineering, Leuven, Belgium, 12-14 September, Session 10, no. 327.
- Okutomi, M. and Kanade, T., 1993. A multi-baseline stereo. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15(4), p. 353363.
- Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47(1/2/3), pp. 7–42.
- Scharstein, D. and Szeliski, R., 2003. High-accuracy stereo depth maps using structured light. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), Madison, WI, USA, pp. 195–202.
- Szeliski, R. and Zabih, R., 1999. An experimental comparison of stereo algorithms. In: Springer, International Workshop on Vision Algorithms, Corfu, Greece, pp. 1–19.

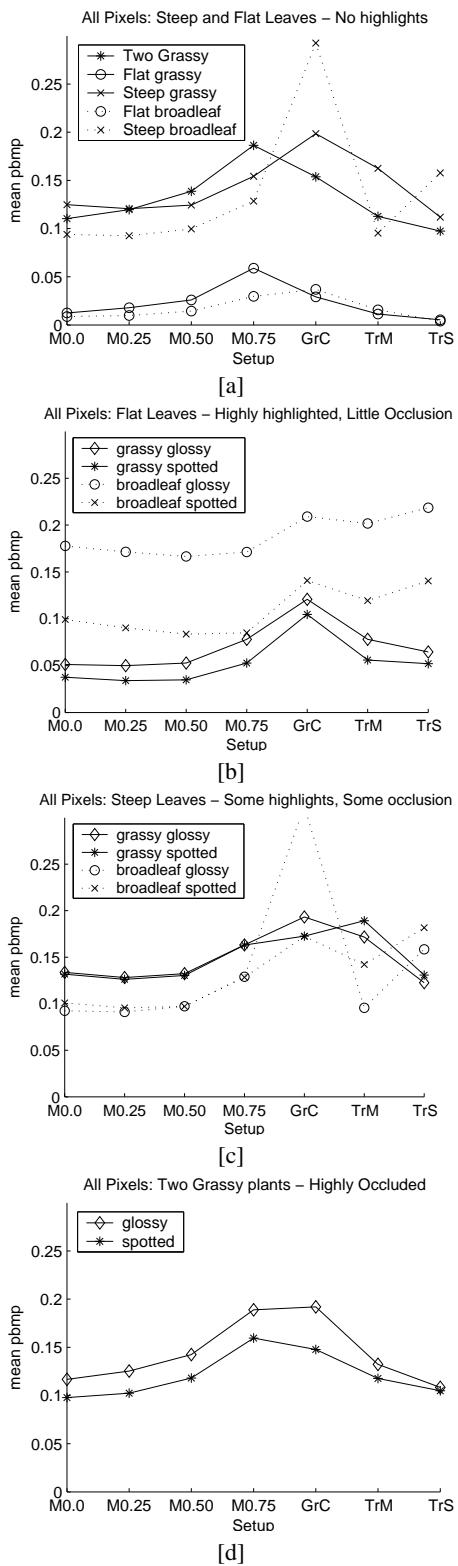


Figure 7: PBMP from all pixels results by object type and leaf orientation. The worst case occlusion is the *Two Grassy Plants* model being 5% occluded. The worst case of highlights were the flat grass-like and flat broad-leaf. 20% of their area suffered from changing specular highlights.

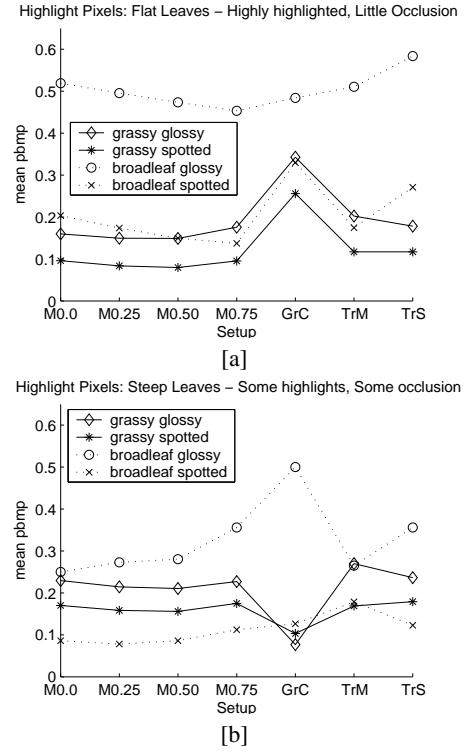


Figure 8: PBMP from specular changing highlight pixels results by object type and leaf orientation. SISSD (M0.25-M0.75) improves performance.

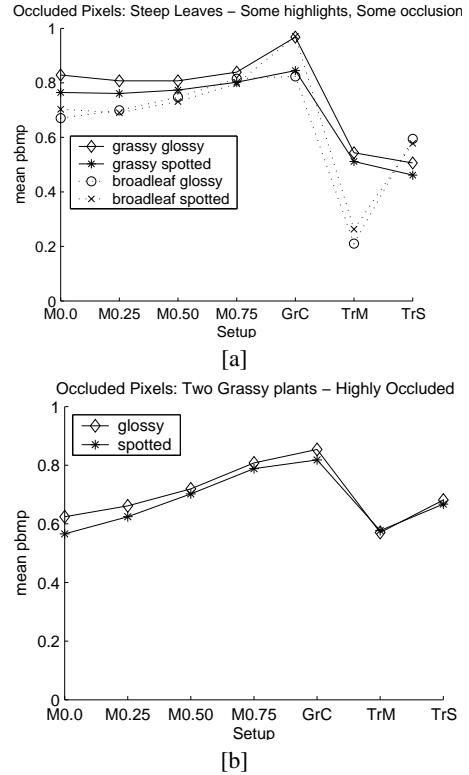


Figure 9: PBMP from occluded pixels results by object type. Trinocular minimum T_m is the best algorithm for occluded areas.