A FOREGROUND OBJECT BASED QUANTITATIVE ASSESSMENT OF DENSE STEREO APPROACHES FOR USE IN AUTOMOTIVE ENVIRONMENTS

Oliver K. Hamilton, Toby P. Breckon

School of Engineering
Cranfield University
Bedfordshire, United Kingdom

Xuejiao Bai, Sei-ichiro Kamata

Grad. Sch. of Inf., Production & Syst.
Waseda University
Kitakyushu, Japan

ABSTRACT

There has been significant recent interest in stereo correspondence algorithms for use in the urban automotive environment [1, 2, 3]. In this paper we evaluate a range of dense stereo algorithms, using a unique evaluation criterion which provides quantitative analysis of accuracy against range, based on ground truth 3D annotated object information. The results show that while some algorithms provide greater scene coverage, we see little differentiation in accuracy over short ranges, while the converse is shown over longer ranges. Within our long range accuracy analysis we see a distinct separation of relative algorithm performance. This study extends prior work on dense stereo evaluation of Block Matching (BM)[4], Semi-Global Block Matching (SGBM)[5], No Maximal Disparity (NoMD)[6], Cross[7], Adaptive Dynamic Programming (AdptDP)[8], Efficient Large Scale (ELAS)[9], Minimum Spanning Forest (MSF)[10] and Non-Local Aggregation (NLA)[11] using a novel quantitative metric relative to object range.

Index Terms— Stereo Vision, Registration, Disparity, Quantitative Assessment.

1. INTRODUCTION

Stereo vision for use in autonomous systems and driver assistance systems is an ever increasing research topic [1, 2, 3]. Previous analysis of stereo correspondence algorithms have focused on static scene tests [12], qualitative comparative analysis [2] and explicit global comparison against ground truth data within the automotive environment [3].

Dense stereo vision essentially recovers pixel wise 3D depth information for a given left-right stereo pair. There is an ever growing pool of dense stereo algorithms [4, 5, 6, 7, 8, 9, 11] of which most can be summarized by four primary steps: pixel matching, matching cost aggregation, disparity measurement and post processing. Stereo correspondence algorithms generally fall into one of two groups, local or global algorithms. Local algorithms use a local support window to compare pixel regions selecting a locally minimal matching cost [4]. By contrast, global approaches compute disparities by using disparity estimates over the entire image by relying on energy minimization techniques, such as graph cuts [13, 12] and dynamic programming [12, 14, 8]. Here, we aim to take a unique look at the real world accuracy of a selection of stereo correspondence algorithms, (BM[4], SGBM[5], NoMD[6], Cross[7], AdptDP[8], ELAS[9], MSF[10] and NLA[11]) as a function of range. Previous comparative studies do not perform such quantitative analysis [1, 2, 3]. Furthermore they use a global scene evaluation, based on correlation against *a priori* ground truth [1, 2], which is inherently bias against foreground objects of interest [3]. Leveraging the recent availability of annotated ground truth data [1] we propose a registration based methodology explicitly analysing the depth accuracy against range on foreground objects. Within the automotive application of stereo vision such foreground objects (e.g. cars, trucks, cyclists and pedestrians) are of primary importance. Despite this, the evaluation on such objects is lacking in prior studies [1, 2, 3]. By

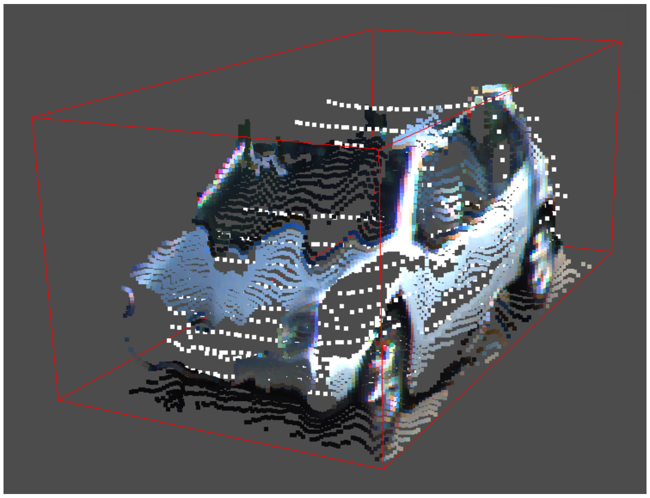

Fig. 1. Visualisation of colour mapped disparity point cloud and laser ground truth point cloud (white points).

This work is supported by TRW Coneckt and EPSRC under case agreement 11330161 (awarded by the UK ICT KTN).

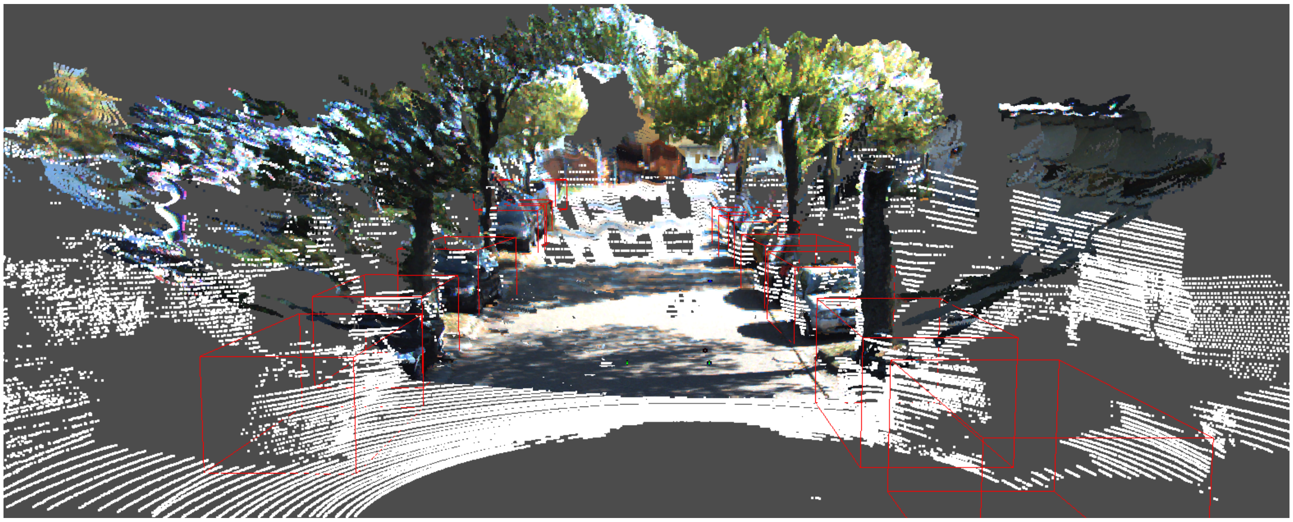

Fig. 2. Colour disparity point cloud generated using [9] with laser scanner point cloud (white dots) and annotated ground truth bounding boxes (red boxes).

contrast, here we present a novel object-wise quantitative assessment of dense stereo accuracy as a function of range, specifically targeting foreground object accuracy to offer new insight into relative algorithm performance.

2. QUANTITATIVE ASSESSMENT METHODOLOGY

Our quantitative comparison methodology has four stages based on:- 1) disparity map generation from dense stereo technique of choice, 2) subsequent stereo point cloud generation via 3D triangulation, 3) foreground object segmentation, 4) registration between segmented stereo point cloud and ground truth. After the disparity map has been generated from a given stereo correspondence algorithm, each pixel is triangulated to real world coordinates (Eqn.1) and added to a disparity point cloud. The corresponding ground truth data, supplied from laser scanner, is transformed into the left camera coordinate system of the stereo set up using the supplied calibration information [1]. This stereo disparity map to point cloud conversion is carried out using the following transformation :-

$$X = \frac{Z(u - cu)}{f} \qquad Y = \frac{Z(v - cv)}{f} \qquad Z = \frac{fB}{d} \qquad (1)$$

where f = focal length (pixels), B = baseline (mm), d = pixel disparity (pixels), $[u, v]$ = disparity map pixel x and y positions respectively (pixels), $[cu, cv]$ = image centre along the optical axis, $[X, Y, Z]$ = real world coordinates in camera reference frame (mm).

A hypothetically perfect dense stereo algorithm with error free measurements of disparity, d, will produce a triangulated 3D depth estimate matching exactly to point $P = (X, Y, Z)$ (i.e. identical to the ground truth). In practice, due to disparity errors in the stereo algorithms this 3D point estimate is imperfect.

By differentiating Eqn.1 w.r.t disparity, d, to recover the derivative of range against disparity we get the following :-

$$\frac{\delta Z}{\delta d} = -fBd^{-2} \qquad (2)$$

We rearrange Eqn.1 and Eqn.2 to form the following two relationships :-

$$\Delta Z = -\frac{fB}{d^2}\Delta d \qquad (3)$$

$$d^2 = \frac{f^2B^2}{Z^2} \qquad (4)$$

By further substituting Eqn.4 into Eqn.3 we recover the range error, ΔZ, as a function of object of interest range, Z, at a fixed disparity error, Δd, Eqn.5.

$$\Delta Z = -\frac{Z^2}{fB}\Delta d \qquad (5)$$

applying this to Eqn.1 yields :-

$$X' = \frac{Z'(u - cu)}{f} \qquad Y' = \frac{Z'(v - cv)}{f} \qquad Z' = Z + \Delta Z \quad (6)$$

We now formulate point $P' = (X', Y', Z')$ as being the actual stereo triangulated point estimate from a disparity map created with a disparity error of Δd. Our assumption is that as the stereo data and ground truth data are registered, we can calculate any difference in stereo disparity to ground truth data as the Euclidean distance between P and P', This can be recovered via Iterative Closest Point (ICP) registration [15]

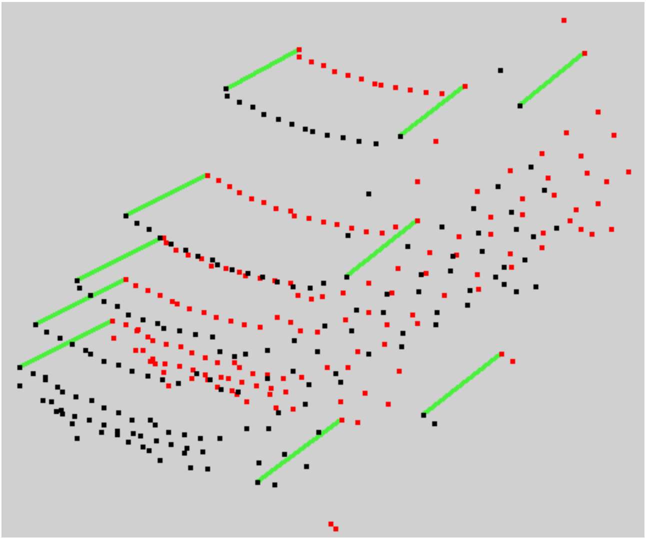

Fig. 3. Raw ground truth data for a car at a range of $25m$ (black points). Ground truth data post ICP registration with disparity point cloud (red points).

between the data and ground truth point clouds of an isolated foreground object.

In Figure 2 we can see the combination of the ground truth data laser scanned point cloud (white points), the point cloud obtained from a given dense stereo approach (coloured points) and the bounding box annotation supplied with the dataset [1] for the car objects (red lines).

From the ground truth annotation (Figure 1) we isolate foreground objects of interest (i.e. cars) in both the disparity point cloud and the laser scanned ground truth point cloud. The bounding boxes are expanded by a fixed ΔD, $(0.25m)$, where the edges of the boxes are defined as $x_{min} = x_{centroid} - width/2 - \Delta D$ and $x_{max} = x_{centroid} + width/2 + \Delta D$ and similarly applied to y_{min} and y_{max}. This is to compensate for any poorly triangulated point positions and ensure they are captured within the bounding box of the relevant object. Maximising the number of points that belong to an isolated object will increase the ICP registration performance.

From the two resulting object of interest point clouds, extracted from stereo disparity and ground truth laser scanner, we perform registration using the Iterative Closest Point [15] approach. This facilitates the recovery of the transformation between the two point clouds. Post registration, we can identify the Euclidean distance offset between the two point clouds. This provides us, via ICP, with a global accuracy metric of the stereo correspondence algorithm for the reproduction of the ground truth 3D information for a given object. Furthermore, as each object of interest is extracted at a known depth from the camera, Z, based on the ground truth annotation we additionally have an accuracy metric relative to object range, (Figure 2). Figure 3 shows an isolated point cloud from the laser scanned ground truth data as the black points and the registered version as the red points. Lines illustrate the offset between point clouds which we denote as displacement error. Performing this over the test sequence yields numerous such object-wise accuracy measurements over a range of distances (range, Z) and angle to target (see Figure 2). The results over the test sequence are presented in Figure 5 (a/b).

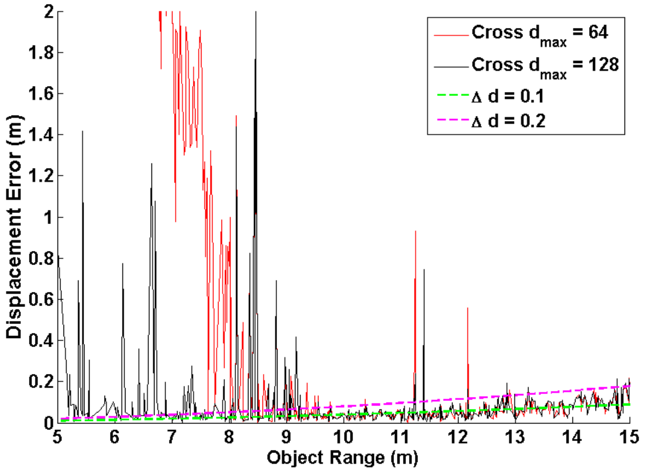

Fig. 4. Short range accuracy difference between disparity parameter settings.

3. RESULTS

This study uses the data set available from [1] which is provided with ground truth 3D laser scan data and annotated bounding box information for cars, trucks, pedestrians and cyclists. Specifically the analysis was carried out on the data set entitled '2011_09_26_drive_0009' containing 89 cars over a sequence of 447 rectified images subsampled at approximately 0.1s intervals. We model this expected drop off in accuracy over range of a given stereo correspondence algorithm (i.e. Eqn. 5) and plot the theoretical range error for two hypothetical disparity errors, $\Delta d = 0.1, 0.2$ (see Figure 5(a/b)). This gives us an base-line expectation of degradation of displacement error against range against which to compare (assuming a constant disparity error over range).

The displacement error, obtained via ICP registration, is plotted for every visible car in every frame in metres and collated for 8 algorithms (Figure 5(a/b)). Here we test BM [4], SGBM [5], NoMD [6], Cross [7], AdptDP [8], ELAS [9], MSF [10] and NLA [11]. From Figure 5(a/b) we can see a distinct difference in performance in relation to the range of the object from the camera. At short ranges it is apparent they can perform very differently. This can be attributed to the parameters used when tuning the algorithms (e.g. setting BM or SGBM to a limited disparity search of 64 pixels artificially limits the minimum effective range, hence the dramatic drop off in accuracy). Increasing this parameter increases the accuracy (decreases the displacement error) performance at short

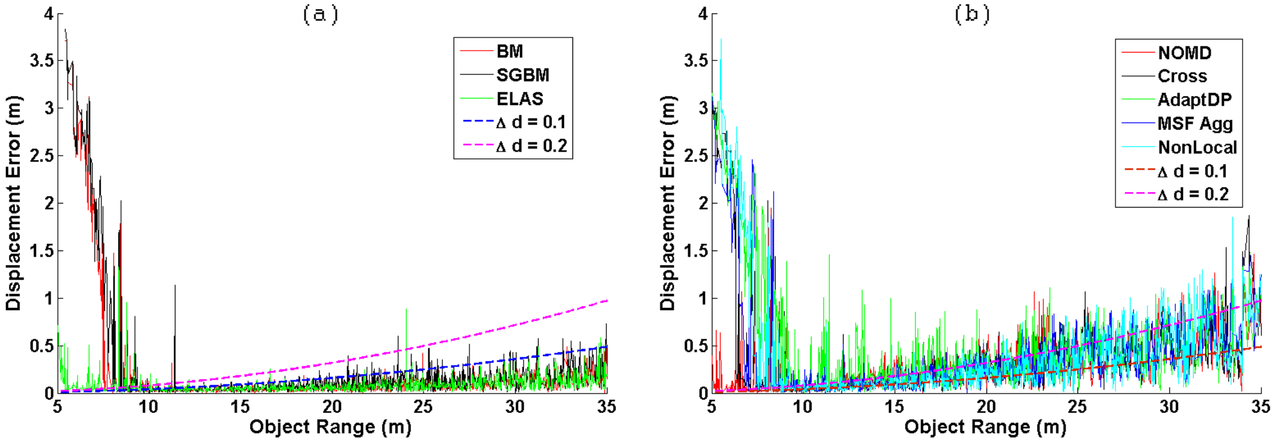

Fig. 5. (a) Most accurate algorithms, (b) least accurate algorithms in this study (Displacement Error (m) versus Object Range (m)).

ranges (see Figure 4).

Algorithms not limited to a disparity search range naturally have improved close range performance (e.g. ELAS and NoMD see Figure 5(a/b)). The rapid increase in the displacement error at small ranges is due to the limitations of ground truth data collection. As the foreground vehicle comes within the minimum range of the laser scanner and the edge of the field of view of the cameras the two point clouds become incomplete. This reduces the registration constraints for the ICP registration causing artificially high registration displacement errors, (see Figure 2).

Figure 5(a) shows the displacement error for the algorithms that performed best in this study. We see that BM, SGBM and ELAS all perform with a $\Delta d < 0.1$. The collated displacement errors for NoMD, Cross, AdaptDP, MSF and NLA are seen in Figure 5(b), we can see they have a matching accuracy of $\Delta d \approx 0.15$.

Due to angular resolution of the laser scanner, objects beyond $\sim 35m$ in range have fewer points for the ICP method to match against. While errors of up to $1.5m$ at a range of $\approx 35m$ may not sound significant, (see Figure 5(b)) a vehicle travelling at the UK motorway speed limit of 70mph ($\approx 31ms^{-1}$) will cover that distance in a little over $1.1s$. Furthermore the typical width of a car within the data set is $\approx 1.6m$ which in itself is close to the magnitude of this error. While in general this does however become significant for real time obstacle avoidance [16].

Figure 5(a) shows the top performing algorithms in this study (BM, SGBM, ELAS). These three algorithms maintain a low displacement error throughout the test ranges indicating a low matching error, Δd. The greater the matching error the poorer the depth estimation and the greater the resulting displacement error, as seen in Figure 5(b). Interestingly BM and SGBM both perform as well as ELAS, only requiring that the maximal disparity search range is increased to cope with close range parts of the scene.

Despite these insights, some notable limitations of this methodology relate to the poor reliability of displacement distance at ranges greater than $\sim 35m$. This is somewhat due to the angular resolution of both the laser scanner and the disparity images, which results in fewer points being used to construct an object of interest at significant ranges, Z.

Overall from Figure 5(a) we can see that BM, SGBM and ELAS perform well over a large range whilst Cross, NoMD, AdaptDP, MSF and NLA are notably seen to perform less well at ranges greater than $\sim 15m$ (Figure 5(b)). The results demonstrate the effectiveness of ELAS by maintaining a low object wise displacement error throughout the test range whilst not itself being inherently limited to a maximum disparity search range (unlike others, e.g. BM, SGBM).

Whilst the limitation of ground truth quality at significant range mildly affects the methodology we still see a clear quantitative insight into relative algorithm performance.

4. CONCLUSION

We present a novel quantitative evaluation approach for dense stereo algorithms considering object-wise foreground accuracy in relation to range. This is used to compare a range of such approaches [4, 5, 6, 7, 8, 9, 10, 11] providing novel insight into complex urban environment performance. From our sample of algorithms, we conclude that ELAS [9] performs with the greatest foreground object accuracy throughout the ranges examined in this study.

Future work will look to increase the breadth of this study in terms of the number of dense stereo algorithms and scene objects considered. Furthermore a comparison of the object wise displacement error against disparity computation time would be useful with regard to the future generation of an efficiency versus accuracy metric.

5. REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Comp. Vision and Pattern Recog.*, 2012, pp. 3354 –3361.

[2] F. Mroz and T.P. Breckon, "An empirical comparison of real-time dense stereo approaches for use in the automotive environment," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 13, pp. 1–19, 2012.

[3] R. Haeusler and R. Klette, "Analysis of KITTI data for stereo analysis with stereo confidence measures," in *Computer Vision – ECCV 2012. Workshops*, pp. 158–167.

[4] K. Konolige, "Small vision system. hardware and implementation," in *International Symposium on Robotics Research*, 1997, pp. 111–116.

[5] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. on Patt. Anal. and Mach. Intel.*, vol. 30, no. 2, pp. 328–341, 2008.

[6] C. Unger, S. Benhimane, E. Wahl, and N. Navab, "Efficient disparity computation without maximum disparity for real-time stereo vision," in *British Machine Vision Conf.*, 2009, pp. 1–12.

[7] J. Lu, K. Zhang, G. Lafruit, and F. Catthoor, "Real-time stereo matching: A cross-based local approach," in *IEEE Int. Conf. on Acou., Speech and Sig. Proc.*, 2009, pp. 733–736.

[8] Wang L., Liao M., Gong M., Yang R., and Nister D., "High-quality real-time stereo using adaptive cost aggregation and dynamic programming," in *Third Int Symp on 3D Data Proc, Vis and Trans*, 2007, pp. 798–805.

[9] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," Asian Conf. on Comp. Vision, 2010, pp. 25–38.

[10] CVPR Paper ID 378, "Accurate stereo matching via an aggregation strategy based on a minimum spanning forest," in *Proc. Comp. Vision and Pattern Recog.*, 2013.

[11] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Comp. Vision and Pattern Recog.*, 2012, pp. 1402–1409.

[12] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. of Computer Vision*, vol. 47(1), pp. 7–42, 2002.

[13] L. Cheng, J. Selzer, and Y.H. Yang, "Region-tree based stereo using dynamic programming optimization," in *Proc. Comp. Vision and Pattern Recog.*, 2006, pp. 2378 – 2385.

[14] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. Int. Conf. Computer Vision*, 2001, pp. 508–515.

[15] P.J. Besl and N.D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. on Patt. Anal. and Mach. Intel.*, pp. 239–256, 1992.

[16] I. Katramados, S. Crumpler, and T.P. Breckon, "Real-time traversable surface detection by colour space fusion and temporal analysis," in *Proc. Int. Conf. on Computer Vision Sys.*, 2009, pp. 265–274.