GENERALIZED DYNAMIC OBJECT REMOVAL FOR DENSE STEREO VISION BASED SCENE MAPPING USING SYNTHESISED OPTICAL FLOW

Oliver K. Hamilton, Toby P. Breckon

School of Engineering and Computing Sciences, Durham University, Durham, UK

ABSTRACT

Mapping an ever changing urban environment is a challenging task as we are generally interested in mapping the static scene and not the dynamic objects, such as cars and people. We propose a novel approach to the problem of dynamic object removal within stereo based scene mapping that is both independent of the underlying stereo approach in use and applicable to varying object and camera motion. By leveraging stereo odometry, to recover camera motion in scene space, and stereo disparity, to recover synthesised optic flow over the same pixel space, we isolate regions of inconsistency in depth and image intensity. This allows us to illustrate robust dynamic object removal within the stereo mapping sequence. We show results covering objects with a range of motion dynamics and sizes of those typically observed in an urban environment.

Index Terms— Stereo Vision, Disparity, Object Removal.

1. INTRODUCTION

Vision based mapping is an active area of research with a wide variety of applications [1, 2]. The most common techniques include structure-from-motion (SFM) [3] and dense stereo [4, 5, 6, 7]. SFM techniques rely on temporally varying image samples which naturally rejects dynamic objects as they do not share the same motion as the observed static background. By contrast, reconstruction from dense stereo is performed on synchronised stereo images therefore there is no temporal discrimination and dynamic objects can not be identified easily. This work provides a generalized method of dynamic object rejection for a stereo vision system, mounted on a moving platform, that is independent of object class and assumes no prior information about object motion characteristics or size.

1.1. Related Work

Prior work addresses the problem in terms of identifying the location and approximate region occupied by a moving object [8, 9, 10, 11]. This is effective for use in autonomous vehicles to aid with obstacle avoidance, however for use in mapping applications the segmented region often fails to completely remove only the moving object [11]. The same limitation is found with detection driven techniques to identify all objects that could be in motion (e.g. car, trucks, bicycles and pedestrians [12]). However such detection based approaches are limited by their generality (in detecting people, vehicles, bicycles, animals, horse drawn vehicles, prams... etc.). The stereo mapping work of [10, 11] uses feature tracking and depth samples to detect dynamic scene elements. Whilst this proves effective in detecting candidate regions, the sparse nature of the feature points means further analysis of the original intensity imagery is required to correctly segment the object increasing both complexity and processing load. Accurate segmentation on intensity images is heavily dependent on lighting variations, shadows and reflections [11]. Results from[13] demonstrate the use of stereo vision for moving object detection. However it is limited to use within a block based stereo matching technique for moving candidate confirmation. Whilst such block based approaches have been shown to perform comparatively to contemporary approaches [14], this dependency limits the wider applicability of [13] to a small subset of stereo algorithms in general [7]. Furthermore the moving object mask produced from [13] is sparse in nature and is insufficient to effectively remove the object from a dense disparity map. Work by [15] demonstrates moving object masks that are more dense than [13], however still require a separate optical flow calculation in intensity space.

It is clear that a limitation exists within the dense dynamic object removal pipeline despite the wealth of recoverable scene information available from stereo vision under platform motion (e.g. depth, odometry, optic flow, structure from motion etc.).

1.2. Overview

We propose a two-step approach to tackle this issue in the general sense whilst imposing limited additional computational load. By using intermediary data from the odometry driven stereo mapping process we can isolate the dynamic objects in the scene such as to remove them prior to the final mapping stage. As outlined in Figure 1, we calculate dense stereo disparity maps (D) on all stereo pairs and obtain the platform motion using stereo visual odometry, (SVO) [6]. Using the calculated platform position and full scene depth information we calculate a re-projection map (RP) allowing

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

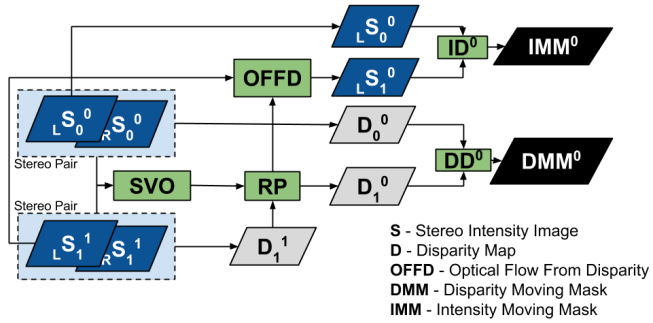

Fig. 1. Processing overview. Diagram naming convention is as follows, ${}_{\mathrm{H}}\mathrm{M}_{\mathrm{T}}{}^{\mathrm{V}}$ – H is frame side (Left or Right), M is the image matrix, T is the time it originates from, V is the viewing time it is projected into. Hence, ${}_{\mathrm{L}}\mathrm{S}_1{}^0$ is the left stereo intensity image from $t = 1$ remapped into $t = 0$.

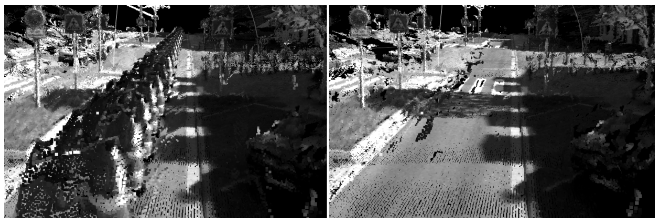

Fig. 2. Left: A car driving forwards, ahead of the logging platform, reconstructed multiple times cluttering up the global point cloud. Right: Moving object removed from the final reconstruction.

the projection of different frames into a common virtual view point. From (RP) we can synthesise an optical flow from disparity (OFFD) map. The dense optical flow map is then used to remap the raw intensity images to the same virtual view point. A 2D projective transform, [16, 17, 12], does not take into account the 3D nature of the scene. By contrast, our approach uses full scene structure and camera motion information to re-project a 2D image with full 3D constraints.

2. 3D SCENE MAPPING

Based on the stereo calibration approach of [18], we recover stereo disparity and hence scene depth based on the approach of [19]. With knowledge of the stereo cameras configuration we can construct a matrix, Q (Eqn. 1) which allows for projection of a 2D disparity image point into 3D world-scaled point clouds (Eqn. 2).

$$Q = \begin{bmatrix} 1 & 0 & 0 & -C_x \\ 0 & 1 & 0 & -C_y \\ 0 & 0 & 0 & f \\ 0 & 0 & a & b \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix} = Q \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix} \tag{2}$$

Fig. 3. Left: Moving object reconstructed multiple times. Right: Moving object masked out of the reconstruction process.

where X, Y, Z are 3D coordinates scaled by W, (x, y) are disparity image pixel locations with disparity value d, $a = \frac{-1}{B}$, $b = \frac{C_x - C_x\prime}{B}$, B is stereo camera baseline in meters, f is camera focal length in pixels, $C_{(x,y)}$ is principal point of left camera and $C_{(x,y)}\prime$ the right. The application of (Eqn. 2) to all recovered disparity values results in a 3D point cloud as per common formulation [18].

Scene mapping in 3D is achieved by creating dense point clouds at multiple camera position. Using SVO[6] we obtain the camera pose for each stereo pair and reconstruct the scene at every camera position and aggregate all the point clouds into a global model. Figures [2, 3] illustrate the output from this process and demonstrate the problem of dynamic objects being reconstructed multiple times.

3. DYNAMIC OBJECT REMOVAL

Central to our approach is the fact that a moving point in space is defined as a rate of change of position in $[\dot{x}, \dot{y}, \dot{z}]$. In order to detect and isolate dynamic objects within the scene we must hence match inter-frame 3D positions on a point-wise basis, between the spatially adjacent stereo camera positions. To enable a point-wise image comparison we perform a scene structure aware projective transform of both disparity and intensity images of consecutive frames into a common camera position.

3.1. Stage 1 - Disparity Projections

The 3D nature of the scene prevents the use of a standard affine transform being used to align disparity maps or intensity images to compare consecutive frames. To align spatially different images we must compensate for the camera motion and scene structure, this is essentially performed by transforming a point cloud by the inverse of the camera motion, then projecting the new motion-compensated 3D points into a synthetic disparity image (Eqn. 3). These three stages can be computed as a single matrix multiplication performed once per stereo pair, encompassing the projection to 3D, the motion compensation and the reprojection back to 2D. Subsequently, we update the new disparity map values to reflect their new distance from the virtual viewpoint (Eqn. 4).

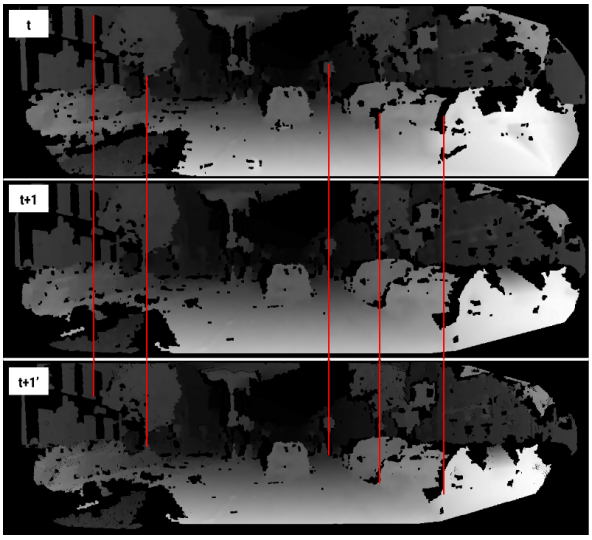

Fig. 4. Top: Disparity map at t. Middle: Disparity map at $t+1$. Bottom: Disparity map at $t+1$ back projected into t now refered to as $t+1'$. Vertical lines illustrate the alignment between scene features.

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \\ 0 \end{bmatrix} = CMQ \begin{bmatrix} x_0 \\ y_0 \\ d_0 \\ 1 \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} r_{31} & r_{32} & r_{33} & t_Z \end{bmatrix} \begin{bmatrix} x_0 - C_x \\ y_0 - C_y \\ f \\ \frac{d_0}{B} \end{bmatrix} = \frac{d_0 f}{d_1} \tag{4}$$

where $M = [R \mid T]^{-1}$, R, T are the rotation and translation components of the camera motion from SVO respectively, C is camera intrinsic matrix, $(x_0,\ y_0)$ are 2D image image coordinates in source disparity image, $(x_1,\ y_1)$ are 2D image image coordinates in synthetic disparity image, r_{rc} are the rotation components from M at row r and column c, f is camera focal length in pixels, $(d_0$,$d_1)$ are source and destination disparity respectively and t_Z is the translation component of platform motion in Z-axis.

Transform (Eqn. 3) creates a new synthetic disparity image that corresponds to a virtual camera at the location of the previous camera position. A scaling transform (Eqn. 4) updates disparity values in the synthetic disparity map to reflect the new distance that points lie away from the virtual camera position. Figure 4, shows three disparity images, from two stereo pairs. The top disparity image is at time t and middle disparity image at $t+1$, the bottom image is disparity at $t+1$ projected (via Eqns. 3, 4) into the virtual camera position of t. Observing the red vertical lines shows how features such as, windows, signs and backs of cars are now aligned allowing for direct point-wise comparison of disparity at t and $t+1$.

Spatial point-wise alignment of temporally separated disparity maps permits us to compute a binary moving object mask. This is done by performing a point-wise difference image between the two projection aligned disparity maps, creating a Disparity Difference map (DD, Figure 1). The work of [14] shows that the non-linear 3D triangulation error from various dense stereo matching algorithms can be represented by a disparity matching error in pixel terms in disparity space. The disparity maps we produce are calculated using Semi-Global Block Matching (SGBM) [19]. The estimated SGBM stereo matching error for real-world data is approximately e=0.2 pixels [14]. We use this accuracy metric to threshold the disparity difference map to populate the binary Disparity Moving Mask (DMM, Figure 1) (Eqn. 5).

$$DMM_{xy} = \begin{cases} 0 & DD_{xy} \geq e \\ 1 & otherwise \end{cases} \tag{5}$$

The DMM is used to reject regions of the disparity map used for the 3D reconstruction. Figure 2 illustrates the aggregated point cloud reconstruction before and after the moving object removal stage.

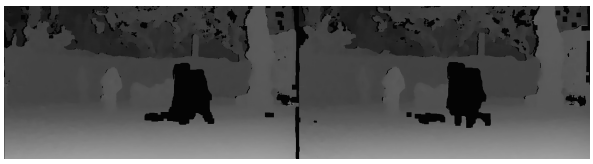

Fig. 5. Disparity maps used to create Figure 3 with moving objects masked out. Shadows cast by the moving objects are also masked out due to intensity image variation.

3.2. Stage 2 - Optical Flow From Disparity

Step two of processing aids object removal by examining the intensity consistency between consecutive frames. Likewise with disparity map projections we reduce the search space by aligning the images so that direct point-wise comparisons can be performed. Traditionally frame alignment would be done by applying a 2D homography transform, essentially performing image stabilization [16, 17]. This approach is insufficient by the fact that a homography transform is a 2D projective transform and does not take into account the 3D nature of the scene. [12] attempts to solve this by applying multiple homograph transforms to image regions of 14x14 pixels. Accurate dense optical flow techniques [20] required for this level of intensity projection are computationally expensive. We avoid dense optical flow calculations entirely as scene structure aware remapping of 2D points at time t into a 2D point at $t+1$ as previously calculated in Eqn 3. Using the optical flow from disparity (OFFD, Figure 1) we can remap intensity images allowing for point-wise comparison of aligned intensity images resulting in an Intensity Difference map (ID, Figure 1). Applying an appropriate threshold yields an intensity moving mask (IMM, Figure 1). Applying

Fig. 6. Top: A slow moving pedestrian with approximately 50% self-overlap between frames. Bottom: Successfully removed from mapping solution.

the DMM and IMM to our dense disparity maps we mask out the regions pertaining to dynamic objects, Figure 5.

4. RESULTS

We used the popular KITTI stereo dataset [21] and our own image sequences captured from a moving vehicle at speeds approximately 10-15mph and sample rate 7.5Hz. Figure 2, shows a typical road traffic scene with a vehicle preceding the camera reconstructed multiple times. Successful removal of the dynamic object is performed primarily via the disparity projection stage as the object has sufficient disparity variation with respect to the static background. Figures 3 and 6, demonstrate a case where a complex object is moving perpendicular to the camera. The objects have approximately 50% frame-to-frame overlap with themselves, therefore some parts remain at constant disparity. Stage 2 successfully removes them from the final map via intensity variation. A large group of people in Figure 7 are mostly removed but some elements that remained static between frames are still present. An interesting result, Figure 8, is where a fast walking pedestrian is removed from the map, however closer inspection reveals the feet are still present in the final point cloud. The feet of pedestrians are static with respect to the road surface over consecutive frames therefore are not classed as dynamic on this timebase. Comparison across a greater time base is required for full removal. Further results can be viewed on https://youtu.be/MAA_Uq0KHoY. Input datasets, in the form of calibrated stereo images, can be downloaded from (http://dx.doi.org/10.15128/1544bp08d).

5. CONCLUSION

This work has demonstrated we can re-use the disparity maps and odometry produced for the mapping solution as a significant data source for the removal of dynamic objects. Our method adds a predictable processing overhead proportional only to image size, unlike previous attempts that use feature points or computationally expensive segmentation algorithms. We demonstrate accurate motion masks can be created in order to enable removal of dynamic objects from 3D maps, this is illustrated upon on two different datasets, KITTI and our own, through varying camera motions and dynamic object characteristics. An extension to this work would be performing a quantitative evaluation and testing images captured in different weather conditions such as rain [22]. Extending this work to platforms that differ greatly [23, 24] will test the robustness and flexability of this approach.

Fig. 7. Top: Large group of people moving at various speeds including some static bystanders. Bottom: Pedestrians are largely removed.

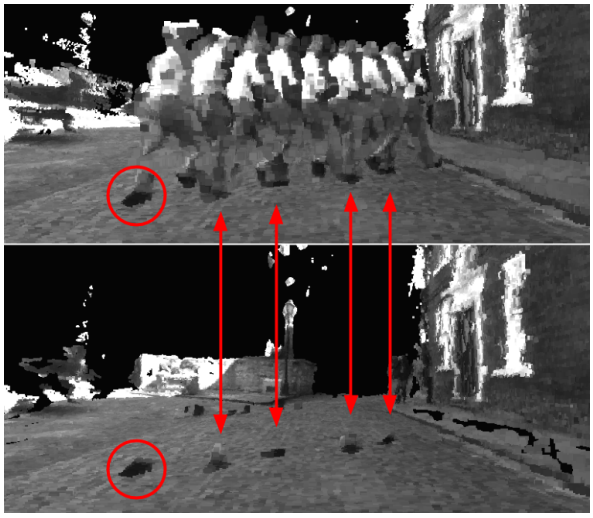

Fig. 8. Top: Fast moving pedestrian with motion perpendicular to camera motion. Bottom: Successful removal of moving components, however the feet remain visible in the point cloud as these are temporarily static during contact with the ground.

6. REFERENCES

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 72–79, Sept. 2009.

[2] S. Thrun, W. Burgard, and D. Fox, "A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping," in *Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference*, 2000, vol. 1, pp. 321–328 vol.1.

[3] S. Song and M. Chandraker, "Robust Scale Estimation in Real-Time Monocular SFM for Autonomous Driving," in *IEEE International Conference on Intelligent Robots and Systems, IROS'14*, June 2014, pp. 1566–1573.

[4] D. Gallup, J. M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1418–1425, 2010.

[5] A. Geiger, M. Roser, and R. Urtasun, "Efficient Large-Scale Stereo Matching," *Computer Vision–ACCV 2010*, pp. 25–38, 2011.

[6] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 963–968, June 2011.

[7] K. Konolige, "Small Vision Systems: Hardware and Implementation," *Robotics Research*, pp. 203–212, 1998.

[8] A. Kundu, K. M. Krishna, and J. Sivaswamy, "Moving object detection by multi-view geometric techniques from a single camera mounted robot," *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, pp. 4306–4312, Oct. 2009.

[9] A. Kundu, K. M. Krishna, and C. V. Jawahar, "Realtime motion segmentation based multibody visual SLAM," *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 251–258, 2010.

[10] J. Klappstein, T. Vaudrey, C. Rabe, A. Wedel, and R. Klette, "Moving object segmentation using optical flow and depth information," in *Pacific Rim Symposium on Image and Video Technology*, 2009, pp. 611–623.

[11] P. Lenz, J. Ziegler, A. Geiger, and M. Roser, "Sparse scene flow segmentation for moving object detection in urban environments," in *Intelligent Vehicles Symposium (IV)*, June 2011, pp. 926–932.

[12] J. Hariyono, V.-D. Hoang, and K.-H. Jo, "Moving object localization using optical flow for pedestrian detection from a moving vehicle.," *The Scientific World Journal*, vol. 2014, 2014.

[13] A. Bak, S. Bouchafa, and D. Aubert, "Detection of independently moving objects through stereo vision and ego-motion extraction," *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pp. 863–870, June 2010.

[14] O. K. Hamilton, T. P. Breckon, X. Bai, and S. I. Kamata, "A foreground object based quantitative assessment of dense stereo approaches for use in automotive environments," in *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings*, 2013, pp. 418–422.

[15] A. Wedel, A. Meiß ner, C. Rabe, U. Franke, and D. Cremers, "Detection and segmentation of independently moving objects from dense scene flow," in *Lecture Notes in Computer Science*, vol. 5681, pp. 14–27. 2009.

[16] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust L1 optimal camera paths," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 225–232, 2011.

[17] A. Censi, A. Fusiello, and V. Roberto, "Image stabilization by features tracking," *Proceedings - International Conference on Image Analysis and Processing*, pp. 665–669, 1999.

[18] R. I. Hartley, "Theory and practice of projective rectification," *International Journal of Computer Vision*, vol. 35, no. 2, pp. 115–127, 1999.

[19] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–41, 2008.

[20] G. Farneb, "Two-Frame Motion Estimation Based on Polynomial Expansion," *Lecture Notes in Computer Science*, vol. 2749, no. 1, pp. 363–370, 2003.

[21] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[22] D. Webster and T. Breckon, "Improved Raindrop Detection using Combined Shape and Saliency Descriptors with Scene Context Isolation," in *Proc. International Conference on Image Processing*, 2015, pp. 4376–4380.

[23] T. Kriechbaumer, K. Blackburn, T. P. Breckon, O. Hamilton, and M. R. Casado, "Quantitative Evaluation of Stereo Visual Odometry for Autonomous Vessel Localisation in Inland Waterway Sensing Applications," *Journal of Electronic Imaging*, vol. 24, no. 5, pp. 1–17, 2015.

[24] P. Cavestany, A. Rodriguez, H. Martinez-Barbera, and T. Breckon, "Improved 3D sparse maps for high-performance Structure from Motion with low-cost omnidirectional robots," in *Proc. International Conference on Image Processing*, 2015, pp. 4927–4931.