Real-Time Automotive Street-Scene Mapping Through Fusion of Improved Stereo Depth and Fast Feature Detection Algorithms

Hossein Javidnia, Peter Corcoran

Department of Electronic and Electrical Engineering, College of Engineering, National University of Ireland, Galway
{h.javidnia1, peter.corcoran}@nuigalway.ie

***Abstract*-- The real-time tracking of street scenes as a vehicle is driving is a key enabling technology for autonomous vehicles. In this work we provide the basis for such a system through combining an improved advanced random walk with restart technique for stereo depth determination with fast, robust feature detection. The enables tracking and mapping of a wide range of scene structures which can be readily resolved into individual objects and scene elements. Thus it is practical to identify moving objects such as vehicles, pedestrians and fixed objects and structures such as buildings, trees and roadside kerb.**

I. INTRODUCTION

In the United States, over 30,000 people are killed in motor vehicle crashes each year. In 2013, fatal vehicle crashes resulted in costs of $44 billion in medical and work loss [1]. Over the past decade, there has been significant progress to realize Intelligent Driver Assistance Systems (IDAS) and autonomous navigation. These systems aim to improve safety on the roads and the next step for IDAS is to enable direct visual monitoring of the on-road environment [2, 3, 4, 5]. In this context, having an accurate depth map of the surrounding environment allows the IDAS to determine and analyze the local road context and safely interact with pedestrians and other vehicles. This can eliminates the need for expensive range sensors on each vehicle and enables more sophisticated visual interpretation (e.g. object recognition) of the road scene.

The most common and cost effective method to compute the depth of a scene is to employ image processing algorithms on cameras. Generally single cameras do not have adequate Field of View (FoV) to monitor the full extent of a road scene and so blind spots will remain in the case of using only 1 camera. That is why in this paper frames from stereo cameras are used for extraction of depth map

This paper presents a new method for refining the accuracy of depth maps for real-time automotive applications based on the post-processing of the Adaptive Random Walk with Restart (ARWR) algorithm and its fusion with Oriented FAST and Rotated BRIEF (ORB) feature detection. At the end the evaluation results of the mentioned method with KITTI[1] and Middlebury benchmark are being presented.

II. ALGORITHMS & IMPLEMENTATION

A. ARWR + Post-Processing

ARWR is a local stereo computation algorithm based on random walk with restart method [6]. ARWR has an acceptable and comparable performance in terms of estimation and speed against other algorithms. The algorithm has some important advantages such as not being affected by illumination variation because of gradient and census transform, having a quite good performance in both outside and inside environment which gives us the option to have an estimation of the depth in low texture scenes. These advantages make this method suitable to be used in different applications such as automotive navigation, scene 3D reconstruction, drones navigation and etc. But the algorithm is suffering from some problems. The depth map generated by ARWR is suffering from speckle noise. Edges and corners are inaccurate especially for objects with a detailed geometry. At some parts of the computed depth map the edges are broken or they are faded into other objects.

These problems brought a challenge to think of increasing the accuracy of the estimated depth by designing a post-processing framework. The framework starts with Adaptive Random Walk with Restart algorithm. To refine the depth map generated by this method, we introduced a form of median solver/filter based on the concept of the mutual structure which refers to the structural information in both images. This filter is later enhanced by a joint filter. Next, a transformation in image domain is introduced to remove the artifacts which cause distortion in the image. Fig. 1 presents the detailed diagram of the mentioned framework.

B. Feature Detection and 3D Reconstruction

In the second part of this study the localization mapping based on the ORB features [7] is considered. ORB is basically a fusion of Features from Accelerated Segment Test (FAST) keypoint detector and Binary Robust Independent Elementary Features (BRIEF) descriptor. It uses an oriented FAST detection method and the rotated BRIEF descriptors which make it scale- and rotation invariant.

This feature has been used in the last couple of years in several real-time feature-based applications like monocular SLAM systems that operates in real time, in small and large, indoor and outdoor environments.

For the localization part of the experiment, a system based on ORB features is employed to map the surrounding environment. The extracted 2D features in each image/frame have been used to solve the PnP problem to estimate the pose of the calibrated camera from n 3D-to-2D correspondence. It enables us to estimate the camera pose based on the 2D points and their distance to camera.

This system is also able to detect the ground plane and differentiate the objects above the ground. Fig. 2 and Fig. 3 show the matched ORB features between 2 frames and the 3D projection of the points respectively.

C. Basic Object & Structure Identification

The goal of this fusion is to estimate the accurate distance of

[1] Karlsruhe Institute of Technology (KIT) and Toyota Technological Institute at Chicago (TTI-C)

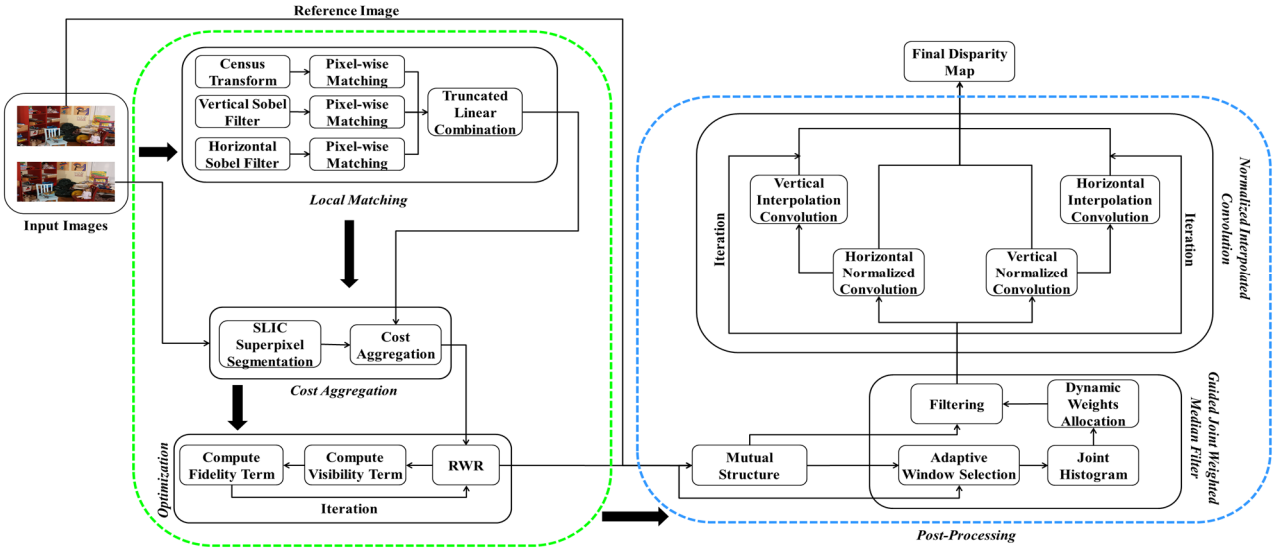

Figure 1. Overview of the ARWR + Proposed post-processing method

different obstacles including pedestrian, cars, ramps and etc to the camera. The basic structure of the objects that are aimed to be detected at the initial phase of the project is limited to their height. Objects higher than 15 cm above the ground will be detected. Later classification techniques will be used to recognize the objects above the ground plane.

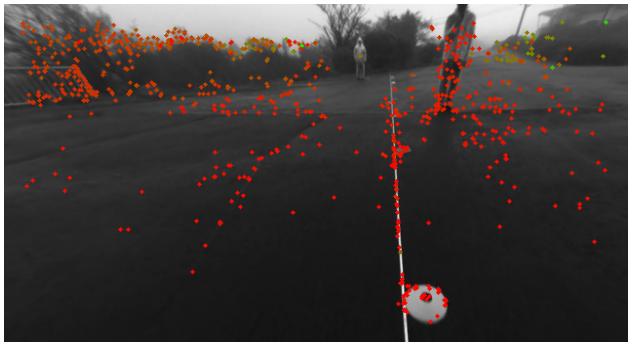

Figure 2. Matched ORB features between 2 frames in a video sequence

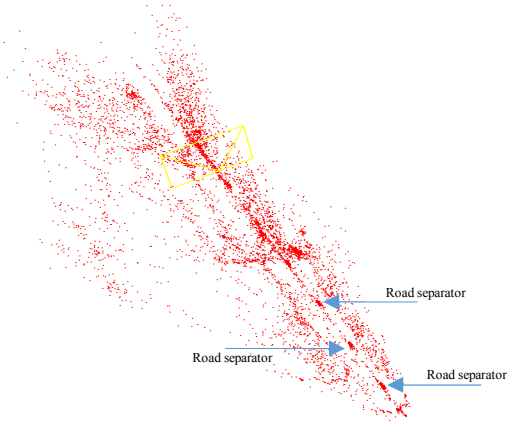

Figure 3. 3D projection of the matched features from the same video sequence as Fig. 2

III. Initial Experiments and Outline of Results

A. Evaluation Metrics for Depth Determination

The Middlebury benchmark [8] is used as a basis for comparing the first step of this approach (ARWR with post processing) with top 8 algorithms evaluated by this benchmark. The key metrics used include (i) the mean squared error (MSE); (ii) the root mean squared error (RMSE); (iii) the peak signal to noise ratio (PSNR); (iv) the averaged signal to noise ratio (SNR); (v) the mean absolute error (MAE); (vi) the structural similarity index (SSIM); and (vii) the structural dissimilarity (DSSIM) a distance metric derived from SSIM. The advantage of using Middlebury is that it provides a publicly available test metric and images with detailed ground truth information.

The scores of the improved ARWR algorithm for the MSE, PSNR, SNR, SSIM and DSSIM metrics has proved that the proposed post-processing method has the best performance among all others currently available in Middlebury. These evaluations were based on the 15 standard images of Middlebury dense training set.

Table 1 presents the average numerical results of the described evaluation.

B. Experiments on real Video Sequences

There are around 14 plots and tables which show the numerical details of the evaluation against Middlebury benchmark and visual results including the real-time application use and depth from stereo images which are ready to demonstrate in details at ICCE 2017.

All the evaluation results are available to download form the following link: **https://goo.gl/97mIR4**

Table 1. Average values of metric/algorithm

	GCSVR	INTS	MCCNN_ Layout	MC-CNN+FBS	MC-CNN-acrt	MC-CNN-fst	MeshStereo	SOU4P-net	Original ARWR	Post-processed ARWR
MSE	0.0235	0.0193	0.0133	0.0194	0.0171	0.0177	0.0195	0.0133	0.0277	**0.0126**
RMSE	0.1456	0.1339	**0.104**	0.1243	0.1199	0.1225	0.1357	0.1069	0.1455	0.1041
PSNR	17.2273	17.8519	20.2902	19.2026	19.028	18.842	17.6704	19.9836	17.7517	**20.3136**
SNR	12.3262	12.9509	15.3891	14.3016	14.1269	13.9409	12.7694	15.0826	12.8506	**15.4125**
MAE	0.112	0.1026	**0.0524**	0.0915	0.0639	0.0666	0.101	0.0739	0.0867	0.0644
SSIM	0.9969	0.9976	0.99849	0.9975	0.9981	0.998	0.9975	0.99847	0.9966	**0.9985**

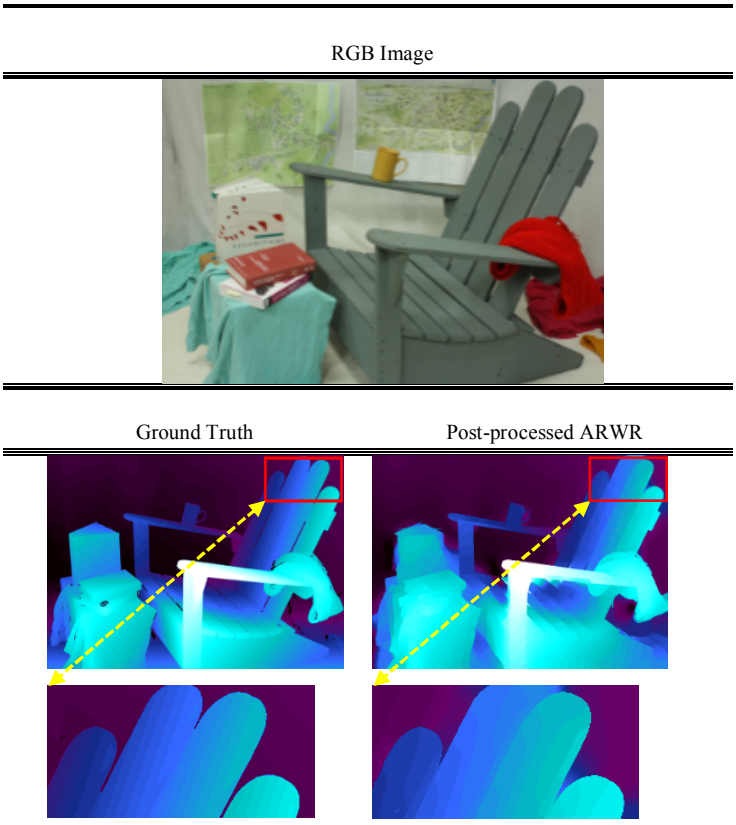

Figure 4. The result of a sample images from Middlebury database

Here we present some parts of our results of the depth from stereo on a test image taken from Middlebury benchmark in Fig. 4 and distance estimation on an image taken from KITTI benchmark by using the fusion of the mentioned methods in Fig. 5 and Fig. 6 respectively. Fig.4 presents the left view of a stereo set from the mentioned benchmark titled as RGB Image. The depth ground truth and estimated depth based on improved ARWR is presented as well. As visual comparison it is clear how close is the improved ARWR to the depth ground truth in terms of preserving the edges and corners.

Fig.5 and Fig. 6 show another left view of a stereo set from KITTI benchmark. The top image shows the normal RGB image and the image in the middle shows the computed depth by improved ARWR. The last image in each figure shows the estimated distance to an object which is computed based on the generated depth map and extracted ORB features.

IV. Conclusions & Discussion

In this paper a post-processing technique to increase the accuracy of the depth map computed by Adaptive Random Walk with Restart method is proposed and evaluated. The generated depth is useful in variety of applications with different purposes. In this experiment a framework is provided to determine the distance to different obstacles in outdoor environment, mainly on the street by fusing the improved ARWR technique and 3D localization using ORB features.

The initial distance by the ARWR is modified by a process based on ORB feature matching. Localization part estimates the distance by solving the P*n*P problem. The final distance is an average value of the estimated depth and the distance from the 3D localized points.

In this research work we found out that keeping the sharp edges and corners along with main structure of the reference image is very helpful in several applications such as segmentation, 3D reconstruction and real-time applications. Also the proposed method is able to estimate the distance to different objects within 2% error which can be useful in automotive navigation.

There are still number of open challenges such as decreasing the processing time with low processing power which motivate us for the future works. In our future work the plan is to employ bilateral filter to decrease the processing time.

The real-time results and the visualization of the fusion part will be presented at ICCE 2017.

Preliminary results show that the various challenges can be addressed and a practical proof of concept will be ready to demonstrate by ICCE 2017.

Figure 5. Distance estimation based on the computed depth. First top image represents the original RGB image from a video sequence. The image at the middle shows the estimated depth map using the proposed method. The last image shows an object labelled with its distance to the camera based the depth map.

Figure 6. Distance estimation based on the computed depth. First top image represents the original RGB image from a video sequence. The image at the middle shows the estimated depth map using the proposed method. The last image shows an object labelled with its distance to the camera based the depth map.

ACKNOWLEDGMENT

The research work presented here was funded under the Strategic Partnership Program of Science Foundation Ireland (SFI) and co-funded by SFI and FotoNation Ltd. Project ID: 13/SPP/I2868 on "*Next Generation Imaging for Smartphone and Embedded Platforms*".

REFERENCES

[1] (2015). *State-Specific Costs of Motor Vehicle Crash Deaths*. Available: https://www.cdc.gov/motorvehiclesafety/statecosts/

[2] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Transactions on Intelligent Transportation Systems,* vol. 14, pp. 1773-1795, 2013.

[3] V. Usenko, J. Engel, J. Stuckler, and D. Cremers, "Reconstructing Street-Scenes in Real-Time from a Driving Car," in *3D Vision (3DV), 2015 International Conference on*, 2015, pp. 607-614.

[4] 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, pp. 1-19, 2012.

[5] J. X. M. K. M.-T. S. A. Geiger, "Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[6] S. Lee, J. H. Lee, J. Lim, and I. H. Suh, "Robust stereo matching using adaptive random walk with restart algorithm," *Image and Vision Computing,* vol. 37, pp. 1-11, 2015.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, 2011, pp. 2564-2571.

[8] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Stereo and Multi-Baseline Vision, 2001. (SMBV 2001). Proceedings. IEEE Workshop on*, 2001, pp. 131-140.