International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

Object Recognition and Tracking for Scene Understanding of Outdoor Mobile Robot

My-Ha Le
Faculty of Electrical and Electronic Engineering
Ho Chi Minh City University of Technology and Education
No.1 Vo Van Ngan Str., Thu Duc Dist., Ho Chi Minh City, Viet Nam

Dinh-Huan Nguyen
Faculty of Electrical and Electronic Engineering
Ho Chi Minh City University of Technology and Education
No.1 Vo Van Ngan Str., Thu Duc Dist., Ho Chi Minh City, Viet Nam

Abstract—**This paper proposes a method for object recognition and tracking in scene understanding system of outdoor mobile robot. Firstly, the moving objects in front of robot are detected using optical flow method based on stereo camera system. The information about the size, the distance of object to the robot are extracted in real-time condition. Secondly, the object categories are recognized base on the training data set and the convolutional neural network classifier. Thirdly, the objects are tracked during their appearance in the region of view of camera. The simulation results will demonstrate the effectiveness of this method.**

Keywords- optical flow, stereo camera, object recognition, Convolutional Neural Network, classifier, machine learning, tracking

I. INTRODUCTION

Moving object detection, tracking, and identification is one of important process in various applications of autonomous mobile robot navigation and advanced driver assistance systems. Other application also can be considered are virtual environment and scene planning. Some progress has been made in scene understanding and 3D scene modeling. The results were obtained during the last few years but they needed a large amount of work done by hand or apparatus, such as laser radar, and airborne light detection and ranging. They are usually expensive and require much more time for data acquisition.

In recent years, many algorithms have been developed for motion estimation, which can roughly be devised into several categories, namely methods using LIDAR (Light Detection and Ranging) system [1-4], methods based on segmentation of single camera [5-6] and the method combined camera and range measurement sensor [7-12]. In the first group, LIDAR system provides information of 3D measurement from scene. This is the time-of-flight device. It emits the beam to the space and measures the time to absorb the reflexed beam when it hit the objects. The absorption depended on the property of objects and the distance to the objects. Therefore, the 3D scene is reconstructed but it is time consuming and limited distance. In the scene understanding system for high speed moving robot/vehicle, the requirement for long distance scene understanding is a compulsory. In the second group, the segmented regions from scene are labeled using context information. However, the using of single camera the depth measurement of scene is not recovered. In the third group, the combination of camera and range measurement may improve the depth information of scene. Nevertheless, the calibration is need to be carried out beforehand when the system is applied in the reality. The non-reflection object or limitation in measurement distance of range device may be the limitation of this method.

Without using any addition range finder device, e.g. laser sensor out of stereo camera, our proposed method overcomes some disadvantages mentioned above. It is much cheaper and compact. Most of experiment presents that stereoscopic yield a better performance for sparse scene reconstruction and SLAM (Simultaneous Localization and Mapping) [13-16]. The optical flow method based on stereo device [17] is applied in this research to detect the moving objects. The objects are then identified by convolutional neural network from the trained data set. The movement of the objects is also tracked during its appearance in the region of view of robot/ vehicle.

The flow chart of proposed method can be seen in figure 1. From stereo system, sequence image are acquired along the scene. SIFT algorithm [18-19] is applied to find invariant feature and matching of each consecutive-overlap pair of views. The optical flow of spare scene is extracted. The optical flow region of moving objects are identified in the consequence frames. The intrinsic parameters of camera are calculated based on Jean-Yves Bouguet [20] method. The region surround the identified objects are extracted. The image region is then feed to convolutional neural network to classify object. The tracking process is applied to follow the object during its movement.

This paper is organized into six sections. The next section summaries moving object detection method using optical flow. Section III presents application of convolutional neural network to object classification. The tracking process is described shortly in section IV. Experiments are showed in section V. Finally, paper is finished with conclusions and point out future works discussed in section VI.

II. OPTICAL FLOW BASED MOVING OBJECT DETECTION

In order to compute the distance to the sparse region of moving we first find intrinsic parameters of camera. In this section, we explain what is camera model, how to extract and match salient features, matching, sparse region of moving

This project was supported by Ho Chi Minh City University of Technology and Education, 2016 (T2016-53TĐ)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

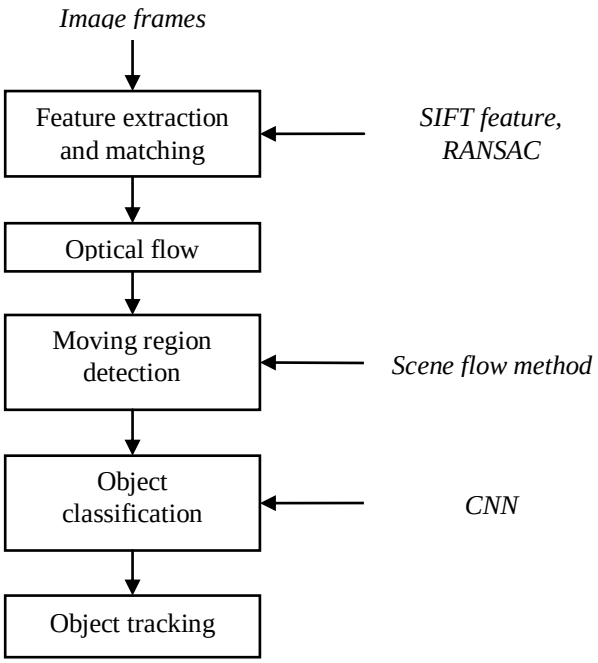

Figure 1. General proposed scheme

object detection. The brief description of this step followed the method in [17] is showed in the over view of figure 2.

A. Camera model

We use the projective geometry throughout this paper to describe the perspective projection of the 3D scene onto 2D images [15]. This projection is described as follows:

$$x = PX \tag{1}$$

where P is a 3×4 projection matrix that describes the perspective projection process, $X = [X, Y, Z,1]^T$ and $x = [x,y,1]^T$ are vectors containing the homogeneous coordinates of the 3D world coordinate, respectively, 2D image coordinate.

When the ambiguity on the geometry is metric, (i.e., Euclidean up to an unknown scale factor), the camera projection matrices can be put in the following form:

$$P = K[R|-RT] \tag{2}$$

with T and R indicating the translation and rotation of the camera and K , an upper diagonal 3×3 matrix containing the intrinsic camera parameters.

$$K = \begin{bmatrix} f_x & s & u_x \\ & f_y & u_y \\ & & 1 \end{bmatrix} \tag{3}$$

where f_x and f_y represent the focal length divided by the horizontal and vertical pixel dimensions, s is a measure of the skew, and (u_x , u_y) is the principal point. The check board used for calibration is present in figure 3.

B. Feature extraction and matching

There are many kind of features are considered in recent research in feature extraction and matching problem including

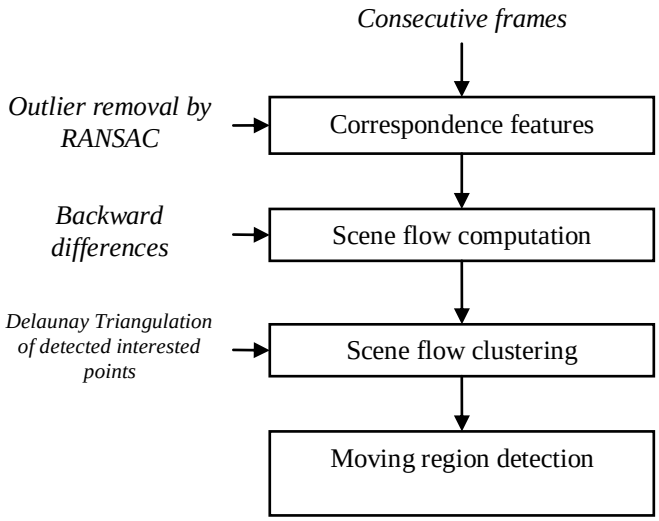

Figure 2. Moving region detection scheme

Harris [21], SIFT, PCA-SIFT, SURF [22], [23], etc. SIFT is first presented by David G Lowe in 1999 and it is completely presented in 2004. As we know on experiments of his proposed algorithm is very invariant and robust for feature matching with scaling, rotation, or affine transformation. According to those conclusions, we utilize SIFT feature points to find correspondent points of image pairs. The SIFT algorithm are described through these main steps: scale-space extrema detection, accurate keypoint localization, orientation assignment and keypoint descriptor. SIFT features and matching is applied for one image pair as showed in Fig. 4. The result of correspondence point will be used to compute spare scene flow.

C. Scene flow computation

The result of correspondence point in previous step will be used to compute scene flow. The proposed method was presented in [15]. From the correspondence points and intrinsic parameter of camera, the real world coordinate X, Y, Z of points are reconstructed in 3D as follows:

$$X = \frac{(x_L - u_{x,L}) \cdot b}{d} \tag{4}$$

$$Y = \frac{(y_L - u_{y,L}) \cdot b}{d} \tag{5}$$

$$Z = \frac{f \cdot b}{d} \tag{6}$$

where b denotes the baseline of the stereo system. This parameter is measured from the system. The velocity of the real world points are calculated by the first derivative of coordinate. The backward differences of delay of occurrence and object detection are used to determine the scene flow.

D. Moving object detection

The moving object detection is based on clustering of scene flow. The sparse scene flow points are connected by Delauney triangulation. The difference of velocity of each

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

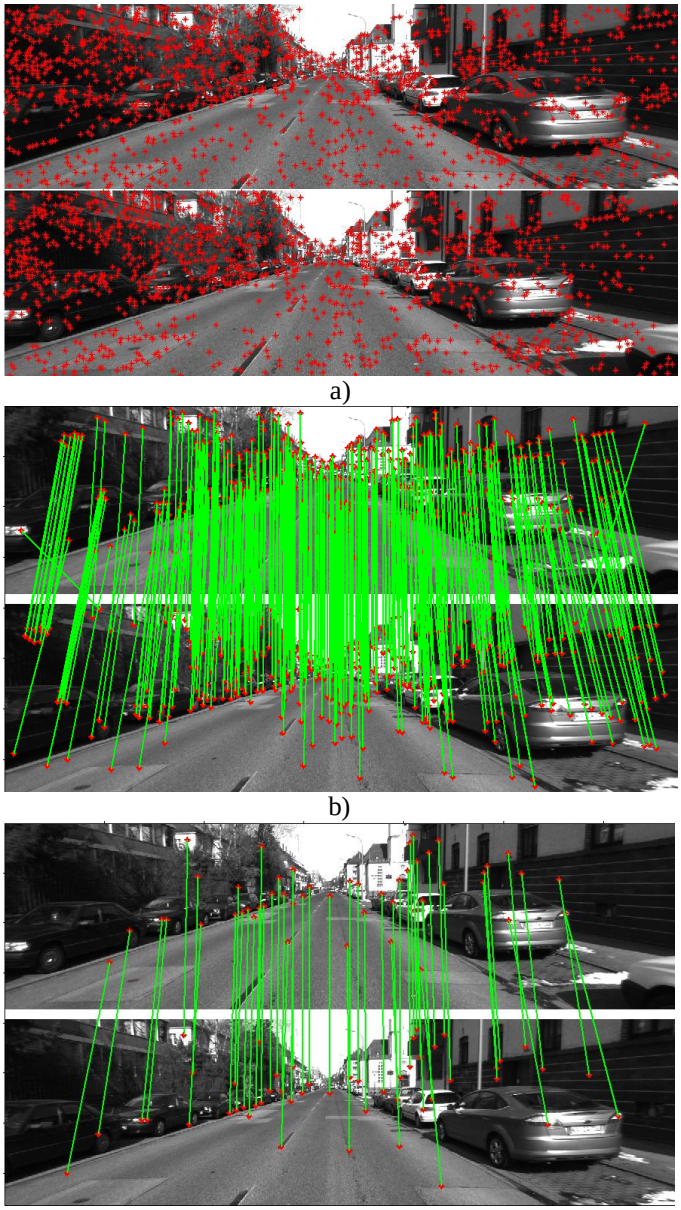

Figure 4. SIFT feature extraction and matching. a) SIFT features, b) features matching before RANSAC, c) features matching after RANSAC

interest point is compare by Mahalanobis distance. The similar moving region will have the difference of velocity smaller than a certain threshold.

III. CONVOLUTIONAL NEURAL NETWORK FOR OBJECT CLASSIFICATION

Object recognition is the interesting topic has been investigated last several decades. Usually, the process consists of two main stage: feature extraction and classifier. The features, such as Histogram of oriented gradient, SIFT, SURF or Harris have been applied for object recognition in indoor as well as outdoor environment. The classifiers, such as Support vector machine, Adaboost, are also researched in variety applications. Among them, Convolutional Neural Network, the hot topic in machine learning nowadays, has been used for classification a variety objects in the visual recognition applications. In this research a CNN have the same structure

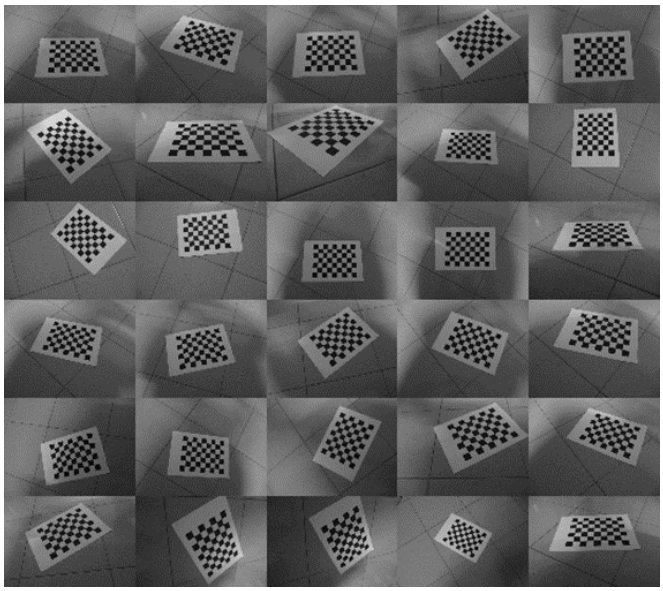

Figure 3. Camera calibration templates

as LeNet5. The method poposed by LeCun et. at. [24] was applied. The structure of the Network is presented in the figure 5. The input image is convolute with the feature to create feature maps. The number of feature is selected based on the object categories. The feature maps are then resampling. There have to common ways for sub-sampling: average pooling and max pooling. In this research, the max pooling method is applied. The convolution and pooling process is repeated several times depended on the number of feature use to train the network. Finally, the output of final feature maps are deed to the fully connected layer, this is multiple layer perceptron. The number of neural and hidden layers depended on the selection of user. This step is treat as conventional neural network. One the structure of network is built, the training process is performed.

IV. OBJECT TRACKING

The object tracking process is performed based on GNN approach proposed in [25]. The idea of this method relies on the prediction of detected objects in the next frame. Once the position of detected object is detected in the previous frame, a similarity is measured in a surround region after one frame. If the similarity comparison is fail, then the detected object is labeled as the new object, and it is also tracked in the next frame.

V. EXPERIMENTS

We experimented on outdoor images which are acquired from two DSLR Nikon camera mounted on the car as shown in figure 6. The trajectory are the large regions. One data set is collected in the campus area, the other one is collected on the urban street. All result were carried on Intel(R) Core(TM) i5 CPU 750@2.67 GHz with 3GB RAM under Matlab environment. In the first and the second experiment, we run 10220 and 11193 frames with size [1280x720], respectively. Figure 7(a), 7(b) are the results of object detection, tracking and recognition. It is easily to figure out from the figures: with the sparse scene, objects can be detected and recognized with high accuracy. In this research the accuracy is 90% for campus data. Figure 8(a), 8(b) are the the results of urban street data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

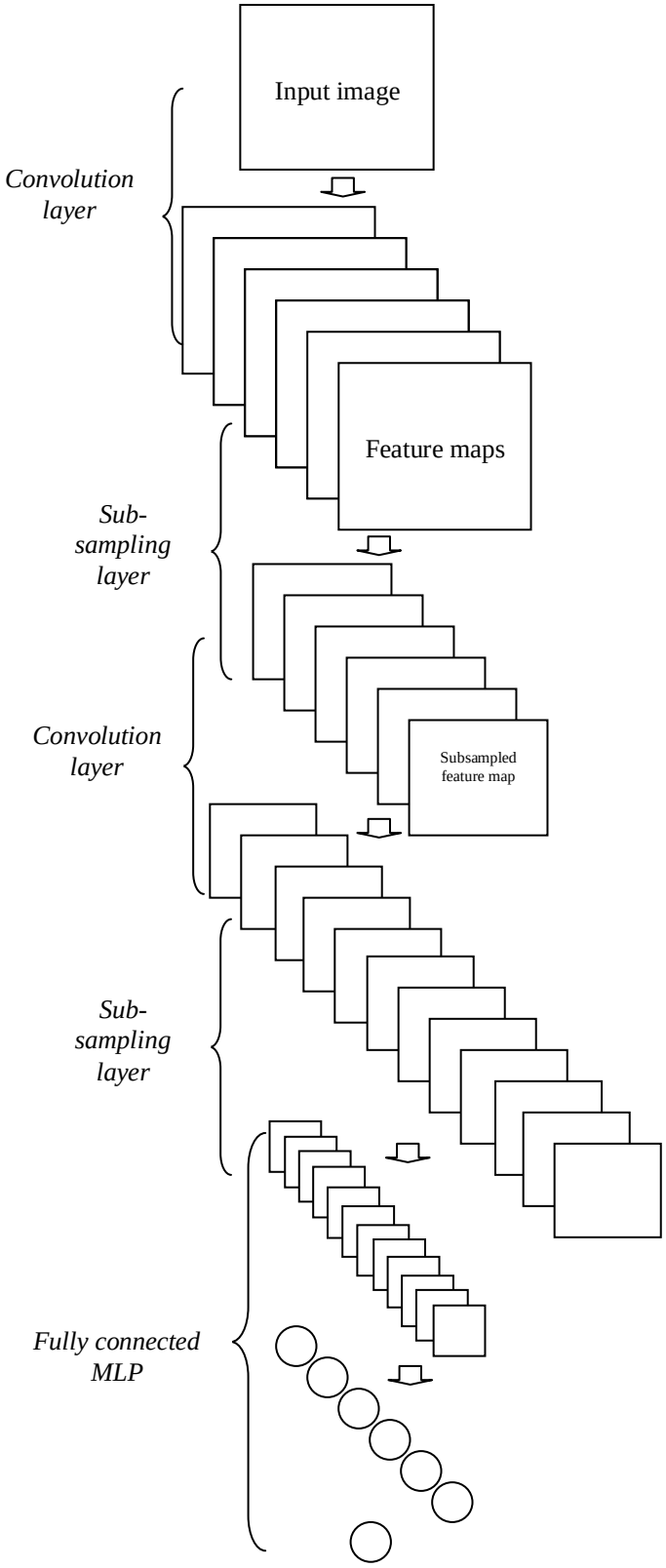

Figure 5. Structure of CNN for object classification (LeNet [24])

When the density of object is dense and exist multiple value of difference velocity of objects, the moving object detection doesn't detect correctly.

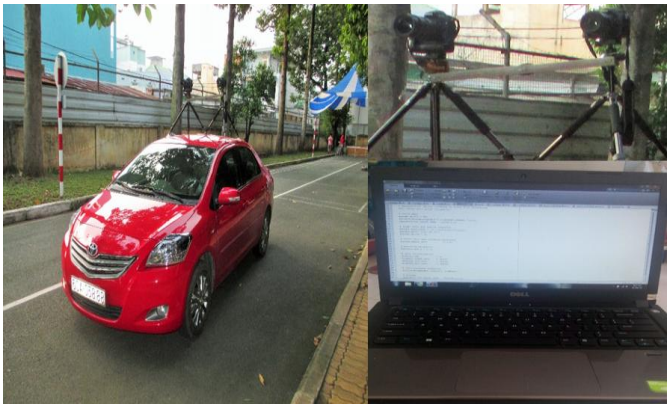

Figure 6. Hardware configuration. Two DSLR Nikon camera mounted on the car.

a)

Figure 7. Object detection, tracking and recognition. a) and b) are results of the campus data.

VI. CONCLUSION

The results of object detection, recognition, and tracking method are presented in this paper. Some advantage points can be realized through our explanation. First, the moving object is detected and tracked separately with the background. The optical flow based method can cluster the similar velocity of interest points. The tracking algorithm applied in this research is also robustness in outdoor condition. Second, object

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

a)

b)

Figure 8. Object detection, tracking and recognition. a) and b) are results of urban street data

recognition base on one kind of deep learning method. The CNN gave the high accuracy in object recognition. Our future woks focus on adaptive detection of moving object when exist multiple difference of velocity. In addition, we will improve and develop this method for smart system can recognize and analyze the scene.

ACKNOWLEDGMENT

This project was supported by Ho Chi Minh City University of Technology and Education, 2016 (T2016-53TĐ)

REFERENCES

[1] R. Triebel, K. Kersting, and W. Burgard. Robust 3d scan point classification using associative markov networks. In Proc. IEEE Int. Conf. Robot. Automat., pages 2603–2608, Orlando, May 2006.

[2] A. Agrawal, A. Nakazawa, and H. Takemura. MMM-classification of 3d range data. In Proc. IEEE Int. Conf. Robot. Automat., pages 2269–2274, Kobe, May 2009.

[3] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard. Unsupervised discovery of object classes from range data using latent dirichlet allocation. In Robotics: Science and Systems V, Seattle, USA, Jun. 2009.

[4] R. Triebel, K. Kersting, and W. Burgard. Robust 3d scan point classification using associative markov networks. In Proc. IEEE Int. Conf. Robot. Automat., pages 2603–2608, Orlando, May 2006.

[5] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. Moving obstacle detection in highly dynamic scenes. In *ICRA*, 2009.

[6] A. Ess, T. Mueller, H. Grabner, and L. van Gool. Segmentation-based urban traffic scene understanding. In *BMVC*, 2009

[7] J. Strom, A. Richardson, E. Olson, Graph-based segmentation for colored 3d laser point clouds, in: roceedings of the IEEE International Conference on Intelligent Robots and Systems, 2010, pp. 2131–2136.

[8] J. P. C. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, P. H. Torr, Mesh based semantic modelling for indoor and outdoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2067–2074.

[9] M. Haselich, D. Lang, M. Arends, D. Paulus, Terrain classification with Markov random fields on fused camera and 3D laser range data, in: Proceedings of European Conference on Mobile Robotics, 2011, pp. 153–58.

[10] S. Laible, Y. N. Khan, K. Bohlmann, A. Zell, 3D LIDAR-and camera-based terrain classification under different lighting conditions, Autonomous Mobile Systems, 2012, pp. 21–29.

[11] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 1–15.

[12] N. Silberman, R. Fergus, Indoor scene segmentation using a structured light sensor, in: Proceedings of the ICCV Workshops, 2011, pp. 601–608.

[13] H. Badino, "A robust approach for ego-motion estimation using a mobile stereo platform," in First International Workshop on Complex Motion, 2004, pp. 198 – 208.

[14] A. Milella and R. Siegwart, "Stereo-based ego-motion estimation using pixel-tracking and iterative closest point," in Proceedings of the Fourth IEEE International Conference on Computer Vision Systems, 2006.

[15] M. Agrawal and K. Konolige, "Real-time localization in outdoor environments using stereo vision and inexpensive gps," in Proceedings of the 18th International Conference on Pattern Recognition, 2006, pp. 1063 – 1068.

[16] M. Agrawal, K. Konolige, and R. C. Bolles, "Localization and mapping for autonomous navigation in outdoor terrains: A stereo vision approach," in Proceedings of the Eighth IEEE Workshop on Applications of Computer Vision, 2007.

[17] P. Lenz, J. Ziegler, A. Geiger and M. Roser, "Sparse Scene Flow Segmentation for Moving Object Detection in Urban Environments", Intelligent Vehicles Symposium (IV), 2011.

[18] D. Lowe: Object recognition from local scale-invariant features. In Proc. of the International Conference on Computer Vision, pp. 1150-1157 1999

[19] D. Lowe, "Distinctive Image Features from Scale-Invariant Interest Points", International Journal of Computer Vision, Vol. 60, pp. 91-110, 2004.

[20] Jean-Yves and Bouguet: Camera Calibration Toolbox for Matlab.

[21] C. Harris, M. Stephens: A combined corner and edge detector, in Proceedings of the 4th Alvey Vision Conference, pp. 147-151, Manchester, UK 1998.

[22] Herbert Bay, Tinne Tuytelaars, Luc Van Gool: SURF: speeched up robust features, in proceeding of ECCV, Vol. 3951, pp. 404-417, 2006

[23] Luo Juan and Oubong Gwun: A Comparison of SIFT, PCA-SIFT and SURF, in International Journal of Image Processing, Volume 3, Issue 5, 2010

[24] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278– 2324, 1998.

[25] S. S. Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5 –18, 2004.