

Search



# IEEE/CSAA GNCC 2018

August 10-12, 2018, Xiamen, China

2018 IEEE/CSAA Guidance, Navigation and Control Conference

Home

Welcome Address

Organizations

Committees

General Information

Plenary Speeches

Technical Program

Author Index

Search

*Proceedings of*

**2018 IEEE/CSAA Guidance, Navigation and Control Conference**

**IEEE/CSAA GNCC 2018**
















August 10-12, 2018  
Xiamen, China

Sponsored by



Technical Sponsored by



 Dong Yiwei	<b>0656</b>	SunB12.35
 Dong Zhuoning	<b>0417</b>	SunA9.4
 Du Bin	<b>0269</b>	SatA6.3
 Du Guangxun	<b>0508</b>	SatB5.3
 Du Liang	<b>0496</b>	SunA12.45
 Du Nannan	<b>0581</b>	SatB9.6
 Du Xiangjun	<b>0518</b>	SunA12.53
 Du Xiao	<b>0249</b>	SatA11.32
 Du Yaoke	<b>0670</b>	SunB9.5
 Du Yongliang	<b>0726</b>	SatB7.7
 Duan Guang-ren	<b>0133</b>	SunA3.4
 Duan Shengqing	<b>0272</b>	SatB10.36
 Duan Xiaojun	<b>0291</b>	SunB8.5
 Duan Xinyao	<b>0480</b>	SunA12.40
 Duan Yongli	<b>0735</b>	SunB12.55

[Home](#)[Welcome Address](#)[Organizations](#)[Committees](#)[General Information](#)[Plenary Speeches](#)[Technical Program](#)[Author Index](#)[Search](#)

## Dense Mapping from Feature-Based Monocular SLAM Based on Depth Prediction

Yongli Duan, Jing Zhang, Lingyu Yang

Beihang Univ.

### ABSTRACT

In recent years some direct monocular SLAM methods have appeared achieving impressive semi-dense or dense 3D scene reconstruction. At the same time, feature-based monocular SLAM methods can obtain more accurate trajectory than direct methods, but only obtain sparse feature point map rather than semi-dense or even dense map like direct methods. With the development of deep learning, it becomes possible to predict the depth map of a scene given a single RGB image. In this paper we demonstrate how depth prediction module via deep learning can be used as a plug-in module in highly accurate feature-based monocular SLAM (e.g. ORB-SLAM). Both accurate trajectory from ORB-SLAM and dense 3D reconstruction from depth prediction can be achieved. Evaluation results show that dense scene reconstruction can be obtained from highly accurate feature-based monocular SLAM.

[Download 0735 Full Text \(PDF\)](#)[◀ back](#)

# Dense Mapping from Feature-Based Monocular SLAM Based on Depth Prediction

Yongli Duan, Jing Zhang\*, and Lingyu Yang

**Abstract**—In recent years some direct monocular SLAM methods have appeared achieving impressive semi-dense or dense 3D scene reconstruction. At the same time, feature-based monocular SLAM methods can obtain more accurate trajectory than direct methods, but only obtain sparse feature point map rather than semi-dense or even dense map like direct methods. With the development of deep learning, it becomes possible to predict the depth map of a scene given a single RGB image. In this paper we demonstrate how depth prediction module via deep learning can be used as a plug-in module in highly accurate feature-based monocular SLAM (e.g. ORB-SLAM). Both accurate trajectory from ORB-SLAM and dense 3D reconstruction from depth prediction can be achieved. Evaluation results show that dense scene reconstruction can be obtained from highly accurate feature-based monocular SLAM.

## I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) is an active research area in computer vision and robotics. Its main goal is to reconstruct three-dimensional scenes and estimate camera pose <sup>[1][2][3]</sup>. Feature-based SLAM generally obtains a sparse landmark map. Sparse maps only model parts of interest, i.e. feature points or landmarks. From the point of localisation, sparse feature point maps can be used to locate the camera or robot. But the spatial structure between several feature points cannot be inferred, so we can not achieve navigation, obstacle avoidance and other tasks which only dense maps can accomplish. Recently, methods to incorporate real-time SLAMs and depth maps obtained from depth sensors have become increasingly popular, because depth information is of vital importance in a lot of engineering applications, such as robotics, augmented reality, computer graphics and autonomous driving.

However, all kinds of depth sensors have their limitations. For example, the 3D LiDAR is very expensive, and can only provide sparse depth values. The depth sensor (e.g. Kinect) based on structured light is sensitive to light, consumes electricity and has a short range. Finally, stereo cameras require long-range baselines and require precise calibration for accurate triangulation, which requires a lot of calculations and often fails in feature-deficient areas.

Due to these limitations, there has been a strong interest in semi-dense and dense SLAMs based on monocular cameras that are small, low cost, energy-efficient, and ubiquitous in consumer electronics. The objective of these methods in [5][6] is to reconstruct scenes in real-time using single camera and

estimate the depth map of the current perspective by stereo matching between some adjacent frames. A necessary condition for these methods to work well is that the camera must translate in space. Stereo matching relies on the intensity-invariance assumption or keypoints extraction and matching.

A major limitation of the monocular SLAM method is the inherent scale-ambiguity. In fact, even if camera pose estimation and scene reconstruction are accurate, the absolute scale of this reconstruction is still ambiguous in nature. This limits the use of monocular SLAM in most applications such as robotics and augmented reality. Some methods use object detection techniques to match the scene with a predefined set of three-dimensional models to solve the problem, but these methods fail if there are no known shapes in the scene. Another major limitation of monocular SLAM is pose estimation under purely rotating camera motion. In this case, stereo vision estimation cannot be used due to the lack of a stereo baseline, resulting in failed tracking.

Recently, a new method was proposed to deal with depth prediction from single images through learning methods. In particular, the use of an end-to-end convolutional neural network <sup>[7]</sup> has demonstrated the possibility of obtaining high-resolution and high-accuracy depth maps, even in scenes that lack mono-clues. One advantage of the deep learning methods is that the absolute scale can be learned from the data so that it can be predicted from a single image without requiring scene-based assumption or geometric constraint.

From the above narrative, we can see that depth prediction based on deep learning and monocular SLAM can be organically combined to achieve complementarity. The main contribution of this paper is that we demonstrate depth prediction based on deep learning can be used as a plug-in module in sparse visual SLAM (e.g. ORB-SLAM) to get accurate trajectory and dense point cloud.

## II. RELATED WORK

In this section we review monocular SLAM and depth prediction that we integrate in this paper.

Yongli Duan is with School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100083 China (e-mail: dylyongli@buaa.edu.cn).

Jing Zhang is with School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100083 China (e-mail: zhangjing2013@buaa.edu.cn).

Lingyu Yang is with School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100083 China (e-mail: yanglingyu@buaa.edu.cn).

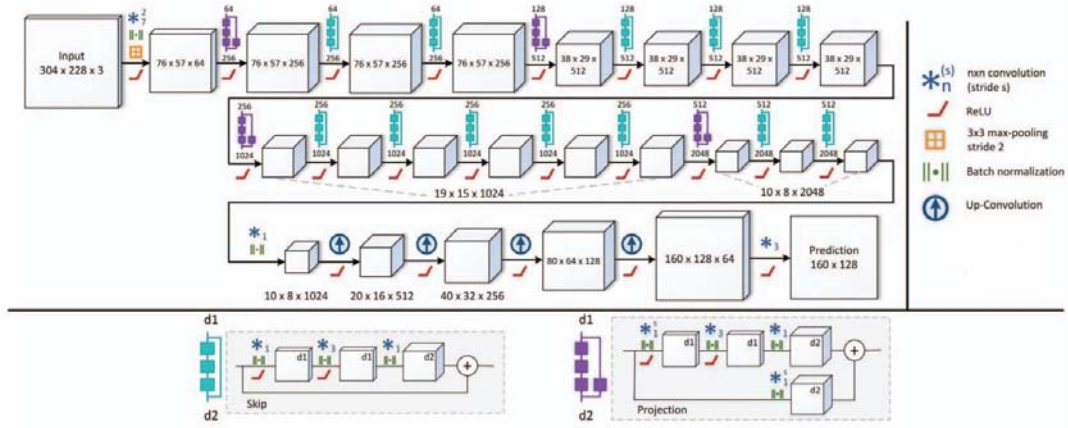


Figure 1. Deep neural network architecture<sup>[7]</sup>.

### A. Monocular SLAM

From a methodological viewpoint, monocular SLAM can be divided into either feature-based and direct. As for feature-based method, ORB-SLAM<sup>[8]</sup> can obtain the most accurate pose estimation. This method depends on sparse ORB features from the input single image. Hence ORB-SLAM can only get sparse feature point maps. This leads ORB-SLAM to be used only for localization and not for other applications such as navigation. Meanwhile, direct methods use all the information of the image. These methods track most or all the pixels in the image. So they can achieve semi-dense or dense reconstruction. However, how to simultaneously estimate dense structure and motion is still an open problem. Direct method has lower tracking accuracy than feature point method.

### B. Depth Prediction

Stereo vision uses paired images for 3D scene reconstruction, which is a traditional method of depth e. In the single-view case, most approaches rely on SFM. These methods usually make strong assumptions on scene geometry. For example, Saxena et al. [9] estimated the absolute scale of local and global image patches and inferred the depth map using a Markov Random field model. Another class of analogous work comprises non-parametric approaches. These methods typically perform feature-based matching (e.g. GIST) between a given RGB image and the RGB-D images database, we can retrieve the depth counterparts which are combined to produce the final depth map.

More recently, revolutionary progress in the field of deep learning drove research of using CNNs for depth prediction. Since depth prediction is very similar to regression task, almost all works built on the most successful architectures of ImageNet Large Scale Visual Recognition Challenge<sup>[10]</sup>, such as AlexNet<sup>[11]</sup> or ResNet<sup>[12]</sup>. Eigen et al. [13] have been the first to propose a two-stack convolutional neural network, with the first stage producing the global coarse output and the other refining local details. Another direction for improving the quality of predicted depth maps is the combined usage of a deep CNN and a continuous conditional random field.

## III. METHODOLOGY

In this section, we describe the pipeline of dense mapping from highly accurate ORB-SLAM based on depth prediction, as shown in figure 2. We can see from the figure that depth prediction is used as a plug-in module of ORB-SLAM. Hence, our system can take both advantages of ORB-SLAM and deep learning, achieving state-of-the-art accurate trajectory and dense map. We will review the three parts used in our system : ORB-SLAM for highly accurate localization, deep learning for depth prediction and methods for constructing dense map.

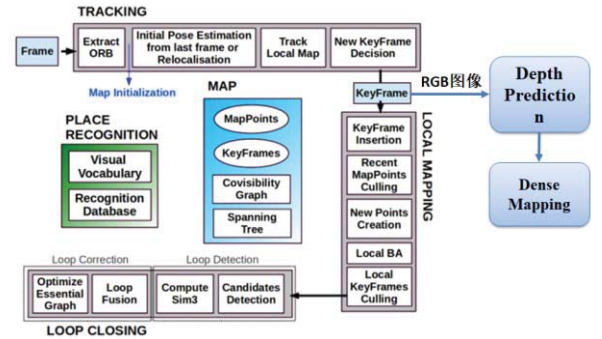


Figure 2. Pipeline of dense mapping from ORB-SLAM based on depth prediction.

### A. Highly Accurate ORB-SLAM

The ORB-SLAM is used to estimate the poses of keyframes. All its tasks utilize ORB features. The ORB feature improves the FAST corners to make it direction-sensitive and uses binary descriptors to speed up matching. ORB is at two orders of magnitude faster than SIFT, and is rotation invariant and resistant to noise. ORB-SLAM consists of three system threads: tracking, local mapping and loop closing<sup>[5]</sup>. Tracking thread process all images from the camera get the camera pose and decide whether to insert a new keyframe. Local mapping thread process every new keyframe and performs local bundle



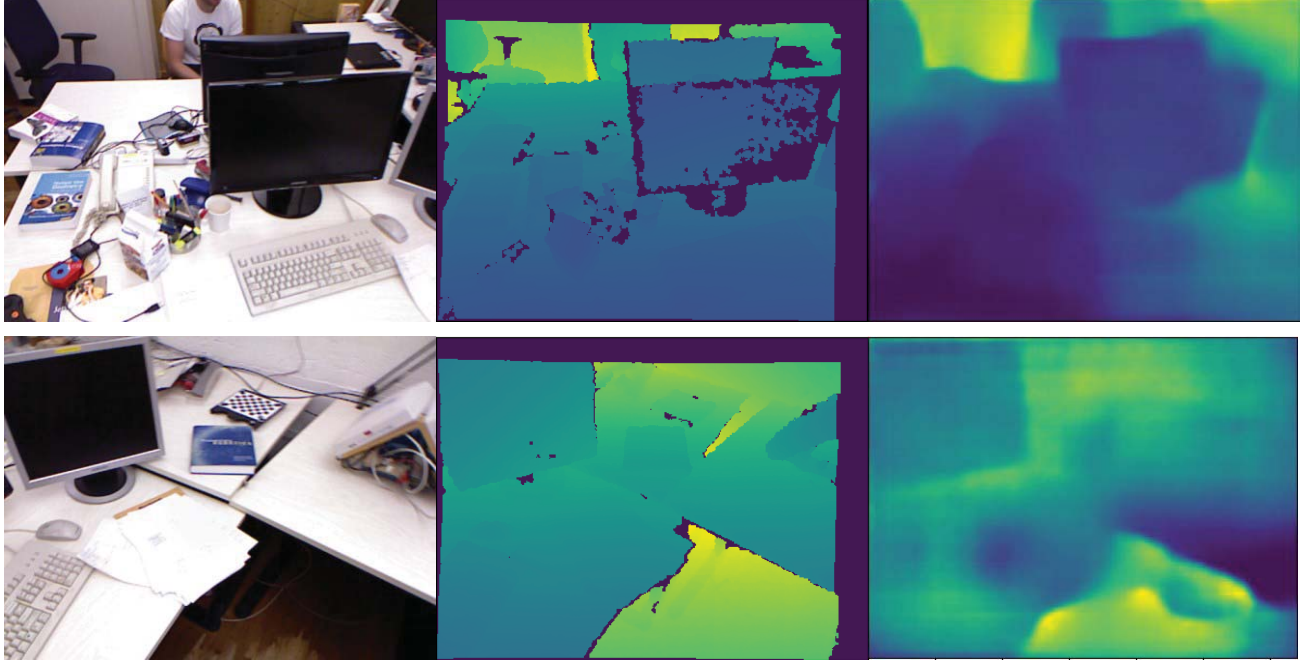


Figure 4. Depth prediction.

The first column is RGB images. The middle column is groundtruth. The last column is our results.

adjustment to achieve local optimal reconstruction in the nearby of the camera pose. The loop closing thread query keyframe database and detect loop closure. Loop closure detection can eliminate accumulated errors and construct globally consistent trajectories and maps.

### B. Deep Neural Network for Depth Prediction

The CNN architecture estimates the depth map of a scene given a single RGB image is shown in figure 2. The method comes from Laina et al. [7]. The architecture depends on ResNet-50 [12] and replace the fully-connected layers with novel up-sampling blocks shown in figure 3.

As we know, fully-connected layer introduces billions of parameters. Replacing the fully-connected layer with an up-sampling layer not only reduces the number of parameters but also gets a high resolution depth map (the predicted depth map is almost half the resolution of the input image). The predicted depth map from DNN can be seen in figure 4.

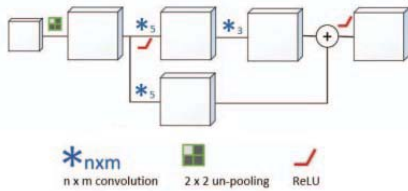


Figure 3. up-sampling blocks

Table I list state-of-the-art results of depth prediction on the NYU-Depth-v2 dataset. Samples indicates the number of depth value which is used for our CNN architecture. We can conclude that sparse depth values can improve the results greatly. RMSE stands for root mean squared error. The unit of RMSE is meter.  $\delta_i$  stands for the percentage of predicted values where the relative error is under a threshold.

Specifically,

$$\delta_i = \frac{\text{card}\left(\left\{\hat{y}_i : \max\left\{\frac{\hat{y}_i}{y_i}, \frac{y_i}{\hat{y}_i}\right\} < 1.25^i\right\}\right)}{\text{card}(\{y_i\})},$$

where  $y_i$  and  $\hat{y}_i$  are the groundtruth and predicted value respectively, and  $\text{card}$  is the set's cardinality.

TABLE I. DEPTH PREDICTION ERRORS

Method	Samples	RMSE	$\delta_1$	$\delta_1$	$\delta_3$
Laina et al. [7]	0	0.573	81.1	95.3	98.8
Ma, Fangchang <sup>[4]</sup>	200	0.230	97.1	99.4	99.8

### C. Dense Mapping

For real-time considerations, we only use keyframes to construct dense map. Based on the pixel depth value and camera intrinsics, we can calculate the position of any pixel in the camera coordinate system. Camera intrinsics which can be obtained from calibration is denoted as follows:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

Unlike the ORB-SLAM, which can only get the depth value of sparse feature points, we use depth prediction to get the depth value of all pixels described in III-B. ORB-SLAM can obtain the accurate pose of keyframes. Based on these poses, the position of the pixels in the world coordinate system

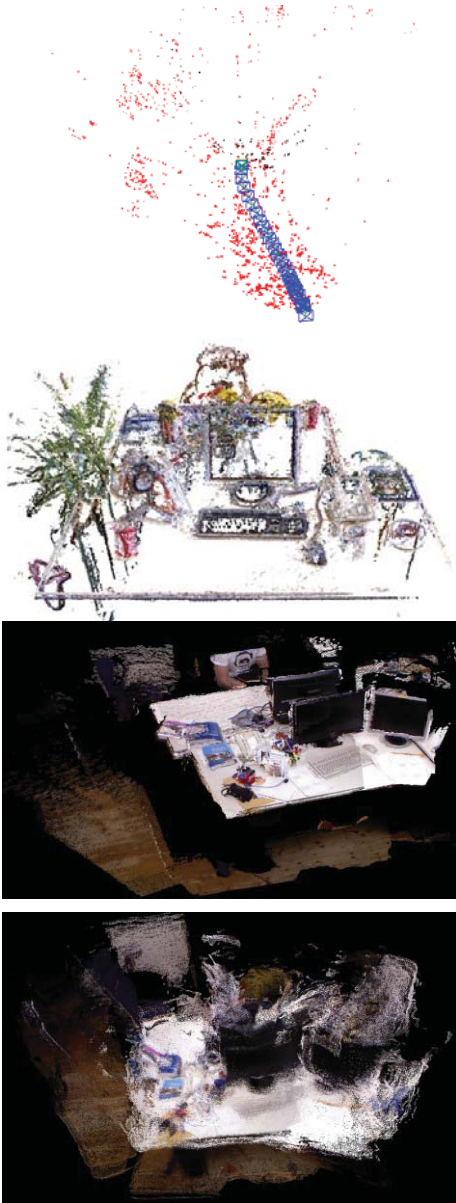


Figure 5. Dense reconstruction.

The first is sparse landmark from ORB-SLAM. The second is semi-dense reconstruction from LSD-SLAM. The third is ground truth point cloud. Real depth values come from datasets. The last is our predicted point cloud. Depth values come from depth prediction module based on deep learning.

can be calculated. The formula is as follows , where  $(u, v)$  stands for pixel coordinates,  $d$  stands for depth value of pixels,  $P_c$  stands for coordinates in camera frame,  $P_w$  stands for coordinates in world frame,  $T_{wc}$  stands for transformation matrix from camera frame to world frame.

$$P_c = \begin{bmatrix} \frac{(u - c_x) * d}{f_x} & \frac{(v - c_y) * d}{f_y} & d \end{bmatrix}^T, \quad (2)$$

$$P_w = T_{wc} P_c. \quad (3)$$

Besides, we take three additional procedures to improve the performance:

- ◆ Remove the point where the depth value is too large, because error at these points is relatively large.
- ◆ Use statistical filter to remove outliers. The filter counts the distribution of the distance values of each point from its nearest  $N$  points, and removes points whose average distance is too large. In this way, we keep the points that gather together and removed the isolated noise points.
- ◆ Use voxel filter for downsampling. Due to the overlap of view fields from multiple perspectives, there will be a large number of closely related points in the overlapping area. The voxel filter guarantees that there is only one point in a certain size cube, which is equivalent to downsampling the three-dimensional space, which can save a lot of storage space.

#### IV. EXPERIMENTS

In this section, we evaluate the performance of our method. Our system embed depth prediction which is based on deep learning within highly accurate ORB-SLAM. We use sequence from TUM RGB-D SLAM [14]. In our experiments, we train our CNN model on the indoor sequences of the NYU Depth v2 dataset, to test the generalization capability of our model to unseen environments. It is noteworthy that the scene of the training dataset are quite different from the testing dataset. Figure 5 lists the results.

We test the system using a laptop with i5-6300HQ CPU. The average time of tracking module is about 30 ms per frame. The time consuming deep learning module runs in a separate time. Therefore it does not affect the real-time nature of the system.

#### V. CONCLUSION

We have demonstrated that the combination of ORB-SLAM and depth prediction based on deep learning is a promising direction to solve the inherent limitations of traditional monocular SLAM. One of the significance is that we can obtain dense point clouds through feature-based monocular SLAM. Features (e.g. ORB) is to a certain degree invariant to illumination and viewpoint, so our system is more robust and accurate than direct methods. Thus, our system obtain both advantages of feature-based method and direct method.

The main limitation of our system is the accuracy of depth prediction. The RMSE of the CNN model used in our system is 57 cm, which is relatively high compared with ORB-SLAM localization accuracy. Fangchang [4] propose a new model which combine RGB images and sparse feature points obtained from monocular SLAM. Our future work can use this model to achieve better 3D scene reconstruction.

#### REFERENCES

- [1] Cadena, Cesar, et al. "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age." *IEEE Transactions on Robotics* 32.6(2016):1309-1332.

- [2] Durrant-Whyte, H, and T. Bailey. "Simultaneous Localization and Mapping: Part I." *IEEE Robotics Automat Mag* 13.3(2006):108-117.
- [3] Bailey, T, and H. Durrantwhyte. "Simultaneous localisation and mapping (slam) part 2: State of the art." *IEEE Robotics & Amp Amp Automation Magazine* 13.3(2006):108-117.
- [4] Ma, Fangchang, and Sertac Karaman. "Sparse-to-dense: Depth prediction from sparse depth samples and a single image." arXiv preprint arXiv:1709.07492 (2017).
- [5] Mur-Artal, Raúl, J. M. M. Montiel, and J. D. Tardós. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System." *IEEE Transactions on Robotics* 31.5(2015):1147-1163.
- [6] Engel, Jakob, T. Schöps, and D. Cremers. "LSD-SLAM: Large-Scale Direct Monocular SLAM." 8690(2014):834-849.
- [7] Laina, Iro, et al. "Deeper Depth Prediction with Fully Convolutional Residual Networks." Fourth International Conference on 3d Vision IEEE, 2016:239-248.
- [8] Mur-Artal, Raúl, J. M. M. Montiel, and J. D. Tardós. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System." *IEEE Transactions on Robotics* 31.5(2015):1147-1163.
- [9] Saxena, Ashutosh, S. H. Chung, and A. Y. Ng. "Learning Depth from Single Monocular Images. " *Advances in Neural Information Processing Systems* 18(2005):1161-1168.
- [10] Russakovsky, Olga, et al. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115.3(2014):211-252.
- [11] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." *International Conference on Neural Information Processing Systems Curran Associates Inc.* 2012:1097-1105.
- [12] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." (2015):770-778.
- [13] Eigen, David, C. Puhrsch, and R. Fergus. "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network." (2014):2366-2374.
- [14] Sturm, Jürgen, et al. "A benchmark for the evaluation of RGB-D SLAM systems." *Ieee/rsj International Conference on Intelligent Robots and Systems IEEE*, 2012:573-580.