Autonomous Salient Visual Feature Detection Using Salient Points from both RGB and HSV Color Spaces for Visual SLAM in Indoor Environments

Yong-Ju Lee
Korea University, Dept. of Mech. Eng.
yongju_lee@korea.ac.kr

Jae-Bok Song
Korea University, Dept. of Mech. Eng.
jbsong@korea.ac.kr

Abstract − For successful SLAM, feature detection becomes an important issue. This paper proposes autonomous detection of salient visual features for visual SLAM in indoor environments. Visual feature candidates are generated by the SIFT keypoints and contour information. Then uniformity maps which measure the level of similarity with the candidates and entropy maps which measure the level of diversity of information are created. Uniformity and entropy maps are combined to make a saliency map which is represented as grayscale values. In the saliency map, it is possible to distinguish the salient visual features from the background. The robot estimates its pose using the detected visual features and builds a grid map of the unknown environment using a laser scanner. The positions of visual features are also represented in the grid map. Experimental results demonstrate that the algorithm proposed in this paper enables a robot to find visual features such as objects or groups of small objects with a stereo camera in unknown environments.

Keywords − Saliency Map, SIFT, SLAM, Visual Attention.

1. Introduction

If a robot moves in an unknown environment, both accurate pose estimation of the robot and mapping of the environment are equally important. Therefore, SLAM (Simultaneous Localization And Mapping) has been one of the most fundamental and challenging issues in the field of mobile robotics. Range sensors (i.e., laser scanners, sonar sensors, and IR scanners) and vision sensors (i.e., monocular and stereo cameras) are usually employed for SLAM. As both range- and vision-based schemes use features to estimate robot pose, it is evident that observation of features is the most important factor of successful SLAM. Feature extraction from the range data is usually simpler than that from the visual data. However, features that can be extracted from the range data are limited to lines and corners. On the other hand, vision sensors offer much more information than range sensors. Although vision sensors require complicated image processing to extract visual features, recent SLAM approaches tend to employ vision sensors as a main sensor.

Most vision-based SLAM methods use feature points extracted from the camera image. The keypoints obtained by the multi-scale Harris corner (MSHC) or scale invariant feature transform (SIFT) are good examples for extraction of visual feature points [1][2]. Since the extracted points are robust to scale and rotation, they are used for matching and recognizing images. However, it is inefficient to use all the points as visual features for estimating the robot pose because too many points tend to cause inefficiency of SLAM or an increase in computational complexity. Thus, a scheme for decreasing the number of points using the clustering algorithm was proposed in [3]. However, the results of feature detection using the clustering algorithm are unstable because the clustered regions in the camera image usually change momentarily.

The neuromorphic vision toolkit (NVT) was proposed for extraction of consistent visual features in [4]. The extracted features are salient regions of an image. The scheme uses local extrema using image pyramids to find salient regions within a natural scene. This scheme aims at searching visual features as humans do and it can successfully extract visual features from the background. However, it focused only on the image analysis and its application to navigation was not attempted.

This paper proposes a novel method to extract visual features that are applicable to SLAM. The proposed method finds useful visual features without any prior information and exploits them as natural landmarks to estimate the robot pose and build an accurate environment map. The proposed scheme consists of generation and evaluation of visual feature candidates. The visual feature candidates are created in the RGB (i.e., red, green, and blue) color space and are evaluated using the saliency map constructed in the HSV (i.e., three properties of color; hue (color), saturation, and value (intensity)) color space. If visual features are determined to be suitable for navigation, they are separated from the camera image and registered in the database. The robot autonomously builds a map of the unknown environment by both autonomous visual feature detection and visual feature-based EKF SLAM. A stereo camera is used for both detection of salient visual features and estimation of the robot pose and a laser scanner is used for mapping. Figure 1 shows the overall procedure of the autonomous visual feature detection algorithm.

The remainder of this paper is organized as follows. Section 2 presents generation of visual feature candidates and calculation of uniformity maps using the candidates. Section 3 deals with calculation of entropy maps and

section 4 presents the saliency map and candidate evaluation. Section 5 describes EKF-based SLAM using extracted features and experimental results are shown in Section 6. Finally, section 7 presents conclusions.

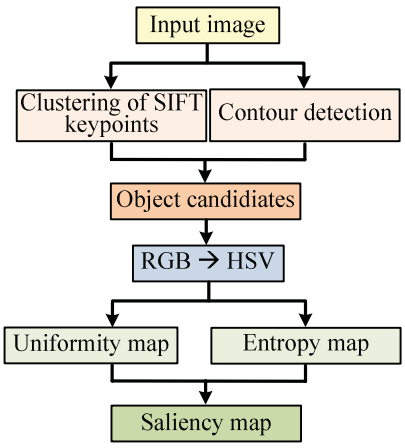


Fig. 1. Overall procedure of proposed scheme.

2. Visual Feature Candidates, Background Candidates, and Uniformity Maps

This section presents the way of obtaining of visual feature candidates and background candidates. Then the calculation of an uniformity map is discussed.

2.1 Visual Feature and Background Candidates

Visual feature candidates are used as clues to the visual features. Because we do not have any information on the visual features, it is assumed that visual features of indoor environments have complex patterns or closed outlines. Therefore, both SIFT and the contour detection algorithm are used to obtain the visual feature candidates.

SIFT extracts the feature points which are invariant to scale, rotation, and viewpoint. The region where many SIFT keypoints exist is useful for navigation because the region is easily recognized by the SIFT keypoints. The SIFT keypoints are clustered into several groups by means of the distance between keypoints in the image. Each group can be considered as a visual feature (although the keypoints of the group come from different physical objects) because the SIFT keypoints tend to exist in the space whose pattern is obvious and remarkable (not flat and monotonous).

Fig. 2. Examples of candidate generation; (a) SIFT keypoints and their clustered regions, (b) detected contours, (c) visual feature candidates, and (d) background candidates.

At the same time, the contours of visual features are detected by the Canny edge algorithm [5]. The contours of visual features can help distinguish them from the background. The detected contours and clustered regions of SIFT keypoints in the input image are shown in Fig. 2(a) and (b), respectively. The visual feature candidates, which are the collections of both the clustered regions of SIFT keypoints and the detected contours, are shown in Fig. 2(c). The background candidates, which are the counterpart of the visual feature candidates, are shown in Fig. 2(d).

2.2 Color Space Conversion of Input Image

The input image in the form of RGB is transformed into the form of HSV. The HSV color space is more intuitive and gives more information than the RGB color space because the HSV color space is similar to the human cognitive system and its three channels are not correlated. In the hue channel, all colors are represented as the values between 0 and 360. The saturation channel represents the degree of purity. For instance, dark blue and light blue are determined by adjusting the saturation channel. The intensity channel represents light information. Figure 3 shows an image which is converted from the RGB color space to the HSV color space.

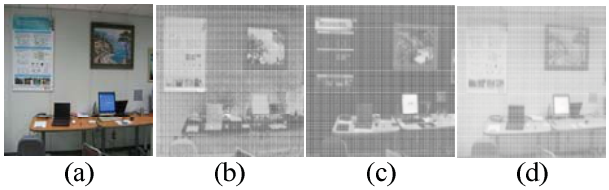


Fig. 3. Conversion of color space; (a) camera image in RGB color space, (b) hue channel, (c) saturation channel, and (d) intensity channel.

2.3 Calculation of Uniformity Maps

Uniformity of a pixel means the level of similarity between the pixel under consideration and all the pixels of the candidates. Since there are two types of candidates, the uniformity maps are also based on either the visual feature candidates or background candidates. In addition, each uniformity map is obtained in association with the hue, saturation, and intensity channels.

A. Uniformity maps using visual feature candidates

The uniformity maps using the visual feature candidates are calculated as follows. In the visual feature candidates, the histograms of the channel values of all pixels in the candidates are generated. Figure 4 shows the visual feature candidates of the hue channel and their histogram.

The weight of Fig. 4(b) is defined as the number of pixels with identical channel values divided by the number of all pixels of the candidates. The uniformity of a pixel is the weighted sum of the likelihoods between the pixel and all pixels of the candidates. It is given by

$$U_{\text{feature}}(u,v) = \sum_i w_i \cdot \exp[-\frac{(G(u,v)-i)^2}{2\sigma_u^{\,2}}] \quad (1)$$

where i is the index representing the channel value (e.g., 0 to 360 in the hue channel) of the visual feature candidates and w_i is the weight associated with i. $G(u, v)$ is the channel value of the pixel (u, v) and $U_{\text{feature}}(u, v)$ is the uniformity of the pixel (u, v) using the visual feature candidates. σ_u is an error bound of the difference between $G(u, v)$ and i. $U_{\text{feature}}(u, v)$ becomes almost zero where the difference is greater than the error bound.

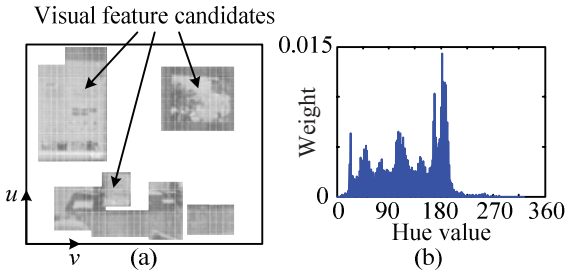


Fig. 4. Examples of feature candidates and generation of histograms; (a) visual feature candidates of hue channel, and (b) their histogram.

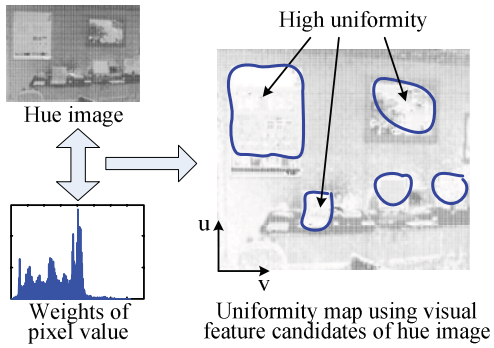


Fig. 5. Uniformtiy map using visual feature candidates in hue image.

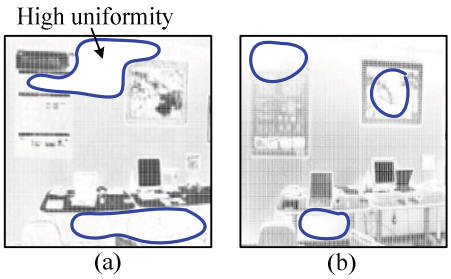


Fig. 6. Uniformity maps using visual feature candidates; (a) saturation channel, and (b) intensity channel.

Figure 5 shows the uniformity map using the visual feature candidates in the hue channel. In general, the pixels of the candidates have higher uniformity values than those of the background. The uniformity maps of the HSV color space can show different tendencies because all the channels have different properties. Figure 6 shows the uniformity maps using visual feature candidates of the saturation and intensity channels.

B. Uniformity maps using background candidates

The uniformity maps using the background candidates are also individually calculated in all the channels of the HSV color space. Figure 7 shows the background candidates of the hue channel and their histogram.

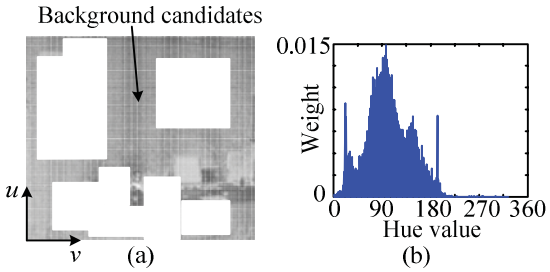


Fig. 7. Examples of background candidates and generation of histograms; (a) background candidates of hue channel, and (b) their histogram.

The uniformity of a pixel using the background candidates is calculated in a similar way to Eq. (1). The only difference is that background candidates are used instead of visual feature candidates.

$$U_{\text{background}}(u,v) = \sum_i w_i \cdot \left(\exp(-\frac{(G(u,v)-i)^2}{2\sigma_u^{\,2}}) \right) \quad (2)$$

where $U_{\text{background}}(u, v)$ is the uniformity of the pixel (u, v) using the background candidates.

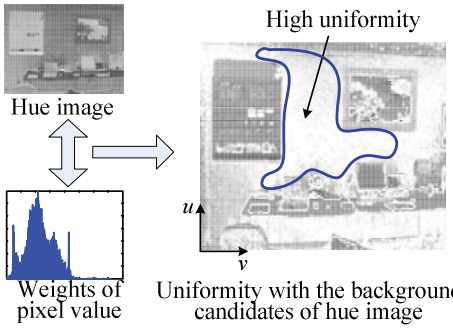


Fig. 8. Uniformtiy map using background candidates in hue image.

Figure 8 represents the uniformity map using the background candidates in the hue channel. The uniformity of other channels are shown in Fig. 9. The values of the uniformity map tend to be high at the background. If the channel values of the visual feature candidates are similar to those of the background candidates, the uniformity values based on the background candidates are high at the visual feature candidates as well as the background candidates. In this case, they are canceled out in the process of combining all the uniformity maps. The combined uniformity map is explained in the next section.

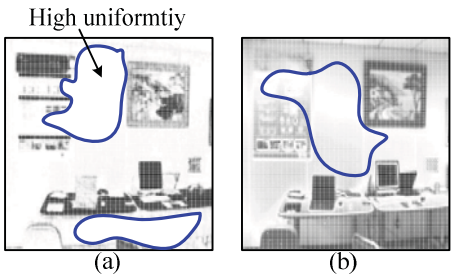


Fig. 9. Uniformity maps using backgorund candidates; (a) saturation channel, and (b) intensity channel.

2.4 Combined Uniformity Map

All the uniformity maps are combined to create a combined uniformity map. The uniformity maps based on the background candidates are subtracted from those based on the visual feature candidates because their properties are opposite and thus the combined uniformity is calculated as follows:

$$U_{\text{combined}} = U_{\text{feature}} - U_{\text{background}} \tag{3}$$

If the uniformity map based on the visual feature candidates is similar to that based on the background candidates (i.e., the case of Fig. 7 and Fig. 9), they disappear through subtraction. Figure 10 shows the combined uniformity map. The uniformity values become high at the visual features, which are described by the solid line.

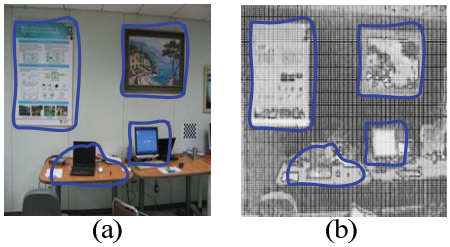


Fig. 10. Experimental results of combined uniformtiy map calculation;(a) camera image, and (b) combined uniformity map.

3. Entropy Maps

3.1 Calculation of Entropy

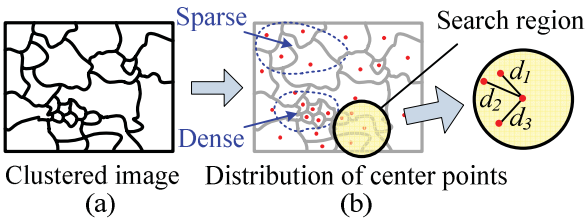


Fig. 11. Calculation of entropy; (a) clusterd image, and (b) one center point and its neighboring center points.

Entropy means the level of diversity of information. The entropy values become high as the diversity increases. In this paper, the level of diversity is related to the level of the difference of the channel values in local regions. At first, the images of hue, saturation, and intensity are clustered according to their HSV values and center points of all the clusters are generated. The entropy maps of all the channels of the HSV color space are calculated from the center points. Figure 11 shows the calculation of the entropy of an image.

The distances between the center point under consideration and the other neighboring center points are used to compute the entropy at this point as follows:

$$E_i = -\sum p_j \log_2 p_j,\ p_j = \exp(-\frac{d_j^2}{2\sigma_e^2}) \tag{4}$$

where E_i is the entropy value of the i-th center point, and d_j is the Euclidean distance between the center point and its j-th neighboring center point. d_j is represented as p_j which shows the level of closeness as the probability and p_j increases as d_j decreases. σ_e is the error bound of the distance. The entropy maps of the image are shown in Fig. 12. The entropy value tends to increase at the place where the density of center points is high. Thus, salient features whose color or intensity information is different from others have high entropy values. However, the backgrounds usually have low entropy values because their changes in color or intensity are not significant. The high entropy value is described by the solid lines in Fig.12.

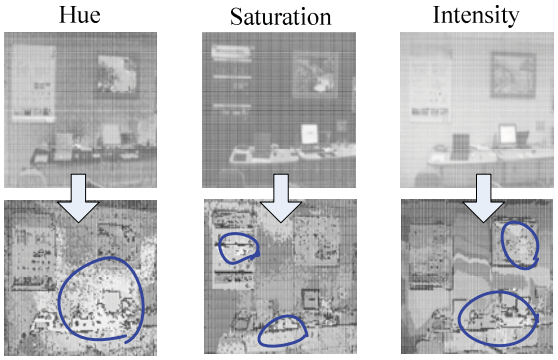


Fig. 12. Entropy maps of HSV color space.

3.2 Combined Entropy Map

All entropy maps from the hue, saturation, and intensity channels are superposed to produce a single entropy map. Figure 13 shows the combined entropy map. The pixels whose entropy values are high in all the channels have a great impact on the combined entropy map.

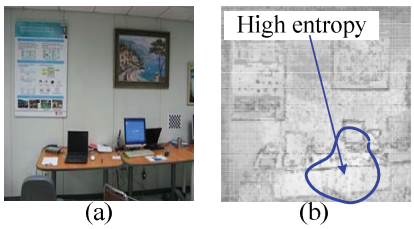
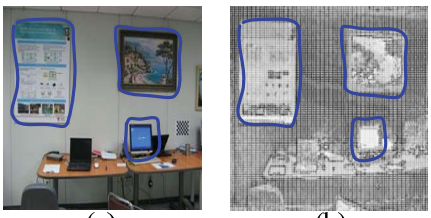


Fig. 13. Experimental results of combined entropy map calculation; (a) camera image, and (b) combined entropy map.

4. Saliency Map and Selection of Candidates

4.1 Saliency Map

A combined uniformity map of section 2.4 and a combined entropy map of the section 3.2 are multiplied to produce a saliency map. Figure 14(a) is the input image and Fig. 14 (b) shows the saliency map of Fig. 14(a). The saliency values are high at the visual features, which are denoted by the solid lines.

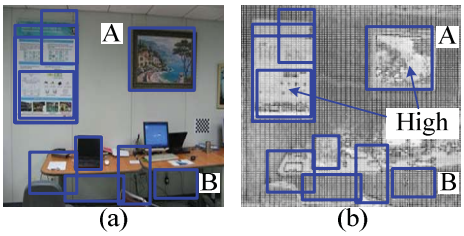


(a) (b)

Fig. 14. Examples of saliency map generation; (a) camera image, and (b) saliency map.

4.2 Selection of Visual Feature Candidates

This section introduces the method to select good candidates among all the visual feature candidates. The main idea is to find the candidate whose average saliency is much greater than that of the surroundings. All the visual feature candidates are shown in Fig. 15. In candidate A, the average of the saliency outside candidate A is much smaller than the average saliency inside candidate A. On the other hand, in candidate B, the average saliency outside candidate B is approximately the same as that of inside candidate B. Therefore, candidate A is selected as a real feature, but candidate B is discarded. Figure 16 shows the results of the proposed visual feature detection method using other images. The camera images, their saliency maps, and the detected visual features are shown in Fig. 16.



(a) (b)

Fig. 15. Selection of visual feature candidates; (a) visual feature candidates, and (b) saliency map.

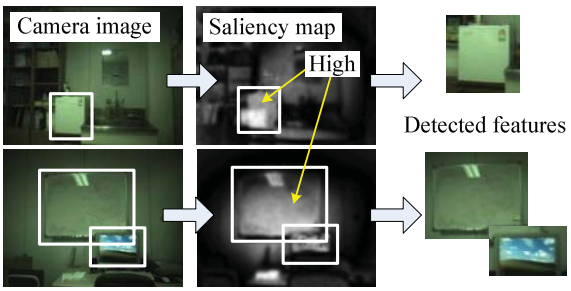


Fig. 16. Camera images, saliency maps, and detected visual features.

5. EKF-based SLAM

The EKF (Extended Kalman Filter) algorithm has adopted for visual SLAM. In EKF-based SLAM, the robot pose and the feature positions are stored in a state vector represented as X, and the position uncertainties of elements of the state vector are stored in a covariance matrix denoted as P. The state vector and the covariance matrix are updated recursively through sensor measurements.

At the prediction stage, the state vector and its covariance matrix at time t are obtained as follows.

$$\hat{X}_t^- = f(\hat{X}_{t-1}, u_t, t) + w_t \tag{5}$$

$$P_t^- = F_x P_{t-1} F_x^T + F_u Q_t F_u^T \tag{6}$$

where $\hat{X}_t^-$ and P_t^- are the predictions of the state vector and its covariance matrix at time t, respectively, and u_t is the displacement of the robot between time t-1 and time t. The vector w_t represents the process noise and Q is the covariance matrix of w_t. The matrices F_x and F_u are the Jacobian matrices of the motion model $f(\cdot)$ with respect to the state vector and the displacement u_t, respectively.

The state variables, the robot pose and landmark positions and the covariance matrix of the state vector are updated by the measurement of the sensor. In this paper, the measurement is obtained from object recognition in the form of a relative range and a relative angle of the object from the robot. The state vector X and its covariance matrix P at time t are updated as follows.

$$K_t = P_t^- H_t^T (H_t P_t^- H_t^T + R_t)^{-1} \tag{7}$$

$$\hat{X}_t = \hat{X}_t^- + K_t (Z_t - \hat{Z}_t) \tag{8}$$

$$P_t = (I - K_t H_t) P_t^- \tag{9}$$

where K_t represents the Kalman gain, and H_t is the Jacobian matrix of the sensor model with respect to the state vector. The error on the pose of the robot due to disturbances is compensated by the Kalman gain which is proportional to the difference between predictions and measurements. If none of landmarks are matched, only the robot pose is calculated by the motion model and the uncertainty of the robot pose increases.

6. Experiments

Various experiments were performed using a robot equipped with a stereo camera and a laser scanner. The camera is used to recognize of visual features and the laser scanner is used to build a grid map of the environment. The experimental environment consists of three rooms and several visual features exist along the walls. The total area of the environment is 10m x 10m. The grid size of the constructed map using SLAM is 10cm x 10cm.

Figure 17 illustrates the mapping process of the experimental environment using both the proposed visual feature detection and the EKF-based SLAM. Figure 17(a)-(e) shows the robot, the scenes from the camera, and the recognized visual features. Figure 17(f) is the CAD data of the environment for comparison. The robot moves

in the environment, builds the grid map and marks the visual features in their own positions. Pictures, a refrigerator, and books on the table are detected as visual features. If the robot cannot detect any visual features, odometric errors are accumulated. However, if the robot recognizes the previously registered visual feature, the accumulated errors can be removed. The SIFT keypoints are used for data association. When the robot arrives at the initial position, the average errors on the position and the orientation are ± 20cm and ±5°, respectively.

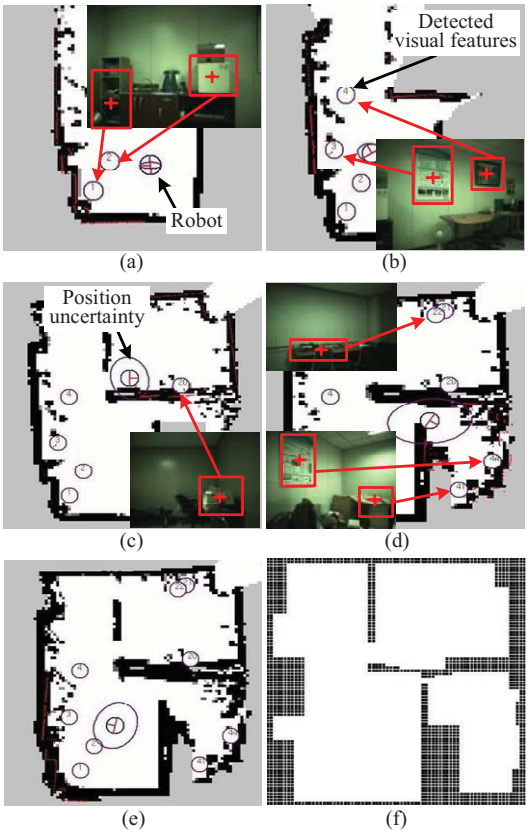


Fig. 17. Indoor SLAM with autonomous object registration.

Figure 18 shows an example of eliminating the odometric error by recognizing the visual features. During the navigation, the estimated robot pose is moved to the left from its real position, which is inferred from Fig. 18(a). In Fig. 18(a), the identical wall is depicted side by side. In Fig. 18(b), however, the robot recognizes the registered feature. As the result, the robot pose is recovered and the two walls become a single wall as shown in Fig. 18(c).

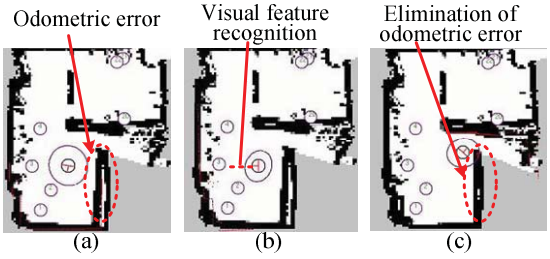


Fig. 18. Elimination of odometric error.

7. Conclusions

Visual features are useful for indoor navigation of a mobile robot because they are easily found in indoor environments. This paper proposed the autonomous detection of visual features which can be exploited for SLAM. From this research, the following conclusions were drawn.

1. The proposed visual feature extraction can autonomously detect visual features such as objects (pictures and posters) or groups of small objects (books on the table) without human interference.
2. The number of extracted visual features is not as many as that of primitive features such as corner or lines. Therefore, the proposed algorithm does not increase computational complexity as much as the primitive features.
3. The experimental results of the proposed scheme and its application to SLAM demonstrate that the proposed algorithm can improve the autonomy of the visual SLAM.

Acknowledgements

This paper was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea.

References

[1] Z. Lin, S. Kim, and I. S. Kweon, "Recognition-based Indoor Topological Navigation Using Robust Invariant Features," *Proc. of Int' Conf' on IROS*, Alberta, 2005.

[2] D.G. Lowe, "Distinctive image features from scale invariant keypoints," *Int'l Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[3] R.H. Luke, J.M. Keller, M. Skubic and S. Senger, "Acquiring and Maintaining Abstract Landmark Chunks for Cognitive Robot Navigation," *Proc. of IEEE/RSJ Int. Conf. on IROS*, Alberta, 2005.

[4] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "On the usefulness of Attention for Object Recognition," *Workshop on Attention and Performance in Computational Vision*, pp. 96-103, 2004.

[5] P. Bao, L. Zhang, and Xiaolin Wu, "Canny Edge Detection Enhancement by Scale Multiplication," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.27, no.9, pp.1485-1490, 2005.