Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

Map Building for Cluttered Environment

Yunusa Ali S., M. Hamiruce Marhaban,Abd Rahman Ramli, Siti A Ahmad, Habibu Rabiu and Arash Toudeshki
Dept. Electrical and Electronic Engineering,
Faculty of Engineering, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia

Abstract—This paper presents a novel approach to scene localization and mapping in indoor environments from the concept of the Image Bag of Words (BOW) technique, where a group of native feature descriptors represents images and are subsequently transformed to a separate set of image words. This approach uses the famed algorithm called Scale invariant Feature Transform (SIFT). To extract distinctive invariant feature for reliable matching, we normalized the images as illumination changes affect the feature extraction. To achieve robust and efficient matching the environment is modelled. Clustering is shown to be appropriate for quantizing these descriptors into clusters based on selected threshold. In this work, we developed an efficient SLAM using images captured from a highly cluttered background. The result indicates a promising trend in using the camera for SLAM implementation.

Keywords—Image Bag of words (BOW), Scale Invariant Feature Transform (SIFT), Clustering, Simultaneous Localization and mapping (SLAM), Images, Environment

I. INTRODUCTION

In recent years, there has been an increase interest in SLAM based on the bag of words (BoW). The technique for map building with one camera for robotic applications is gaining more popularity among many researchers [1,2,3,4,5] It successfully navigates through visually poor environments despite the inconsistent camera movement, tainted position estimates, whole disorientation, or large accumulated errors.

A map of the scene that has been explored is constructed, refined and improved once places measure are re-visited. The most interesting thing about SLAM based on BoW is its ability to navigate and build a map of even troublesome environments measure [6]. Based on the review on bag of word algorithm, the codebook is generated using K-means to cluster centers of the extracted descriptors. However, it fails to code other informative region, and leads to unsatisfying results as only a feature is selected to represent the centre of the cluster, however, this is a major problem or limitation to the approach. The aim of this study is to improve the Algorithm by designing a model that gives a better performance with robust matching and recognition. The novelty here, is the robustness of modelling the environment to guide the mapping process that leads to a better matching of the clusters. The system is tested on a dataset, which consist of 420 images of the environment.

II. METHODOLOGY

In image classification, visual dictionary building and classifier coaching are two core units that are performed offline on image database. Wordbook construction, entails clustering the similar image attributes into distinct groups. The main idea is to take a series of images from the corpus of different object categories and form cluster that represent such groups. K-means algorithms is the popular method usually employed for the clustering [7].

A. Environment/database

All the image sequences included in the database were acquired from the Electrical and Electronics Engineering Department corridor, of the Universiti

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

Figure 1. Exemplary Pictures taken to Present the surroundings

Putra, Malaysia. The images cover the main entrance to the department, and extend to the departmental office, kitchen, meeting room, and toilet which are all connected by the corridor as shown in Fig 1. A general map of the environment is also presented in Fig 2. Places like corridor, can be regarded as public, which implies that various people and activities may be present. Furthermore, other items such as furniture, pigeon cabinets, notice board, floor and uneven illumination have chartered the scene, thus more challenging for the algorithm.

TABLE I. Parameters and Settings

Frame rate	30fps
Resolution	315x240
Exposure	Auto
Focus	Auto
Height	100cm

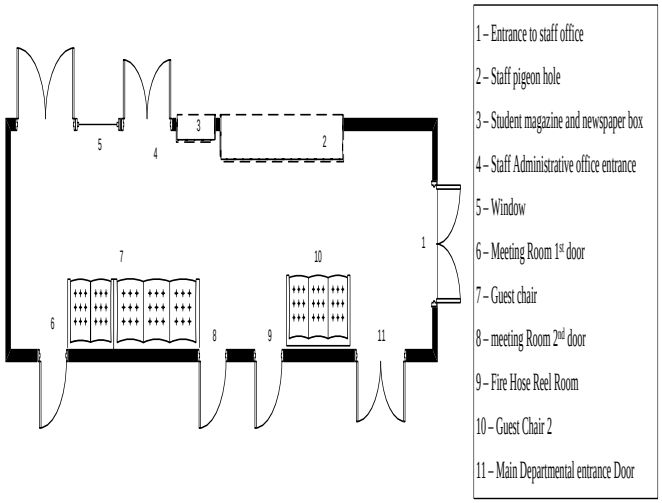

Figure 2. General Map of the environment

The database used in this work was recorded using a digital Camera mounted on a camera stand, the digital camera acquired image at a speed of 30 frames per seconds to generate database of the environment (Fig 3). The camera was mounted at a height of 100cm. Detailed parameters and settings are shown in Table I.

B. Feature Extraction

Generally, feature extraction is the first step into an image classification unit. Features are salient points in the image which make it distinctive and thus easy to capture. Two images with similar features are most likely to have come from the same location. The Scale Invariant Feature Transform (SIFT) [8] The detector and descriptor system provides a good start line. This has been utilized in several previous works and shown to be fairly sturdy. Features extracted using SIFT are invariant to image scale and orientation and therefore suitable for the task at hand. SIFT algorithm is protected beneath patent, but it is liberal to use for tutorial functions, therefore for this project we tend to use it in map building [9].

Firstly, we use a pyramid of Difference of Gaussian (DoG) in order to make our feature invariant to scale and orientation. Secondly, we determine from the result of DoG, a stable keypoints with a more accurate model. Thirdly, the orientation information is then assigned to each keypoint image gradient direction. Lastly, for illumination variation and

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

local image gradient then transforms to stable features [10].

The continuous two dimension difference of Gaussian convolution kernel is mathematically given as:

$$f(\varepsilon,k,\sigma)=ae^{-\varepsilon}-\frac{a}{k^2}e^{-\varepsilon} \quad (1)$$

Where k is a constant bigger than one which scale the standard deviation, and $a=\frac{1}{2\pi\sigma^2}$

While the magnitude and orientation of the gradient

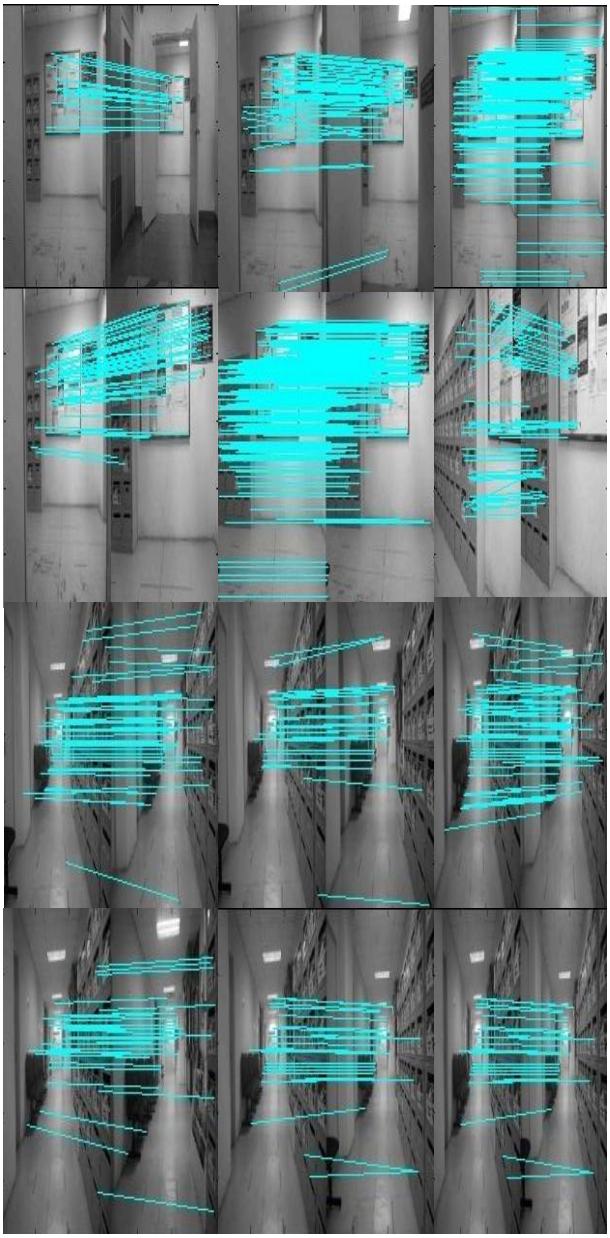

Figure 4. Sample of matched images

Figure 3. Photographic view showing the data acquisition process using a digital camera.

are represented as A_{ij}, M_{ij} and R_{ij}

$$M_{ij}=\sqrt{\left(A_{ij}-A_{i+1,j}\right)^2+\left(A_{ij-Ai,j+1}\right)} \quad (2)$$

$$R_{ij}=\arctan 2\left(Aij-A_{i+1,j}-Aij\right) \quad (3)$$

Employing these procedures leads to correction of illumination changes [11].

C. *Dictionary Building*

Building the Dictionary is by applying agglomeration on the extracted SIFT features from the training pictures. K-means algorithm attempt to bring together similar image patches into one group. Thus allow extracted features to be group into n distinct groups. Next, the centroid of each group is then determined and used in building the visual dictionary. Therefore the list of image words is selected in such a way that identical mark view in different pictures ought to map to identical word representation. Two pictures with several image words in common have several similar-looking features, and thus probably pointout identical place. Determining number of cluster is always a difficult task, however we have device a model which help in projecting the appropriate number of cluster for a giving training set, the least square technique was implemented, from the curve fitting approach.

Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013

ISBN: 978-981-07-6261-2

III. EXPERIMENTAL RESULTS

In this section we demonstrate novel strategy base on the concept of the BOWSLAM to generate a correct and stable map, the information set are unit process offline on a laptop with Intel core i3 central processor M380 @ 2.53GHZ. The mapping strategy has been enforced and tested in indoor environments. The surroundings model is obtained within the off-line. These pictures were captured, employing a digital camera to generate the database. The pictures were captured from a completely different position, The length of our database consists of 420 image frames, each frame has spatial dimension of 315x235. Fig. 4 show a sample of matching image between the query image and prime stratified area image.

The dataset demonstrates that the approach base on the concept of BOW can build an accurate and stable map of typical indoor environment, when loop closure was detected the position of the final and the initial landmark connected and terminates the map building. Fig. 5. shows the constructed map.

The most obvious finding that emerges from this study is that, not all the features that were extracted from the image are relevant to the mapping task. Therefore, in order to strenghten the discriminatory capability of our feature set, the less important features need to be removed. One of the ways to do that is by setting a reasonable threshold that will force all less informative features out of the feature set. Using global threshold does not always perform well, as some features are more discriminating than others. Traditionally, trial and error method is used in determining the threshold value, however this method is time consuming. In this work we proposed a simple and robust method using linear regression that automatically determines the most appropriate clusters for a given data set. For a scene recognition under investigation, selected parameters are passed to which in turn gives the appropriate cluster numbers (4).

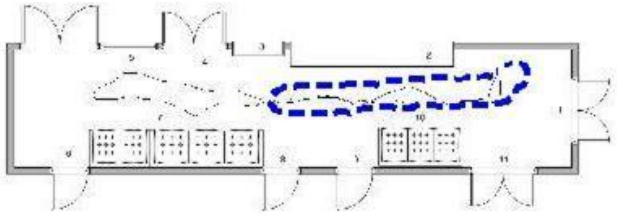

Figure 5. Showing map of the environment

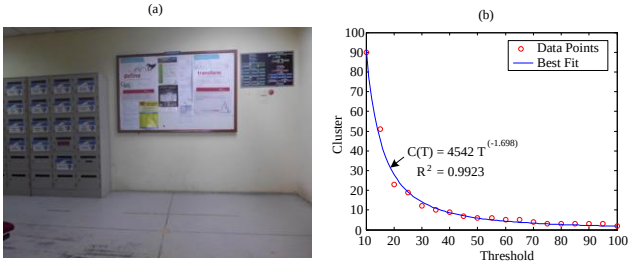

Figure 6 The example image shows the good fit base on cluster and threshold.

$$C(T) = \alpha T^{-\beta} \qquad (4)$$

Where C is cluster, T is variable threshold, coefficient α varies between ma minimum of 266.2 to ma maximum of 13970 and coefficient β varies between minimum of 0.886 to maximum of 2.137 with median of R-squared of 0.9946 for all fits

IV. Validation

The proposed (4) is fitted to the data of thresholds based on cluster which was extrcated via selected image, when the coefficient α has the optimum value of 45542 and the optimal value for the coefficint β was 1.698. The R-square of 0.9923 determined the goodness of fit to the datapoints which is presented in Fig 6(b). It must be noted that for different image the optimal coefficients α and β have different value and need to employ the suggested least-square technique to difine them. However, the general behaviour of the fit curve is performing as it was proposed in (4).

V. CONCLUSION

The paper described the proposed novel approach based on the concept of BOWSLAM, a scheme for offline SLAM using a single camera, it shows an accurate and stable map of an environment. The new approach base on the concept of BOW algorithm is used to represent each image. Experimental results using indoor images show that

ISBN: 978-981-07-6261-2

tracking accuracy depends on the characteristics of the environment. As it was mentioned in Ref. [14], it was proved that clustering can be avoided entirely once an appropriate dictionary for the environment is formed, in some environment a general purpose dictionary could be used to avoid the need to cluster entirely. The loop closure shows that this approach has solved the problem of data association.

ACKNOWLEDGEMENTS

The authors would like to thank MohammadReza Shoorangiz a postgraduate student in University Putra Malaysia for his valuable contribution in this project.

REFERENCES

[1] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1403–1410, 2003

[2] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1052–1067, 2007.

[3] L. A. Clemente, A. J. Davison, I. Reid, J. Neira, and J. D. Tardós, "Mapping large loops with a single hand-held camera," in *Robotics: Science and Systems*, 2007.

[4] I. Ohya, A. Kosaka, and A. Kak, "Vision-based navigation by a mobile robot with obstacle avoidance using single-camera vision and ultrasonic sensing," *Robotics and Automation, IEEE Transactions on*, vol. 14, no. 6, pp. 969–978, 1998.

[5] A. J. Davison and D. W. Murray, "Simultaneous localization and map-building using active vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 865–880, 2002.

[6] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 1470–1477.

[7] D. Fillet, "A visual bag of words method for interactive qualitative localization and mapping," in *Robotics and Automation, 2007 IEEE International Conference on*, 2007, pp. 3921–3926.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] B. Deavin, "Bag of Words: Automated Classification of Images," 2010.

[10] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[11] F. Hjelmare and J. Rangsjö, "Simultaneous Localization And Mapping Using a Kinect In a Sparse Feature Indoor Environment," Linköping, 2012.