The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

A Hull Census Transform for Scene Change Detection and Recognition Towards Topological Map Building

Min-Liang Wang and Huei-Yung Lin
Department of Electrical Engineering
National Chung Cheng University
168 University Road, Min-Hsiung, Chiayi 621, Taiwan, ROC
{d94415018, hylin}@ccu.edu.tw

Abstract— **This paper presents a novel encoding method for scene change detection and appearance-based topological localization framework. The relation computation over convex hull points is used to compare the similarity between the scenes. It relies on the relative ordering of the feature strength, not directly on the feature vectors. We first deal with multiple convex hulls over the detected features and then compile statistics for coding on the hull points through a vector magnitude comparison. Finally, the hull points are formed by binary codes. The codes are suitable for scene change detection and visual place recognition by statistical analysis. The experimental results show the coding method is robust under the varying environment.**

I. Introduction

The localization and place recognition have become important issues for mobile robot applications in recent years. They are mainly used for robot navigation in the indoor environment. Some research trends include using image based technique instead of traditional range sensors, to solve the problems.

In this paper, we consider the problem of detecting scene change events using an omnidirectional camera mounted on a mobile robot. The events are then employed to construct a topological map for recognizing the nodes of the visual places in some other time. In order to let the mobile robot have the ability of autonomous navigation and learn to recognize the environment, the robot must be able to label the areas of an environment automatically. We suggest to use the scene change events to perform this task. The scene change events are typically defined as the video paragraphs segmented from a consecutive video sequence in multimedia community [7]. When such tasks are applied to robotic applications, a topological map is commonly used to efficiently represent each scene change location [9].

To establish an appearance based topological map, we design a binary code transform for scene representation called “Hull Census Transform” (HCT). The method is used to handle the scene change conditions, and also takes the varying environment into account while the illumination or small objects change at the same place. This is helpful for further topological map building and visual place recognition, especially for the catadioptric vision sensors. The proposed technique is fast and robust under illumination change. It enables the mobile robot to recognize scenes based on the image appearance and automatically add nodes into the existing topological map.

For example, if there are two image features, one lies on the corner of a window, the other lies on the corner of a door. The illumination is changed from cloudy to night, and the two features can still be detected in the same location. In this example, the values of each image feature vector is changed in the two weather conditions, thus this will affect the feature matching. However, we compile the two features to two HCT bits by calculating the relative vector length between them. The HCT code is not changed with two different weather conditions. The advantage of HCT is that it can tolerate a variation of feature vectors under varying environment.

In this work, we make the following contributions:

1) We propose a novel HCT descriptor for scene representation, which is in average about $10-30$ bit codes for an image frame (one hull case). It contains sparse data with respect to image features or images themselves for visual place representation.
2) The scene change detection problem for topological map building is tackled efficiently using the proposed HCT descriptor.
3) The proposed visual place recognition is robust under the varying environment, in particularly for the omnidirectional vision system.

For clear explanation, we define the place recognition problem for a mobile robot as follows:

- a video sequence for place descriptors and topological nodelists construction,
- the other sequences of the same places with different paths for recognition.

This work will test the place recognition based on the HCT encoding approach.

II. Related Work

In mobile robotics, topological maps provide a concise description of the environment for navigation. Angeli et al. [2] represented each topological node using the *bag of visual words* paradigm. Booij et al. [3] suggested that the robot could robustly navigate on the topological map through epipolar constraint and they employed image edges and SIFT features [10] as topological nodes. Daniele et al. [5] proposed a weighted graph to indicate the similarity metric related to image features. Their node representation also used the image features, and such graph was represented as their topological map.

For visual place recognition, Hemant et al. [15] employed a stochastic model of image perturbations in order to decide whether the image of a place was a location near a place with previously captured images. Pronobis et al. [1] adopted a high dimensional histogram to represent the visual scene and employed SVM to classify the visual places. The other highly related approaches is the context-based place recognition [16].

A similar topic but with different goal is the *visual loop-close detection*. The researchers employed the similar techniques as the visual place recognition to perform the repeated scenes detection. Chanop et al. [14] used a high resolution omnidirectional camera to build a feature-based topological visual map and then used the SIFT descriptor to detect previously observed scenes. Mark et al. [4] proposed a *bag-of-words* based loop-close detection. They used a probabilistic model to represent each scene, and the model was also used to calculate whether the current scene was similar to previously observed scenes. Viverk et al. [11] proposed a metric-based topological map and detected loop-close for outdoor scenes. In contrast to the previous approaches, our scene representation method is a code-based topological map. In the proposed technique, the location is not only a concise description, but also robust under the varying environment.

Our approach relies on the local feature transforms based on non-parametric measures that are able to tolerate the varying environment. The non-parametric measure "census transform" is first proposed by Ramin et al. [18], which was used to deal with the dense matching problem for disparity map construction. Recently, the researchers have also used it for visual place categorization [17] and face recognition [13]. In contrast to the previous census transforms, our approach is a sparse feature-based census transform and only process the convex hull points extracted from the image features.

The omnidirectional camera system has been used increasingly in robotic applications due to its capability of capturing the rich information (e.g. full 360° field of view) and low cost. In this work, we adopt only the omnidirectional camera for evaluating the encoding based scene representation and its capability in visual place recognition. Due to the omnidirectional imaging geometry, the image content of the border region does not vary as rapidly as the nearby region. Thus, it is able to provide the stable features[1] to build the binary codes and achieve acceptable scene change detection performance.

III. Visual Topological Localization System

We describe the original formulation of the hull census transform, and then introduce our application of this framework for scene representation and visual place recognition.

A. Descriptor of Hull Census Transform

This section discusses the HCT coding method. Let X^i be the total features in the ith image frame of an omnidirectional video sequence. Suppose Y_1^i is a set of features which forms the convex hull of X^i, then the features points in Y_1^i are located close to the border of the omnidirectional image. The rest features $X^i - Y_1^i$ are then used to derive the second convex hull Y_2^i. We repeat this procedure to construct multiple convex hulls from the set X^i such that

$$X^i = \bigcup_{l=1}^{n} Y_l^i \tag{1}$$

where n is the number of constructed convex hulls.

In order to explain how to build a binary code of a layer of the multiple convex hulls, we represent the feature points of a given convex hull as $Y_{l,p}^i$, where p is the index of feature points and l is one of the total n layers of the multiple convex hulls.

Suppose the set of δ_l^i feature points is a small set of the image features and the points lie on the lth layer convex hull of the ith image frame. The set can be denoted as

$$Y_l^i = \bigcup_{p=1}^{\delta_l^i} Y_{l,p}^i \ , \ Y_{l,p}^i \in Y_l^i \tag{2}$$

Suppose an image feature is represented by a vector, e.g. the vector dimension of SURF might be 64 or 128 [6]. If the magnitude of the vector of a feature point, $\parallel Y_{l,p}^i \parallel$, is less than its consecutive neighbor[2], then the bit code $B_{l,p}^i$ of this feature point is set as 0 and otherwise is set as 1. That is,

$$B_{l,p}^i = \begin{cases} 0, & \text{if } \|Y_{l,p}^i\| < \|Y_{l,p-1}^i\| \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

The proposed hull census transform relies on the relative ordering of connected feature points on the convex hull. The HCT code in a layer l is given by

$$hct_l^i = \bigcup_{p=1}^{\delta_l^i} B_{l,p}^i$$

where the code length may vary for different layer of convex hulls. The complete scene descriptor for an image frame i with multiple hulls is as follows:

$$HCT^i = \bigcup_{l=1}^{n} hct_l^i \tag{4}$$

An example of 2-layer HCT descriptor for an omnidirectional image is shown in Fig. 1.

[1] The definition of the stable feature is that a group of image features which can be detected and stayed in video frames for a period of times.

[2] This comparison can be processed by clockwise or counterclockwise of the points which lie on a convex hull. In other words, the HCT scene coding is also a cyclic coding with no specific starting point.

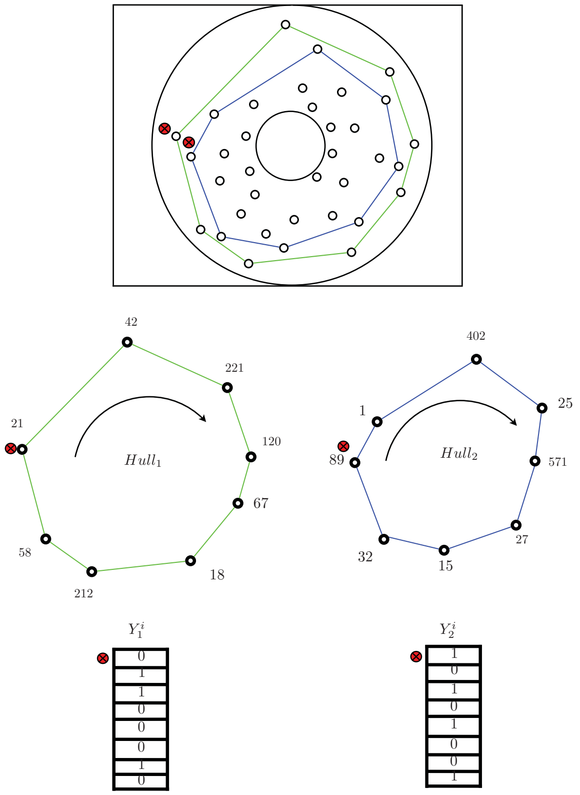


Fig. 1. An example of constructing a two-layer hull census transform. The black circles indicate the detected image features in the omnidirectional image. The connected line segments are then used to generate a layer of convex hull. In this example we consider only the image features forming two convex hulls. For each layer, we count the magnitude from each of the feature vectors and compare it with its neighbor feature. Finally, we form a two-layered HCT descriptor through Eq. (4).

B. Scene Change Detection

In this system, we use the HCT to detect scene change events and also generate the topological nodelists. To acquire suitable images for the robot vision system, we suggest that the robot should move at a normal speed (e.g. at a speed of roughly 0.3m/s), and then the video sequences are used in this system.

To detect the scene change events, the first step is to compute the rank transform of the HCT codes in two consecutive image frames. A rank transform in a layer of the HCT $\Re(hct_l^i)$ is to count the number of bits "1", which is defined as,

$$\Re(hct_l^i) = \sum_{p=1}^{\delta_l^i} B_{l,p}^i \ , \ \text{for} \ \parallel Y_{l,p}^i \parallel \geq \parallel Y_{l,p-1}^i \parallel \qquad (5)$$

We then check the HCT codes by the next two statistic formulations as follows based on the rank transform of the HCT codes. The second step is the "Cost value $\Omega(x,y)$", modified from "*Two-sample pooled t-test*" [8] to compare the HCT codes of two consecutive image frames, i and $i-1$. Let

$$\varnothing_i = \prod_{l=1}^{n} \frac{\Re(hct_l^i)}{\delta_l^i} \qquad (6)$$

and

$$\Omega(i,i-1) = \sqrt{{\varnothing_i}^2 - {\varnothing_{i-1}}^2} \qquad (7)$$

where $\varnothing_i \in [0,1]$ is a normalized rank sum value of HCT bits.

The last step is to calculate the "Score $\Psi(x,y)$" of the HCT code. The score value is converted to the format of an HCT from binary code to decimal value [17] in its ranking number of bits. Two HCT codes are then given as follows:

$$\Psi(i,i-1) = \prod_{l=1}^{n} \mid \Re(hct_l^i)_2 - \Re(hct_l^{i-1})_2 \mid \qquad (8)$$

$$\Phi(i,i-1) = \frac{\Psi(i,i-1)}{\Sigma_{l=1}^{n}(2^{\delta_l^i} + 2^{\delta_l^{i-1}})} \qquad (9)$$

If Ω and Φ exceed some pre-defined ratios then there is a scene change between images $i-1$ and i,

$$\textit{Scene change}: \ = \begin{cases} \Omega(i,i-1) \geq & \eta_1 \\ \Phi(i,i-1) \geq & \eta_2 \end{cases} \qquad (10)$$

In our experiment the two thresholds η_1 and η_2 are given by 0.75 and 0.25, respectively. Finally, we collect the calculated information (Eqs. (5) – (9)) to form the nodelists which located at the scene change locations and use Eq. (10) to build the topological map.

The HCT codes must be rotated in cyclic order to find the minimum Hamming distance before comparing the HCT codes. This said, when the robot rotates at a fixed position without forward/backward motion, the HCT codes also rotate with the robot self-rotation in the omnidirectional images. Thus, the minimum Hamming distance makes the HCT code unchanged. This makes the HCT preserve high similarities in the rotation omnidirectional images.

C. HCT Matching Scheme

The scenario for scene change detection and place recognition is that the robot navigates and captures the sequences with different trajectories possibly under varying illumination conditions. One of the image sequences is used to build a series of HCT codes by detecting the scene change signals for the topological map and form the nodelists. The rest video sequences with different paths are used to recognize each image frame by counting the highest similarity node which belongs to the nodelists.

For scene recognition, it should be noted that the number of total bits is not the same between the currently detected HCT code and the nodelists. The following equation is used to match two HCT codes and only the highest score node in the nodelists is selected to represent the scene,

$$\underset{j}{\text{argmax}} \{\Psi\langle i,j\rangle \ , \ \text{for} \ \Omega(i,i-1) \leq \eta_1\} \qquad (11)$$

where i is the detected HCT codes in the current omnidirectional image and j is an HCT code from the constructed scene change nodelists.

For a moving robot in an environment, the scene recognition procedure is as follows:

1) Rotate each of the HCT codes to find the minimum Hamming distance between the current node and the node from the nodelists.
2) Find the highest similar node using Eq. (11) by calculating the score with Eq. (8) to recognize the current scene.

IV. Experiment and Performance Evaluation

We have implemented the non-parametric local transforms, and explored their behavior based on the image features through multiple convex hulls. The proposed HCT is coded based on the image features, yet it can tolerate the environment change. All results are processed in omnidirectional video sequences. We also provide the analysis of HCT similarity and SURF feature matching under three different weather conditions.

A. Dataset

The COLD dataset [12] is adopted to evaluate our approach. Three different robot platforms with two heterogeneous cameras (catadioptric and perspective) are used to gather the image data under varying conditions and times in different environments. The videos are also captured under human motions, and different weather conditions (e.g. cloudy and night). The dataset is useful for testing the visual place recognition algorithms because the 3 different environments have similar rooms such as print rooms, one-person offices, etc.

B. Scene Change Detection and Topological Framework

The SURF feature extraction technique [6] is used in this work and the proposed HCT is compiled based on the extracted image features. The scene change detection is done with the HCT ranking and score described in Section III-B.

In this work, the topological nodelists construction rule is as follows: A video sequence from the COLD dataset of a robot is used to detect the scene change places and assign private labels. The unchanged places are assigned a label as the surrounding node. If a signal of the scene change event is rising, the color label will change in the topological map and automatically add a node to the topological nodelists. Fig. 3(a) shows the scene change detection ability using 1459 image video frames in the cloudy weather path-1 of the COLD-freiburg robot. The topological map is constructed with no refinement strategy, Fig. 5 shows the 12 detected scene change nodes plotted with a big color dot. The average recognition rate is about 66.7%. Due to the environment structure and the corresponding walls of the captured omnidirectional images between the places are quite similar in the dataset (e.g. the printer room, two-person office and the corridor), we think the recognition rate is acceptable for some mobile robot applications.

TABLE I

The data capacity for constructing the 12 topological nodes. The comparison of the images (size 512×384), SURF features with 128-dimension vector and HCT codes.

Methods	capacity
Images for 12 nodes	6912kB
SURF features for 12 nodes	240kB
HCT (2 layer) for 12 nodes	1kB
HCT (2 layer) for total 1459 frames	130kB

The data capacity is also an important issue for long-term working of a mobile robot in an environment. Due to the high sparse coding method of the proposed HCT, we show the comparative results between the HCT, image features and images themselves in Table I. In the toleration testing of different weather conditions, we list a table (see Table II) to compare the similarity and feature matching between three different omnidirectional images captured at the close locations in the cloudy, sunny weather and at night, respectively.

In Table II, row (a) lists 3 weather conditions and their images are shown in Fig. 2. Rows (e) and (l) are the sub-item of weather conditions for showing the image feature matching, execution time, etc. Row (c) is the total detected SURF features of each image. Row (f) shows the feature matching between different weathers, for example, “16” means there are 16 features matched between cloudy and sunny images, and so on. Row (m) is the similarity calculated by the proposed method, which shows the high similarity between the three weather conditions.

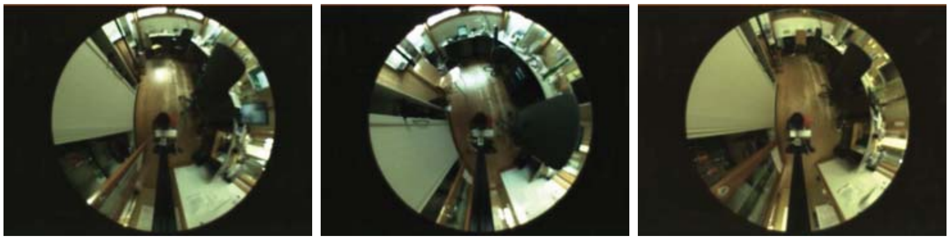
C. Scene Recognition

The recognition rule in this work is as follows: When the topological nodelists are built, the rest videos of the same robot with different paths are used to recognize the highest similarity node from each image frame. The visual place recognition does not need any refinement stages. Each omnidirectional image frame is considered as a visual place and given a label with the highest similar node in the nodelists. The average recognition rates are shown in Fig. 4 and the recognized areas of the robot paths in different weather conditions with color labels are plotted in Fig. 3(b) - 3(d).

V. Conclusion and Future Work

This paper presents an approach for topological visual place recognition based on a modification of census transform. Our method, with repeatedly generating the convex hull from the image features and computing the relative magnitude between image features over the convex hull, has shown promising results on the COLD datasets.

We have evaluated the performance of the topological place recognition by using an omnidirectional camera. The major limitation of the HCT is the moving speed of the robot and the frame rate of the camera. If the frame rate is too slow (e.g. 2 – 3 fps) or the robot moves too fast, such as 1 m/s,



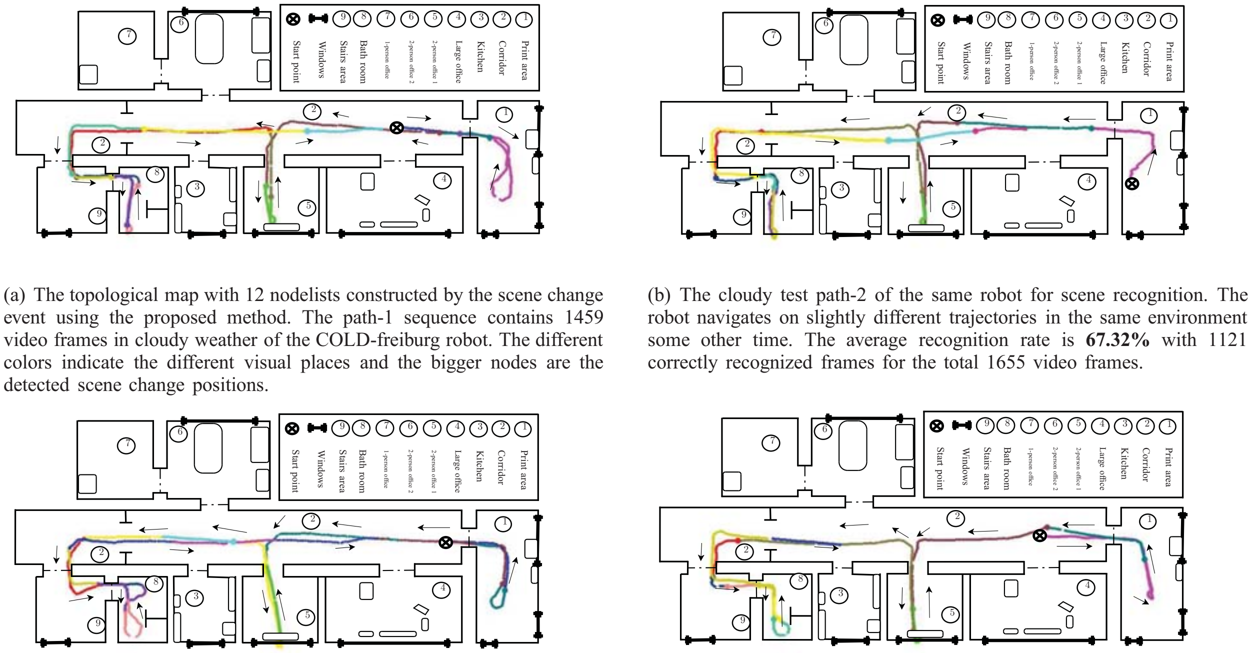
(a) Cloudy.
(b) Sunny.
(c) Night.

Fig. 2. The three images are captured under different weathers at close positions. In this test case, the analysis result is shown in Table II.

TABLE II

THE PERFORMANCE EVALUATION OF HCT FOR CHANGING WEATHER. THE HCT OFFERS HIGH SIMILARITY USING IMAGE FEATURES.

Weather (a)	Cloudy		Sunny		Night	
Image size (b)	512 × 384		512 × 384		512 × 384	
SURF features (c)	317		397		289	
Extraction time (d)	600.1ms		614.4ms		637.2ms	
(e)	Sunny	Night	Cloudy	Night	Cloudy	Sunny
Matching number (f)	**16**	**14**	**16**	**6**	**14**	**6**
Matching time (g)	133.3ms	100.5ms	133.3ms	128.7ms	100.5ms	128.7ms
HCT codes (h)	0110000011010110100010		01101010110110010110		10101001110100011011	
Total HCT bits (i)	22		20		20	
$\varnothing$(j)	0.272157		0.846578		0.412924	
$\Re$ (k)	0.41		0.55		0.55	
(l)	Sunny	Night	Cloudy	Night	Cloudy	Sunny
Similarity (m)	**Ω:96.2%,Φ:43%**	**Ω:99.7%,Φ:86%**	**Ω:96.2%,Φ:43%**	**Ω:95.9%,Φ:56.7%**	**Ω:99.7%,Φ:86%**	**Ω:95.9%,Φ:56.7%**
HCT build time (n)	13.2ms		4.0ms		9.3ms	



(a) The topological map with 12 nodelists constructed by the scene change event using the proposed method. The path-1 sequence contains 1459 video frames in cloudy weather of the COLD-freiburg robot. The different colors indicate the different visual places and the bigger nodes are the detected scene change positions.

(b) The cloudy test path-2 of the same robot for scene recognition. The robot navigates on slightly different trajectories in the same environment some other time. The average recognition rate is **67.32%** with 1121 correctly recognized frames for the total 1655 video frames.

(c) The sunny test path-1 for scene recognition. The robot navigates on different trajectories in same environment but another weather. The average recognition rate is **68.67%** with 1098 correctly recognized frames for the total 1599 video frames.

(d) The night test path-1 for scene recognition. The average recognition rate is **64.45%** with 1126 correctly recognized frames for the total 1747 video frames.

Fig. 3. The map is downloaded from the COLD dataset [12] with the odometry trajectory. The odometry information is used only for showing the robot movement. It is not used in the place recognition tasks. The colored paths indicate different areas of the environment regions which are detected by the HCT. Both of the scene change detection and recognition are tested with no refinement strategies.

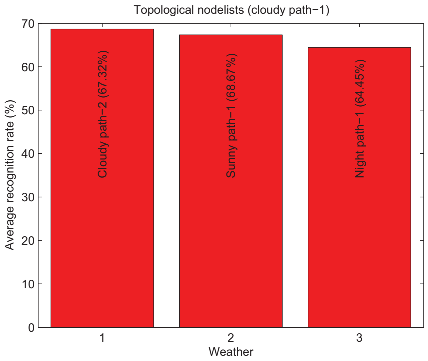


Fig. 4. The average recognition results of the topological localization system. The topological nodelists are constructed by cloudy path-1.

each frame will be considered as a signal of scene change. The HCT can also operate with the image features extracted from the perspective camera. However, it is not suitable for visual place recognition. In this work, we chose the SURF as the image features because it is faster than SIFT features for video sequences. Furthermore, the HCT coding method depends on the image features, but is not limited to the high quality image features. In other words, the consistency of video images is more important than the quality of the video sequences.

We are also interested in testing how many layers of HCT are enough for the topological localization framework. In our testing, 2 − 3 hulls are suitable for the topological nodelists construction using the COLD dataset. If we use more than 3 layers, the average hit score and cost value does not usually increase. For further applications, the multiple HCT descriptor can also be applied for solving the scene categorization problem, which will be investigated in the future research.

VI. Acknowledgments

The support of this work in part by the National Science Council of Taiwan, R.O.C, under Grant NSC-96-2221-E-194-016-MY2 is gratefully acknowledged.

References

[1] P. J. A. Pronobis, B. Caputo and H. I. Christensen., “A discriminative approach to robust visual place recognition,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[2] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, “Visual topological slam and global localization,” International Conference on Robotics and Automation (ICRA)., 2009.

[3] O. Booij, B. Terwijn, Z. Zivkovic, and B. KrLose, “Navigation using an appearance based topological map,” International Conference on Robotics and Automation (ICRA)., pp. 3927–3932, 2007.

[4] 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. [Online]. Available: http://ijr.sagepub.com/cgi/content/abstract/27/6/647

[5] D. Fontanelli, P. Salaris, F. A. W. Belo, and A. Bicch, “Visual appearance mapping for optimal vision based servoing,” Experimental Robotics: The 11th Intern. Sympo., STAR 54, Springer Berlin/Heidelberg, pp. 353–362, 2009.

[6] T. T. Herbert Bay and L. V. Gool., “Surf: Speeded up robust features,” in *ECCV ’06: Proceedings of the 9th European Conference on Computer Vision-Part I*. Springer-Verlag, 2006, pp. 404–417.

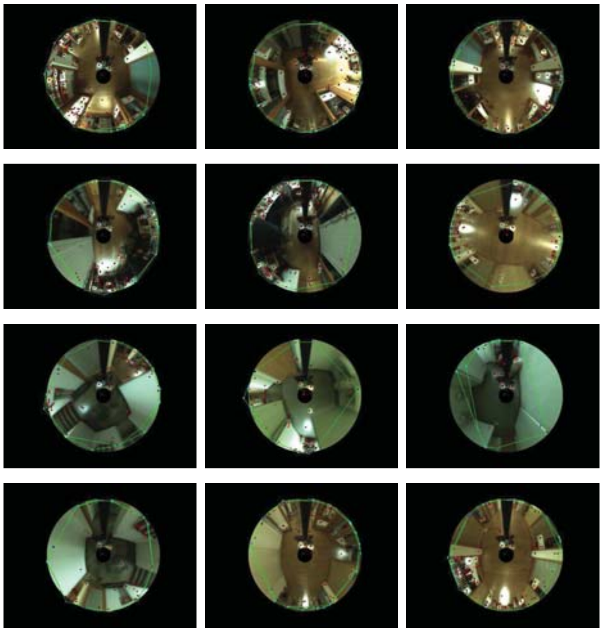


Fig. 5. The 12 nodes of scene change positions are detected by the proposed HCT method with no refinement strategy using the path-1 video sequence in the cloudy weather. The video sequence has totally 1459 video frames.

[7] S.-W. Lee, Y.-M. Kim, and S. W. Choi, “Fast scene change detection using direct feature extraction from mpeg compressed videos.” *IEEE Trans. on Multimedia*, vol. 2, no. 4, 2000.

[8] E. Lehmann and J. P. Romano, “Testing statistical hypotheses (3e ed.).” in *New York: Springer.*, 2005.

[9] M. Liu, D. Scaramuzza, C. Pradalier, R. Siegwart, and Q. Chen, “Scene recognition with omnidirectional vision for topological map using lightweight adaptive descriptors,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009), St Louis, Missouri, USA, October, 2009.

[10] D. Lowe, “Local feature view clustering for 3d object recognition,” *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–682–I–688 vol.1, 2001.

[11] V. Pradeep, G. Medioni, and J. Weiland, “Visual loop closing using multi-resolution sift grids in metric-topological slam,” CVPRW ’09: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, 2009.

[12] A. Pronobis and B. Caputo, “COLD: COsy Localization Database,” *The International Journal of Robotics Research (IJRR)*, vol. 28, no. 5, May 2009.

[13] J. A. Ruiz-Hernandez, J. L. Crowley, A. Meler, and A. Lux, “Face recognition using tensors of census transform histograms from gaussian features maps,” The British Machine Vision Conference, 2009.

[14] C. Silpa-Anan and R. Hartley, “Visual localization and loop-back detection with a high resolution omnidirectional camera.” Workshop on Omnidirectional Vision, 2005.

[15] H. Tagare, D. McDermott, and H. Xiao, “Visual place recognition for autonomous robots,” in *Robotics and Automation, 1998. Proceedings. 1998 IEEE International Conference on*, 1998.

[16] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, “Context-based vision system for place and object recognition,” in *Proceedings of the IEEE International Conference on Computer Vision.*, 2003.

[17] J. Wu, H. I. Christensen, and J. M. Rehg, “Visual place categorization: Problem, dataset, and algorithm,” IEEE/RSJ International Conference on Intelligent Robots and Systems, St Louis, Missouri, USA, October, 2009.

[18] R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” in *In Proceedings of European Conference on Computer Vision*, 1994, pp. 151–158.