

# Topological map construction and scene recognition for vehicle localization

Huei-Yung Lin<sup>2</sup> · Chia-Wei Yao<sup>1</sup> · Kai-Sheng Cheng<sup>1</sup> · Van Luan Tran<sup>1</sup>

Received: 5 June 2015 / Accepted: 22 April 2017 / Published online: 8 May 2017  
© Springer Science+Business Media New York 2017

**Abstract** This paper presents a vehicle localization method to assist vehicle navigation based on topological map construction and scene recognition. A topological map is constructed using omni-directional image sequences, and the node information of the topological map is used for place recognition and derivation of vehicle location. In topological map construction and scene change detection, we utilize the Extended-HCT method for semantic description and feature extraction. Content-based and feature-based image retrieval approaches are adopted for place recognition and vehicle localization on the real scene image dataset. The proposed technique is able to construct a real-time image retrieval system for navigation assistance and validate the correctness of the route. Experiments are carried out in both the indoor and outdoor environments using real world images.

**Keywords** Scene recognition · Topological map · Scene change detection

## 1 Introduction

Scene recognition and autonomous navigation for vehicle driving assistance and mobile robot applications are active research topics in recent years. In general, vehicles or mobile

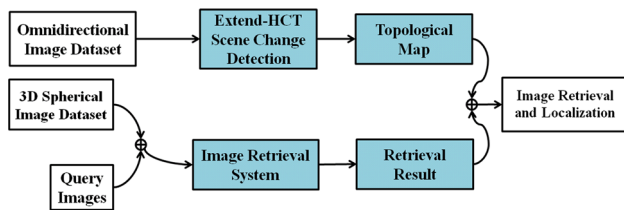
robots need to recognize the scenes and derive their location information in the environment using exteroceptive sensors such as cameras, laser range finders, sonars or GPS, for the purpose of autonomous navigation. Although GPS is commonly used in navigation, it does not function well under several circumstances due to data loss or communication failure. For instance, bad weather conditions, mountains, or tall buildings in the cities may reduce or block GPS signals. To enhance the performance of GPS based positioning, several techniques such as differential GPS (DGPS), assisted GPS (AGPS), or combined with inertial navigation system (INS) have been developed (Maier and Kleiner 2010; Toledo-Moreo et al. 2007). Nevertheless, they are not suitable for vehicle navigation in general driving environments. From the human assisted technology perspective, the visual information is considered as a good alternative to improve the GPS localization techniques (Larnaout et al. 2014; Conte and Doherty 2009).

To design a vehicle navigation and localization system, we propose a navigation assistance concept. It uses an on-board camera in the vehicle to capture images for image retrieval in a pre-established image database, and the driver can then check whether the location and direction are correct based on the scene recognition information. In this work, a vision assisted navigation technique based on topological map construction and scene recognition is presented. As shown in Fig. 1, the dramatic scene change locations are described by the Extended Hull Census Transform (Extended-HCT) based on omni-directional image sequences (Wang and Lin 2011) and used to construct the topological map. To increase the robustness of scene change detection and place recognition, we combine the CBIR and FBIR (feature-based image retrieval) techniques to improve the weak points of the individual image retrieval methods. In general, it is possible to find the regions with similar colors but result in a misjudg-

✉ Huei-Yung Lin  
lin@ee.ccu.edu.tw

<sup>1</sup> Department of Electrical Engineering, National Chung Cheng University, 168 University Rd, Min-Hsiung, Chiayi 621, Taiwan

<sup>2</sup> Department of Electrical Engineering and Advanced Institute of Manufacturing with High-Tech Innovation, National Chung Cheng University, 168 University Rd, Min-Hsiung, Chiayi 621, Taiwan



**Fig. 1** The system flowchart of the proposed vision assisted navigation system. It is composed of two parts: topological map construction (*top*) and image retrieval (*bottom*)

ment of objects in their shapes for some specific images if only the CBIR method is used. In contrast, if only the FBIR method is adopted, the recognition could result in different objects with similar shapes. Consequently, an image retrieval algorithm taking both the content and feature characteristics might benefit the overall recognition results.

Generally, the main contribution of this work lies on the two aspects of the proposed navigation technique as follows:

- Scene change detection and topological map construction: We show that the Extended-HCT is a sound way for constructing topological maps and defining scene change nodes.
- Hybrid image retrieval by combining CBIR and FBIR methods: We build a robust image retrieval system and greatly reduce the image retrieval time.

To perform the scene change detection and topological map construction, we use the SURF descriptor and the Extend-HCT method. First, We extract the key points of the omni-directional images by the SURF detector. The extracted key points of each image is described by six parameters, and the relationship among feature magnitude, feature point structure and color information is also included. The six types of parameters are: features of convex hulls, cost, score, color histogram index, average distance of center and average distance of feature. Then, the scene change nodes are defined by these six parameters.

To perform the image retrieval for localization, we use the combination of the Content-based image retrieval (CBIR) and feature-based image retrieval (FBIR) methods. The CBIR method is based on shape, color, texture, spatial correspondence or other information in the images. The compact composite descriptors (CCD) is used to compute similarity between dataset and the ground truth to pick up a few candidates, and then determine the one with the highest score as output. This method describes the information of different scenes using lower-level feature points, and combines the fuzzy system to classify and index the images. The CCD descriptor includes 84 bytes of float number with color, brightness, texture direction and edge direction information. Then, The FBIR method use feature matching by SURF and

classification by  $k$  nearest neighbors ( $k$ -NN) to calculate the descriptor distance of the query and candidate images. If the descriptor distance is less than a threshold value, then we add one point to the similarity rank score. Finally, the image of highest similarity rank score in the candidate images is the retrieval result. As the result, the image retrieval time can be reduced significantly by combining the CBIR and FBIR methods. Finally, we compare the proposed technique with the FAB-MAP system (Cummins and Newman 2008) using the same datasets.

The rest of this paper is organized as follows. Section 2 shows some related works. Section 3 presents the method for scene change detection and topological map construction. Section 4 describes the hybrid image retrieval algorithm for robust place recognition. The implementation and experimental results are provided in Sect. 5, followed by the conclusion of this work in Sect. 6. Synthetic motion blur generation for both.

## 2 Related work

In our work, the navigation and localization system is used to support the driver to verify the location and direction based on the scene recognition information. For many applications of navigation and localization with topological map, a topological map which contains the information of important nodes is constructed for motion planning and navigation. It can also be used to provide the semantic description of a scene for vision assisted driving techniques. In the previous work, Ranganathan presents a method called “place labeling through image sequence segmentation” (PLISS), for place recognition and categorization from visual cues (Ranganathan 2012). Liu et al. use omnidirectional cameras for scene reconstruction (Liu and Siegwart 2012, 2014). They extract major vertical lines from panoramic images to derive the ROI in the neighborhood. For the proposed adaptive descriptor, it is more proper than the field of view (FOV) to describe the primitive features (Liu et al. 2013). Liu et al. also present several topological navigation techniques, e.g. the homing robot application in the visual approach using uncalibrated omnidirectional camera within image based visual servoing (IBVS). Given a topological map embedded with visual information, robot navigation can be guided by recognizing the scene images. In the past few decades, solving the general scene recognition problem has been one of the most important robot vision research topics. To deal with image retrieval and localization, Smeulders et al. (2000) review over 200 references in content-based image retrieval (CBIR) methods, e.g., the types of pictures, the role of semantics and sensory gaps. Philbin et al. (2007) present a large-scale object retrieval system, where a user can select a query image for the query object and the system is able to return several

possible candidates. [Eakins et al. \(1999\)](#) propose a CBIR concept based on shapes, colors, textures, and other features in the image content for retrieval. [Milford and Wyeth \(2012\)](#) develop a feature-based image retrieval technique called sequential simultaneous localization and mapping (SeqS-LAM), which is able to adapt to the changing environment conditions. On the other hand, some cloud services such as Google Street View and Google Earth are getting popular on mobile and smart phone applications. Thus, it is highly desirable to develop a visual navigation system which takes the street view images as the primary information source. Finally, Stone et al. proposed a novel combination of sensory modality and image processing to obtain a significant improvement in the robustness of sequence-based image matching for place recognition, use a UV-sensitive sheye lens camera to segment sky from ground, providing illumination invariance, and encode the resulting binary images using spherical harmonics to enable rotation-invariant image matching. This method enables robust place recognition during aggressive zigzag manoeuvring along bumpy trails and at tilt angles of up to 60°. Like all camera-based methods, their method is sensitive to extreme lens are, such as that caused by direct sunlight very early or late in the day ([Stone et al. 2016](#)).

More recently, many works focus on the navigation and localization system, as it is important in the applications of vehicles or mobile robots. As in the introduction, our methods and targets are generality different from the localization of Mobile Robot as mentioned previously. The common localization of Mobile Robots with the topological map are usually to estimate the position and angle for mobile robot. In our system, we use to assist the vehicle navigation system based on topological map construction and scene recognition.

### 3 Scene change detection and topological map construction

Scene change detection was mainly used for MPEG video compression in early years for multimedia applications ([Shahraray 1995](#); [Lee et al. 2000](#); [Huang and Liao 2001](#)). In the field robotics research, it is used for topological map construction in the identification of significant locations or landmarks such as street intersections, buildings, etc. In order to let a vehicle or mobile robot can automatically detect and construct a topological map of the environment, we adopt the Extended-HCT technique to build the descriptors for omnidirectional image sequence. It first transforms the images to a series of binary codes for representation, and the information associated with these image descriptors is used to detect the scene change frame and location in the image database. A simple topological map composed of nodes and links is then

constructed from the complex data map. During the navigation, it is easier to observe the correct path by the continuity relationship among the nodes. An overview of the proposed system is briefed below.

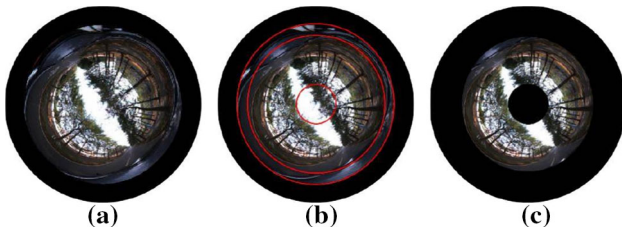
**Image dataset acquisition.** We use a multi-complex omnidirectional camera to capture dome-like images of 360° outdoor scenes. Two image datasets, the CCU campus and street scenes, are used to verify the proposed method. Each image in the datasets contains the GPS coordinates which can be used to check the correctness of localization results.

**Extend-HCT codebook generation.** The extended-HCT encoding is designed to obtain the descriptors for omnidirectional images. A series of binary codes is generated to describe the features associated with an image frame. **Scene change detection.** To define a scene change location in an image sequence, we analyze the HCT codes and find the common crests and troughs in the feature space to identify the consecutive image frames with a significant content difference.

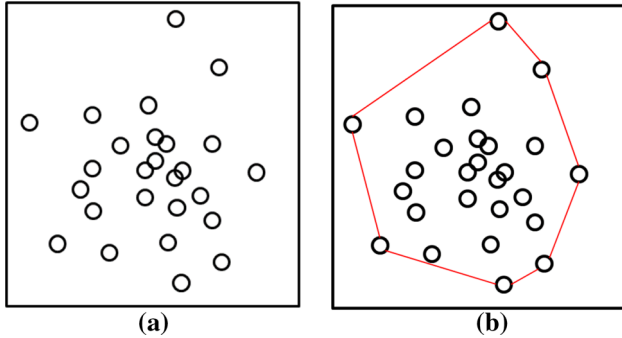
**Topological map construction.** The major difference between a topological map and a metric map is the former only retain the node and connection information of a route. With the scene change detection results, we only set a few pin marks to represent the important nodes.

#### 3.1 Omni-directional image dataset

To collect the image dataset, we use a spherical vision Ladybug2 camera. It contains six CCD sensors, which are used to resemble an omnidirectional camera for panoramic scene acquisition. The Ladybug2 camera allows our navigation system to record the image frames for more than 75% of the 360-degree spherical imagery and map the six images obtained from different viewpoints to a 3D spherical surface. Compared to the previous works ([Pronobis and Caputo 2009](#); [Wang and Lin 2010](#); [Rituerto et al. 2014](#)), the multi-complex camera has more advantageous terms than the traditional catadioptric lens, in terms of the image quality and resolution. The approximate pinhole images and several types of high quality images, e.g., dome-like and panoramic, are readily available. In the experiments, we mount a Ladybug2 camera on the roof of vehicle and connect it to a laptop through an IEEE1394 port. The image sequence and GPS information are acquired simultaneously during the vehicle motion. To reduce the computation time of the Extended-HCT encoding, a circular region of interest (ROI) is extracted for further processing (see Fig. 2). We set the original image size as 512 × 512, and the outer and inner mask radii are 220 and 65 pixels, respectively.



**Fig. 2** A circular region of interest (ROI) is extracted for processing to reduce the computation. **a** The original omnidirectional image. **b** Region selection is shown in red. **c** The ROI used for processing (Color figure online)



**Fig. 3** The construction of a convex hull. **a** A group of sparse points. **b** A convex hull constructed by the outer points

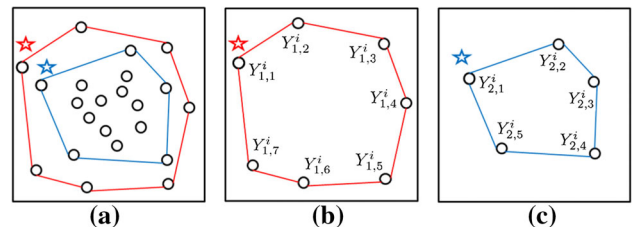
### 3.2 A HCT-Based semantic description

In this section, we introduce the convex hull feature description for omnidirectional images, and consider the problem of detecting the scene change events for navigation. We adopt the Extended-HCT technique to build the omnidirectional image descriptors (Wang and Lin 2011). The method is developed for the scene change detection using omnidirectional images, especially it can adapt to the dramatic and rapid environment illumination change. To improve the scene change detection results and make it suitable for the outdoor environment, a more sophisticated parameter estimation method is proposed for scene change node analysis. The feature point extraction for convex hull description is based on the SURF algorithm. In addition, the color histogram index is added to increase the robustness of the descriptor.

Figure 3 shows an example of a convex hull defined on a hyperspace with a group of sparse points. The outer points are connected by a closed contour or polylines. Any of the connection lines cannot cross each other, or equivalently, any connecting point of the lines is not allowed to appear in a convex hull.

#### 3.2.1 Hull census transform

HCT is a binary code for describing the feature information of images. First, we extract the key points of omnidirectional



**Fig. 4** An illustration of two layers of convex hull. **a** The star indicates the starting position of HCT. **b** The first layer convex hull  $Y_1^i$ . **c** The second layer convex hull  $Y_2^i$

images by the SURF detector, and use  $X^i$  to represent a group of feature points from the  $i$ th frame in a video sequence. Let  $Y_1^i$  be the outermost convex hull deriving from the set  $X^i$ . By eliminating the first layer convex hull  $Y_1^i$ , the remaining points form a new set of feature points,  $X^i - Y_1^i$ . A second layer convex hull  $Y_2^i$  can then be generated from the feature set  $X^i - Y_1^i$ . Repeating this process, a set of multi-layer convex hulls can be established, and

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

where  $n$  is the number of extracted convex hulls.

To convert a convex hull to the binary code representation, we denote the feature point in a layer as  $Y_{l,p}^i$  and the feature set  $Y_l^i$  in Eq. (1) can be written as

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

where in the  $i$ th frame:

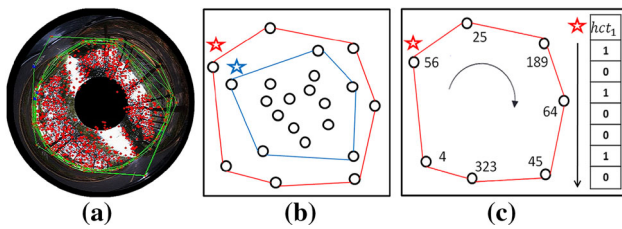
- $p$  is the index of the feature points in a convex hull.
- $l$  is the layer index of the  $n$  convex hulls for HCT encoding.
- $\delta_l^i$  is the number of feature points which lie on the  $l$ th layer.

Consequently,  $Y_{l,p}^i$  represents the  $p$ th feature point in the  $i$ th frames and  $l$ th layer (see Fig. 4).

The feature vectors extracted using SURF can be represented as a 64 or 128 dimensional vector. We take the norm of  $Y_{l,p}^i$  on each layer of the convex hulls to generate a binary sequence. Starting at an arbitrary point on a convex hull, if the norm of the current feature point is larger than the norm of the next feature point, then  $B_{l,p}^i$  is set as 1. Otherwise, it is set as 0. More specifically,

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





**Fig. 5** **a** A multi-layer convex hull in an omnidirectional image. **b** The *star* indicates the starting position of HCT. **c** The coding process of the convex hull  $hct_1$

Note that,  $B_{l,p}^i$ 's can be obtained either clockwise or counter-clockwise. The HCT code of the  $l$ th layer is then given by

$$hct_l^i = \bigcup_{p=1}^{\delta_l^i} B_{l,p}^i \tag{4}$$

Finally, the encoding of the image with all layers is written as

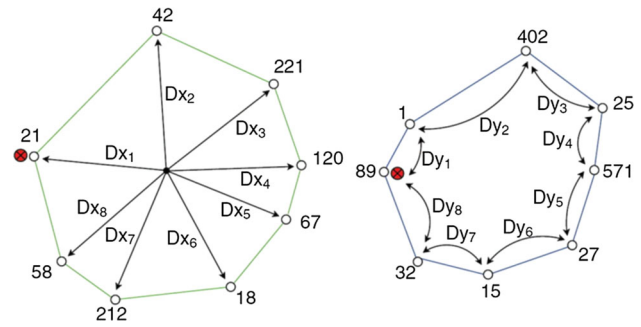
$$HCT^i = \bigcup_{l=1}^n hct_l^i \tag{5}$$

where  $i$  as the frame index and  $n$  is the number of convex hulls used to generate the HCT code.

By definition, the length of different HCT codes may be different, and an image frame can contain several  $hct_l^i$ . The number of HCT code depends on the quantity of layers (see Fig. 5). Figure 5a shows the extracted HCT feature points of an omnidirectional image. Figure 5b is an example of constructing a two-layer hull census transform. It also indicates the detected SURF features in the omnidirectional image. According to the convex hull transform experiments of real world scenes, the magnitude change of the inner circles is smaller than the change of the outer circles. As a result, the outer circles can provide more stable environment information. Figure 5c illustrates a series of binary codes describing the relationship between the feature points in the same layer built for image representation.

### 3.2.2 Extended-HCT

In the Extended-HCT algorithm, each image is described by six parameters, and the relationship among feature magnitude, feature point structure and color information is also included. The six types of parameters are (1) *Features of convex hulls* the total feature points in each convex hull. (2) *Cost* showing the discriminative property of each feature vector. (3) *Score* evaluation by converting the Extended-HCT code to a decimal value. (4) *Color histogram index* the color information of images represented by four channels– red, green, blue and gray. (5) *Average distance of center* computing the



**Fig. 6** An illustration of the definition of  $D_x$  and  $D_y$  given by Eqs. (9) and (10)

average distance between each feature point and the centroid. (6) *Average distance of feature* computing the average distance between the neighboring feature points.

The color histogram index includes four channels,  $\theta_r, \theta_g, \theta_b$  and  $\theta_{wg}$ . The channel values are normalized as  $r = R / (R + G + B), g = G / (R + G + B), b = B / (R + G + B), wg = 0.299R + 0.587G + 0.114B$ , where  $wg, R, G$  and  $B$  are gray, red, green and blue color intensities, respectively. The color channel histograms are then given by

$$C_i = \begin{cases} H_r(i) = \sum_x^W \sum_y^H r(x, y) \\ H_g(i) = \sum_x^W \sum_y^H g(x, y) \\ H_b(i) = \sum_x^W \sum_y^H b(x, y) \\ H_{wg}(i) = \sum_x^W \sum_y^H wg(x, y) \end{cases} \tag{6}$$

$$I_i(\theta_r, \theta_g, \theta_b, \theta_{wg}) = \left[ \frac{H_r(i)}{WH}, \frac{H_g(i)}{WH}, \frac{H_b(i)}{WH}, \frac{H_{wg}(i)}{WH} \right]^T \tag{7}$$

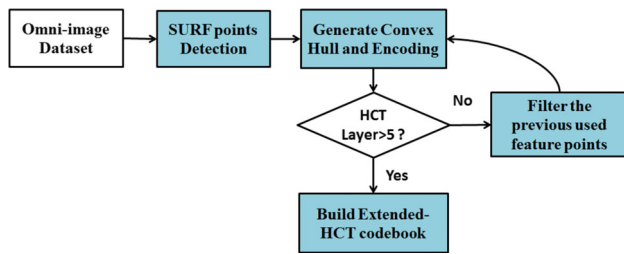
where  $H_r(i), H_g(i), H_b(i), H_{wg}(i)$  are the histograms of four channels and  $WH$  is the image size.

The average distance between the features and their centroid, and the average distance between the neighboring features are included in the Extended-HCT to describe the structure of the feature points. For each feature points, we calculate the distance  $D_{xi}$  and  $D_{yi}$  in the same convex hull, as illustrated in Fig. 6. In our scene change detection model, the equations are given by

$$Cx_l^i = \frac{1}{\delta_l^i} \sum_{p=1}^{\delta_l^i} B_{l,p}^i(x, y) \tag{8}$$

$$Dx_i = \frac{1}{\delta_l^i} \sum_{p=2}^{\delta_l^i} \left( B_{l,p}^i(x, y) - Cx_l^i(x, y) \right)^2 \tag{9}$$

$$Dy_i = \frac{1}{\delta_l^i} \sum_{p=2}^{\delta_l^i} \left( B_{l,p}^i(x, y) - B_{l,p-1}^i(x, y) \right)^2 \tag{10}$$



**Fig. 7** The flowchart of the Extended-HCT codebook construction. In this work, five layers of the convex hulls are extracted for feature representation

where  $B_{l,p}^i(x, y)$  are the feature points on a convex hull and  $Cx_l^i(x, y)$  is the centroid of the convex hull. Figure 7 depicts the entire process of the Extended-HCT codebook generation, five layers of the convex hulls are extracted for feature representation in this work.

### 3.3 Scene change node analysis

In the implementation, we convert the HCT results to a text file before the analysis for scene change locations. The scene change nodes are defined by six parameters as discussed previously. Each group of the parameters are extracted from five layers of HCT convex hulls, so the total parameters in an image is 29, as shown in Table 1. After the encoding of descriptors is completed, the method in Lin et al. (2013) is used for scene change detection.

Figure 8a shows a typical example of the parameter values over time. Different colors are used to illustrate different parameters. In the figure, the close relations and high dependencies among the parameters can be observed, e.g., some peaks and valleys appear at the nearby locations. We analyze the common peaks and valleys to derive the images with significant scene changes. Figure 8b shows the gradient detection result using a one-dimensional convolution mask  $G_m = [-1, 0, 1]$ . To make the zero-crossing on the parametric descriptors for scene change detection feasible, we employ a sliding window to search the extremes of the descriptors within a fixed time interval. As shown in Fig. 8b, the blue lines correspond to the scene change frames and are used to mark the nodes on a topological map. The following procedure is carried out to obtain the scene change nodes from the Extended-HCT code.

**Table 1** The Extend-HCT parameters

Parameter types	Feature of hull	Cost	Score	Color <sup>a</sup>	A.D. feature <sup>b</sup>	A.D. center <sup>b</sup>
Total	5	5	5	4	5	5
Order in code	1 ~ 5	6 ~ 10	11 ~ 15	16 ~ 19	20 ~ 24	25 ~ 29

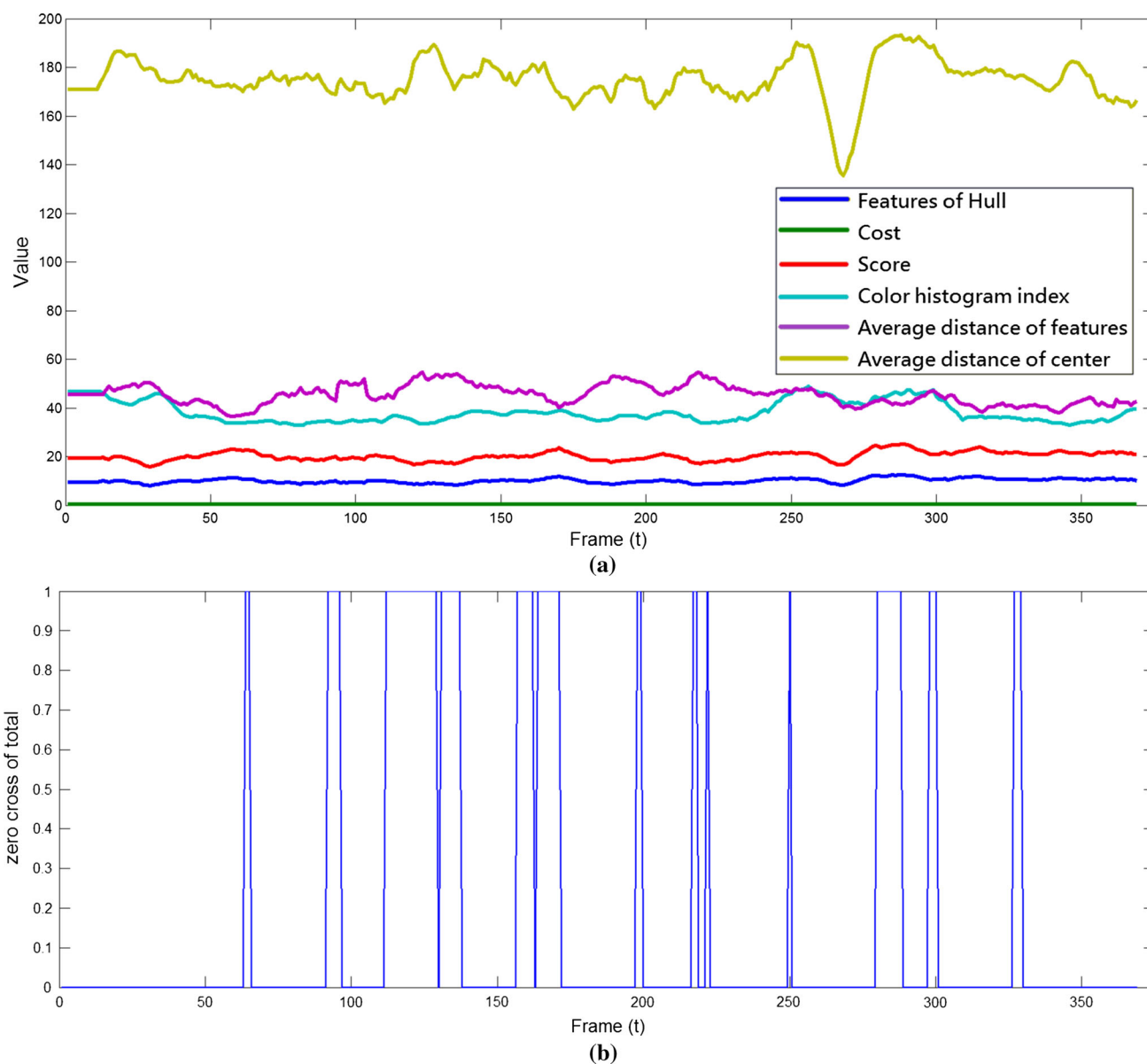
<sup>a</sup> Color represents the color histogram index (R, G, B, Gray)

<sup>b</sup> A.D. is the average distance

- (1) *Remove outliers and fill in the missing data* We employ the SURF descriptor for feature detection. In some scenes with less feature points, the convex hull construction will be affected. Consequently, some parameter values will decrease and be encoded as zeros. To avoid the missing information and maintain a stable binary code, the features on the outer layers are duplicated to make the size of convex hulls consist in five layers.
- (2) *Compress the parameters and filter the plot* There are 29 parameters obtained in process (1). In order to reduce the noise in the Extended-HCT codes, we take the average value of 29 parameters in 6 groups (each group contains four or five parameters) and perform an information reduction. As illustrated in Fig. 8a, it can be observed that the noise is reduced while the original attribute is still maintained.
- (3) *Detect the gradient information using a sliding window* We use a one-dimensional mask  $G_m = [-1, 0, 1]$  to derive the gradient information, and convert the result in process (2) to a gradient plot. Each parameter curve comes across the  $x$  axis. The crest/trough locations in the parameter space correspond to the frames with a zero value in the gradient plot.
- (4) *Generate the zero-crossing result* The zero-crossing indicates that the crest/trough has been detected through a sliding window. A threshold is given for each of the six parameters. In general, the crest/trough does not appear in the same frame. If the zero-crossing output contains six of them (maximum), it means that the zero-crossings of six parameters are all positive.

To summarize, we introduce the Extend-HCT method to define the scene change nodes, and use the SURF descriptor to extract feature points in the omni-directional images to build five layers of convex hulls for each image frame. In addition, several processes are used to analyze the Extend-HCT codebook and output the zero-crossing result. If the value is larger than the a pre-set threshold, a scene change node will be created for the topological map. Table 2 shows the typical settings of the scene change detection parameters.

We use the COLD dataset (Pronobis and Caputo 2009) to verify the performance of the Extend-HCT algorithm. The dataset contains 1459 frames of omni-directional images captured in the indoor environment (see Fig. 9). The experimental results on the average distances of feature/center for



**Fig. 8** Scene change nodes derivation based on the HCT results. **a** The change of the Extended-HCT code over time. **b** The scene change detection results

which the scene changes are listed in Table 3. Based on the experiments on the COLA dataset, we are able to build a topological map for the outdoor scenes. The marked position in the map derived by the Extended-HCT is used to represent the significant location such as an intersection or a landmark.

The major difference between the experiments on the indoor and outdoor scenes is the map scale. In the indoor experiments using the COLA dataset, it is easier to obtain the location information according to a pre-defined trajectory or using the sensors on the robot. However, the localization in the outdoor experiments heavily depends on the GPS data, which, in general, can only be used to verify the rough location due to its accuracy. The nodes marked by the Extend-

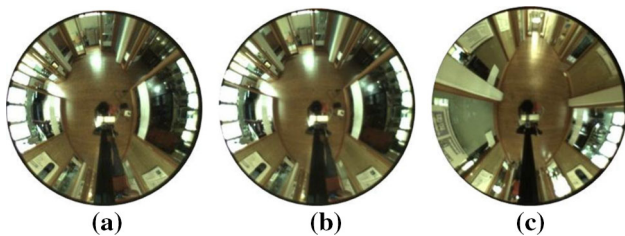
HCT along with the travel route generate a graph. A pair of nodes close to each other usually represent nearby geographic locations. They are condensed on the graph to create the topological map.

#### 4 Image retrieval and localization

In this section, we present an approach to use the pre-established image database for image retrieval and localization. To use the images obtained from a car dash camera, we capture the street view images using a Ladybug2 spherical camera. We set the viewpoint of the spherical image at

**Table 2** The setting of scene change parameters

Process for scene change detection	Parameters
Layers of convex hull	5
Average sample smoothing (29 parameters) <sup>a</sup>	10
Average sample smoothing (6 parameters) <sup>b</sup>	30
Gradient mask	$G_m = [-1, 0, 1]$
Average gradient smoothing <sup>c</sup>	30
Sliding window size	25
Threshold in zero-crossing result <sup>d</sup>	5

<sup>a</sup> Indicate the process (1)<sup>b</sup> Indicate the process (2)<sup>c</sup> Indicate the process (3)<sup>d</sup> If  $zero\text{-crossing}\ result > threshold$ ; then set a scene change node**Fig. 9** One of the COLD datasets, COLD-Freiburg Part B, used in the experiment for the illustration of Extended-HCT code similarity comparison. **a** COLD-Freiburg dataset frame 1. **b** COLD-Freiburg dataset frame 3. **c** COLD-Freiburg dataset frame 401

the center of the sphere and the field of view with a suitable image size. An approximate perspective image derived from a pinhole camera is then simulated. With this approach, we can deal with the feature point matching problem more properly, and control the image size for the required field of view. The combination of content-based and feature-based image retrieval methods are used for scene recognition. For the content-based image retrieval, we use the color space information to build descriptors and find those with the most similar descriptors as the candidate images. For the feature-based image retrieval, we take the candidate images for SURF feature point matching and find the images with the best correlation. This two-stage approach is used to mark the nodes on the topological map by the retrieval and localization results for place recognition.

**Table 3** The Extend-HCT code similarity comparison for the images shown in Fig. 9

	Features of hull	Cost	Score	Color	A.D. of feature	A.D. of center
Value in Fig. 9a	9.60	0.40	20.00	28.00	41.82	162.63
Value in Fig. 9b	9.60	0.47	18.80	27.25	45.30	154.38
Diff. with Fig. 9a (%)	0.00	15.35	6.00	2.67	7.66	5.07
Value in Fig. 9c	10.60	0.63	18.40	26.75	47.41	148.60
Diff. with Fig. 9a (%)	10.41	37.60	8.00	4.67	11.78	9.44

**Fig. 10** 3D spherical surface mapping used to synthesize a perspective image from six outward camera views. **a** The different colors represent the mesh points of individual cameras. **b** An image cropping result from the Ladybug2 six-camera imaging

#### 4.1 Spherical mapping

To generate a perspective image, the OpenGL texture mapping is first carried out to map the six images obtained from different orientations onto a 3D spherical surface. As shown in Fig. 10a, different colors indicate the mesh points of individual cameras. We only need to handle the overlapping problem among the images on the spherical surface. In the implementation, we utilize an alpha mask on the overlapping area and add blending edge seams among the images, followed by setting the viewpoint at the center of the sphere to synthesize a perspective image.

Since the correspondence matching between the omnidirectional images is less reliable in general, our method can solve the feature point matching problem more easily in the cylindrical image space. An image data index and the 3D coordinates file are built for the image retrieval system. The retrieval image dataset here is for the feature point matching by SURF to identify the imaging location, but the omnidirectional image dataset we use in Sect. 3 is for scene change detection. The matching results will fail if the omnidirectional images are used to match the feature points. We let the six images to map onto a 3D spherical surface using the 3D coordinates file, and verify the result on the 3D spherical surface. Figure 10b shows an example of the partial image cropping results obtained from the captured images.

#### 4.2 Image retrieval for localization

We expect our system to handle the real time image processing task on a vehicle. The reduction of system load and



computational complexity is therefore an important issue in the image retrieval phase. In this work, we use the combination of CBIR and FBIR methods for image retrieval. To reduce the effect caused by the changing lighting conditions, we take the fast CBIR method as a filter to pick similar image candidates, and then use SURF to match the feature points to identify the most similar scenes. Since the feature extraction generally requires more computation, the combined method can significantly reduce the overall image retrieval time.

#### 4.2.1 Content-based image retrieval

The CBIR concept is based on shape, color, texture, spatial correspondence or other information in the images. The descriptions of these visual characteristics are used for image retrieval. In our CBIR, three types of descriptors, CEDD (Chatzichristofis and Boutalis 2008), FCTH (Chatzichristofis and Boutalis 2008) and BTDH (Chatzichristofis and Boutalis 2010), are adopted for the query image encoding. The combination of these algorithms is called CCDs. It is used to compute the similarity between the dataset and the ground truth to pick up a few candidates, and then determine the one with the highest score as output. This method describes the information of different scenes using lower-level feature points, and combines the fuzzy system to classify and index the images. The description of the CCD approach is as follows:

1. Color and edge directivity descriptor (CEDD) contains the color information and edge direction. The edge direction defined by MPEG-7 (edge histogram descriptor, EHD) is widely used in texture classification and indexing. The image descriptor is built by computing the regularity and direction vector. The direction filters are used for the detection of texture information by dividing the texture region into several sub-image blocks. Each image block contains four sub-block defined as  $a_0(i, j)$ ,  $a_1(i, j)$ ,  $a_2(i, j)$ , and  $a_3(i, j)$ , with directions  $F_{vertical}(k)$ ,  $F_{horizontal}(k)$ ,  $F_{diagonal-45}(k)$ , and  $F_{diagonal-135}(k)$ , where the coefficient  $k$  represents the location of sub-blocks. The magnitudes are denoted by  $M_v(i, j)$ ,  $M_h(i, j)$ ,  $M_{d-45}(i, j)$ , and  $M_{d-135}(i, j)$ .
2. Fuzzy Color and Texture Histogram (FCTH) contains fuzzy color information and texture information. FCTH includes  $8 \times 24 = 192$  regions, where the descriptor has a 24-bin color histogram generated by a 24-bin fuzzy texture linking machine in an HSV color space. The texture unit is transformed by Haar Wavelet transform in YIQ color space.
3. Brightness and Texture Directionality Histogram (BTDH) contains brightness information and texture direction histogram.

In general, CEDD and FCTH are the descriptors for environment color and texture information, and BTDH is the descriptor that mingles image brightness and texture features. We combine the CEDD and FCTH as joint composite descriptor (JCD), and build the CCD. The CCD descriptor includes 84 bytes of float number with color, brightness, texture direction and edge direction information. To calculate the distance of the most similar descriptors between different images, the Tanimoto (a.k.a Jaccard coefficient) similarity and Tanimoto coefficient (Chi et al. 1996) are used for evaluation:

$$S(M, N) = \frac{M \cdot N}{|M|^2 + |N|^2 - M \cdot N} \quad (11)$$

$$|M|^2 = \sum_k M_k^2 = \sum_k M_k \quad (12)$$

$$M \cdot N = \sum_k M_k N_k \quad (13)$$

where  $S$  is the score,  $M$  and  $N$  are the values of the descriptors, and  $k$  is the length of the descriptors. We select the most similar descriptors as candidate images by argument of the maximum, and take these candidate images for feature point matching to find the most similar image:

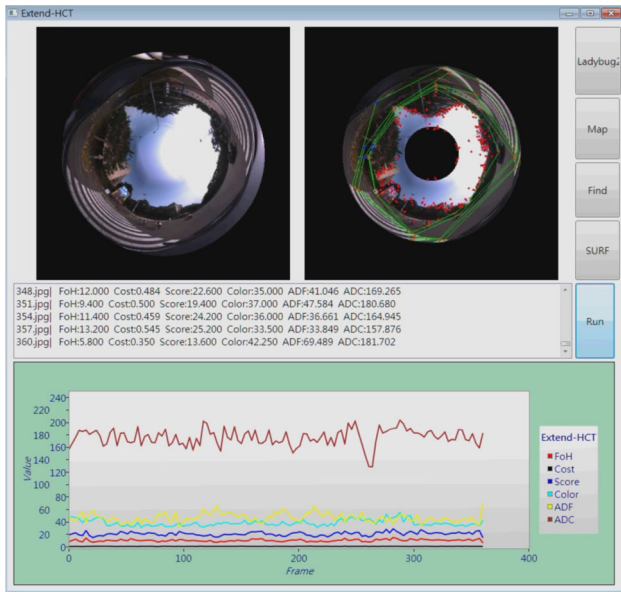
$$D_s = \arg \max \{S_1(M_1, N_1), S_2(M_2, N_2), \dots, S_n(M_n, N_n)\} \quad (14)$$

where  $D_s$  is the set of candidate images.

#### 4.2.2 Feature-based image retrieval

In the feature point matching phase, we first adopt the SURF descriptor as described in Sect. 2 for feature extraction. The FLANN library and  $k$  nearest neighbors ( $k$ -NN) clustering method are then used for feature matching and classification. We calculate the descriptor distance of the query and candidate images. If the descriptor distance is less than a threshold value, then we add one point to the similarity rank score. Finally, the image of highest similarity rank score in the candidate images is the retrieval result. In our application we need to retrieve the current position in a short time, thus we pick up five candidates by content-based image retrieval and then extract the SURF key points to match with the query images.

Nearest neighbor search (NNS) method is also called proximity search or similarity search. There are two search strategies, linear search and space partitioning. In linear search algorithm, it computes the Euclidean distance for all feature points to determine the similarity, and its time complexity is  $O(n^2)$  for a large feature point set. In space partitioning, we use the  $K$ -dimension tree (KD tree) to search



**Fig. 11** Extended-HCT code book construction and scene change detection UI

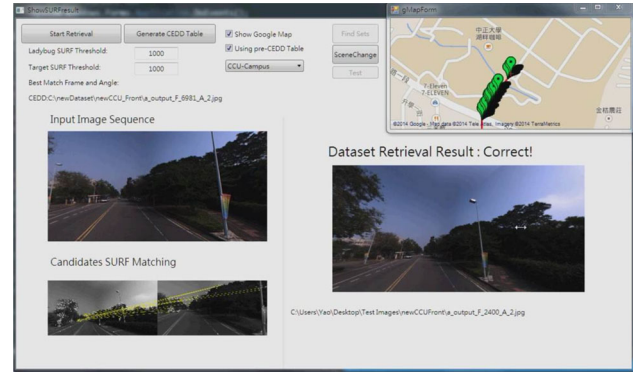
each feature point and build the KD tree structure for matching a pair of feature points at root to subset leaves. The time complexity is  $O(\log n)$ . In the implementation, we build a binary KD tree for  $k$ -NN classification, and set a threshold  $d = 0.6$  to evaluate the similarity score (Cantone et al. 2005). The final result is one of the five candidates closest to the dataset images.

## 5 Implementation and experiments

In this section, we first introduce the user interface of the proposed system, followed by the experiments presented in two parts: (1) topological map construction and (2) image retrieval and localization.

### 5.1 System and user interface

To visualize the Extend-HCT coding process, a graphical user interface (GUI) with convex hull construction and codebook status is developed. As illustrated in Fig. 11, the top-left shows the input omni-directional image and the top-right is the output Extend-HCT result. The processing of each frame contains six parameters as shown in the central block, and a graph of parameter changes over time is shown in the bottom. After the codebook generation, we output the zero-crossing results and scene change nodes. The main window of image retrieval and experiment evaluation are divided into three parts: CBIR, FBIR and the experimental results. In the image retrieval GUI, we integrate the content-based and feature-based methods, and use Google Map to show the localization



**Fig. 12** Image retrieval and localization UI



**Fig. 13** The camera is mounted on top of a vehicle for image acquisition

result (see Fig. 12). The GUI contains the image sequence and the input query image. The retrieval result is shown in GMap.Net SDK tool (Radioman 2013) at the top-right corner with the red and green markers presenting the missing and correct retrieval results, respectively.

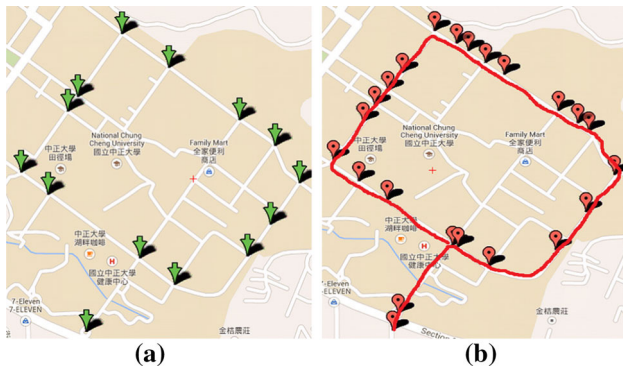
### 5.2 Topological map construction

In the topological map construction stage, we connect a GPS receiver to a laptop to record the GPS information associated with each image frame. The Point Grey Software Development Kit is used in our system development. We install the camera on top of an SUV (see Fig. 13), and capture an image dataset around the CCU campus. The dataset contains 5080 frames and every frame consists of six images. It is used to build an omni-directional image database by fusing the images frame-by-frame. The detail of the dataset (CCU campus) is shown in Table 4.

To reduce the system load, we limited the number of images according to the travel distance. Based on the GPS data associated with each image frame, only one frame is used for processing per 10 m of moving distance. This step can remove the images with similar scenes, and the size

**Table 4** The details of pre-recorded image datasets and test image datasets

Dataset name	Frames	Testimage name	Frames
CCU Front Dataset	369	CCU Front test	204
CCU Back Dataset	369	CCU Back test	204
Bigeat Front Dataset	116	Bigeat Front test	76
Bigeat Back Dataset	116	Bigeat Back test	76

**Fig. 14** The results of scene change node analysis. **a** The ground truth of each intersection (marked with a *green arrow*) around the CCU campus. **b** The scene change nodes are marked in *red*, and the *red lines* indicate the travel route on the topological map (Color figure online)

of the dataset is reduced from the original 5080 frames to 386 frames. We then built the topological map using the reduced image dataset. As shown in Fig. 11, we utilized the Extended-HCT codewords for each omni-directional image in the dataset, and used the codewords for scene change detection. In each intersection, we set a green marker on the map as ground truth (see Fig. 14a). Figure 14b shows the topological map with the scene change nodes (red markers) and the travel route (red lines). We highlight the significant nodes on the topological map and the navigation system only needs to record the order of the nodes. If the navigation system detects an incorrect node order, an error message will be reported to the user. In this experiment, 23 scene change nodes are detected and used for the navigation system.

### 5.3 Image retrieval result

In the image retrieval and localization stage, we implemented the algorithms using C# and Windows Presentation Foundation for GUI development. In content-based image descriptor extraction, the CCD method only needs a set of low level features to describe the different image types. We use Joint Composite Descriptor (JCD) of the CCD method as our content-based image descriptors for feature extraction. The JCD method mainly applies on the HSV color space and fuzzy system as the basis of classification. The descriptors of JCD contain 168-bin codewords by a 168-bin histogram. In

**Table 5** The function parameters setting for feature-based image retrieval

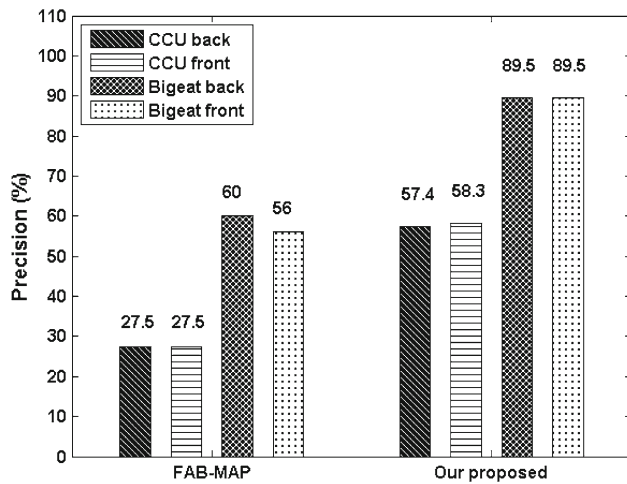
Function parameter	Value
Number (K)	2
SURF Hessian threshold	300.0
KD-trees	4 (trees)
Distance between pairs	0.6
Number of checks	24

our experiments, we built the JCD descriptors in the reduced 3D spherical image dataset with 369 frames. The descriptor data are written in a text file beforehand. Therefore, we only need to use the pre-established descriptor data to calculate the similarity score along with the current input image by Tanimoto similarity and distance. This method can help us to find the most similar candidate images very quickly. According to our experiments and considering the tradeoff between the execution time and accuracy, we set the most similar five candidate images for the later feature-based image retrieval stage.

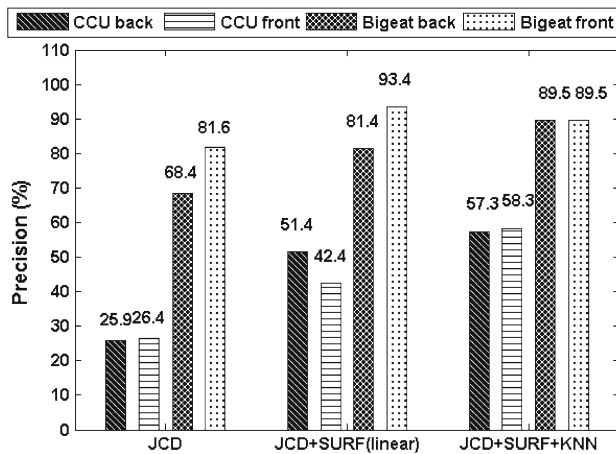
In the feature-based phase, the SURF feature extraction method is adopted as presented in Sect. 4.2.2. We applied Emgu CV 2.4.2 image processing library, and extracted SURF descriptors for FLANN and  $k$ -NN clustering. We built the index value of SURF descriptors of every image by FLANN, and set  $k = 4$  in  $k$ -NN method. In our implementation, the  $k$  value in  $k$ -NN means that we find the  $k$  nearest descriptors of a query image among the candidate images. The experiments show that the descriptor distance value of 0.6 is the best distance threshold (see Table 5).

In our application scenario, the objective is to develop a street view image system like Google StreetView. The experiments were carried out on a vehicle. Two datasets, CCU Front and CCU Back, are created accordingly, and both contain 369 frames. The test images were captured on the same route with a lower image quality and the vertical viewing angle less than  $10^\circ$ . The test dataset is obtained from the CCU campus dataset (with 5080 frames) for every 25 frames, so the two input test datasets (called CCU Front Test and CCU Back Test) both contain 204 frames. The hardware used in the experiments is a PC with Intel Core2 Quad Processor Q8200 and 4GB memory. We set the retrieval result as positive if the location is within 20 m (e.g. two frame intervals) of the GPS coordinates.

The proposed technique is a combination of JCD and SURF with FLANN. We compare the performance with the methods using only JCD and JCD with SURF. The experimental results are shown in Figs. 15 and 16. The system processing time is shown in Table 6. It can be seen that the accuracy of our method is better than others, but the retrieval time is not as fast as the original JCD method (see



**Fig. 15** The image retrieval results obtained using FAB-MAP and our approach



**Fig. 16** The image retrieval results obtained from JCD, JCD+SURF (linear) and JCD+SURF+KNN

Table 7). We also compare the proposed method with FAB-MAP (Cummins and Newman 2008) using our four datasets. The FAB-MAP system only relies on the appearance information for navigation and mapping, which leads to several retrieval instance failures in the experiments. The vocabulary words built by FAB-MAP use Chow Liu tree to compute for the dataset (sps1054142). The results depend on the probability based PDF table. If there are a few dataset images not proper for FAB-MAP, e.g., continuously similar scenes or occluded buildings, then the PDF table will change dramatically. In CCU dataset, it contains a lot of similar scenes such as aligned street trees (see Fig. 17). The results demonstrate that our method can perform well in the city town (Bigeat) with several similar scenes (see Fig. 15). On the other hand, although the system can successfully retrieve more scenery than FAB-MAP, it's still hard to achieve a high precision rate in CCU dataset. Considering the trade-off between the exe-

**Table 6** The system processing time

Process status	Processing time (s)
JCD descriptor searching	0.1
Extract SURF key points from query images	0.15
Extract SURF key points from dataset image	$0.15 \times 5$
KNN matcher	$0.2 \times 5$
Total searching time	2.0 (per image)

**Table 7** The average search time for JCD, JCD+SURF (linear), and JCD+SURF+KNN

Dataset name	Total search time (mm:ss:ms)	Average time per image (s)
<b>CCU Front Dataset</b>		
JCD	03:12:45	0.94
JCD+SURF (linear)	27:43:11	8.15
JCD+SURF+KNN	00:06:36	<b>1.94</b>
<b>CCU Back Dataset</b>		
JCD	03:19:74	0.98
JCD+SURF (linear)	26:36:83	7.64
JCD+SURF+KNN	00:06:43	<b>1.97</b>
<b>Bigeat Front Dataset</b>		
JCD	01:08:02	0.89
JCD+SURF (linear)	07:38:52	6.03
JCD+SURF+KNN	00:01:58	<b>1.55</b>
<b>Bigeat Back Dataset</b>		
JCD	01:08:54	0.90
JCD+SURF (linear)	07:08:98	5.64
JCD+SURF+KNN	00:01:51	<b>1.46</b>

Best is shown in boldface

cution time and accuracy, they are in an acceptable range for a mobile system.

## 5.4 Evaluation

To evaluate the proposed technique, additional experiments are carried out under different environment conditions. We setup the system as described in the previous section, and perform the data acquisition around the CCU campus and a small town. The images are captured with the vertical viewing angle less than  $10^\circ$ , and four datasets, CCU Front, CCU Back, Small Town Front and Small Town Back, are created accordingly. Figures 18 and 19 show the map-represented localization results of CCU Front and Small Town Front, respectively. The scene change node at a correct location is marked in green, otherwise is marked in red. Detailed results are shown in Table 8.

For the results presented in Table 8, the experiment is carried out at about 5 p.m. under a cloudy weather condition. It is seen that the performance is not affected too much, as



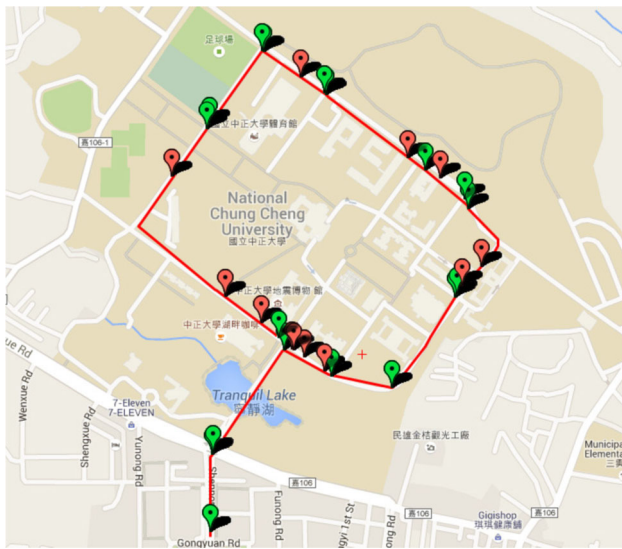


(a)



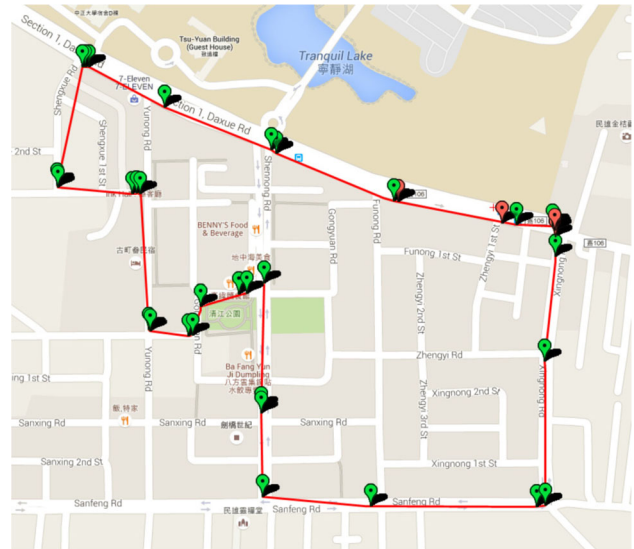
(b)

**Fig. 17** Two similar scenes with road trees in the CCU dataset. They were captured in different locations, and can be distinguished by the proposed method



**Fig. 18** The results of the scene change nodes with CCU Front Dataset

illustrated in the table and Figs. 18 and 19. The extended-HCT depends on the image features, but is not limited to high quality image features. The advantages of our method include the low dimension, fast calculation, using only a vision sensor, requiring less storage capacity and promising visual place recognition results under strict testing on stable and varying illumination conditions. The main limitation of our system is the capability of operating at night or raining



**Fig. 19** The results of the scene change nodes with Small Town Front Dataset

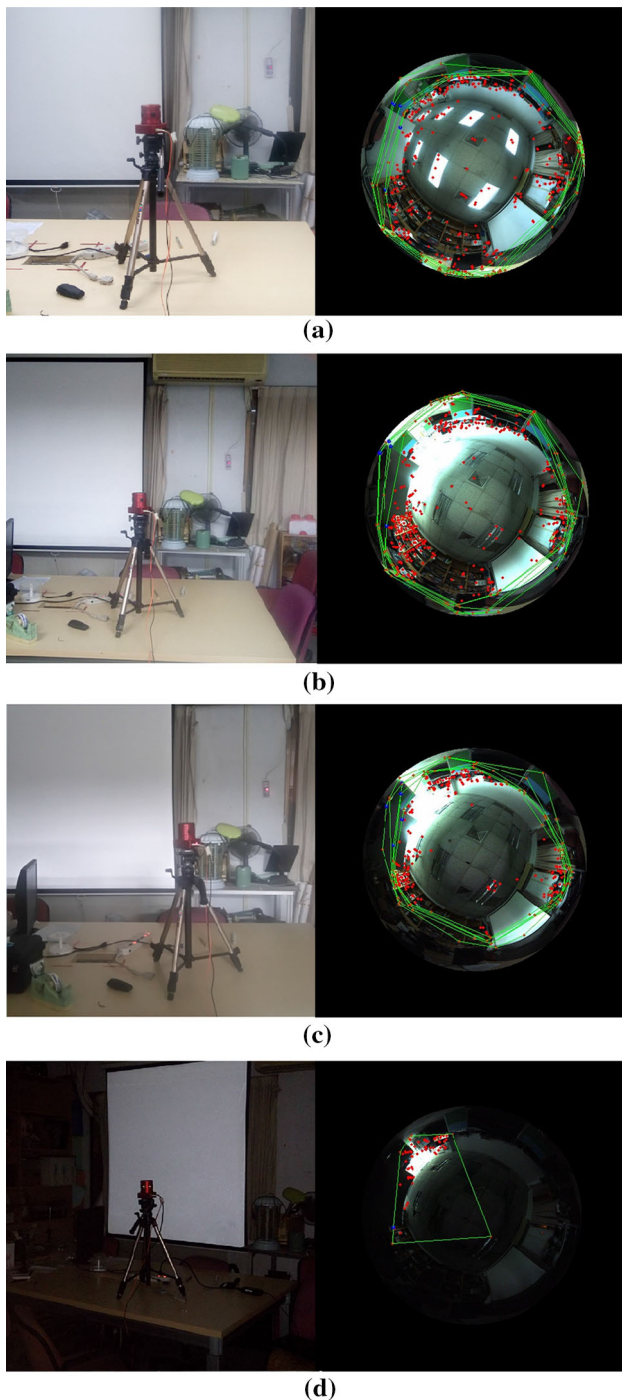
**Table 8** The results of the CCU Campus Dataset and the Small Town Dataset

Dataset name	Frames	Testimage name	Frames	Precision%
CCU Front Dataset	660	CCU Front test	34	61.7
CCU Back Dataset	660	CCU Back test	34	64.7
Small Town Front Dataset	740	Small Town Front test	47	85.5
Small Town Back Dataset	740	Small Town Back test	47	71.4

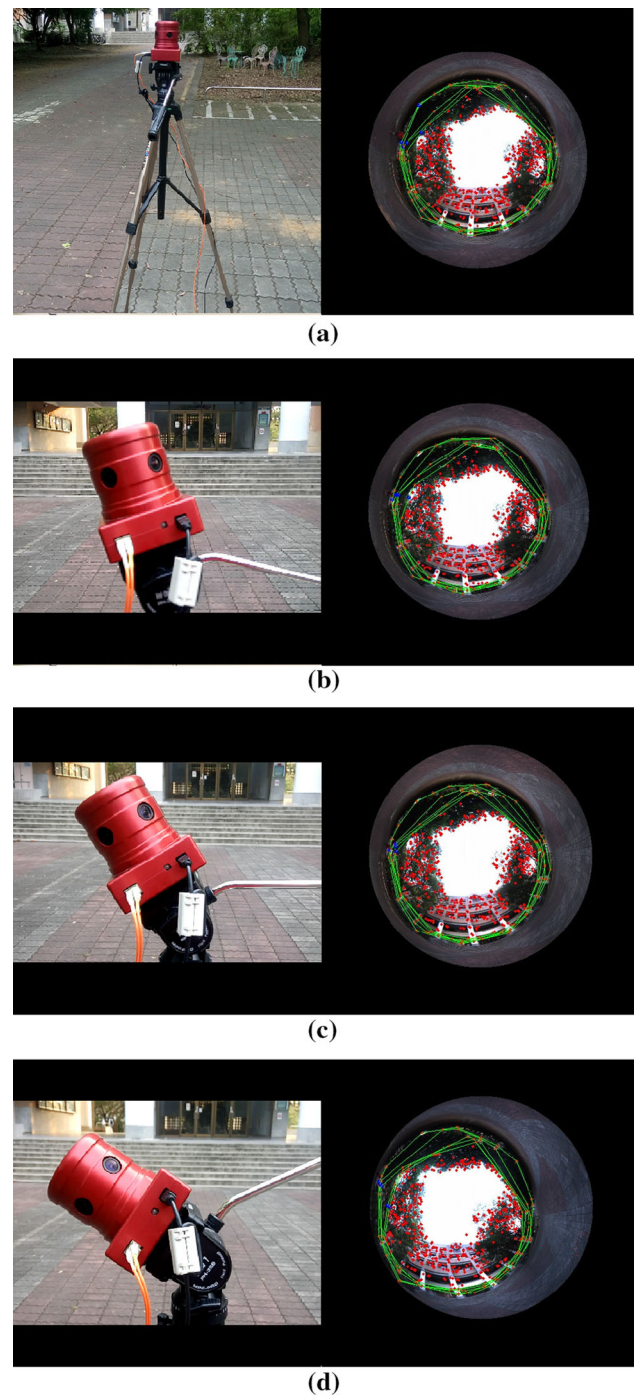
day. To evaluate the the stability of the extended-HCT, we test the algorithm with changing luminance, different camera poses, and dynamic objects in the scene.

### 5.4.1 Evaluation with luminance change

We have evaluated the performance of the topological place recognition using an omnidirectional camera. One major limitation of the Extended-HCT and most recognition techniques is the image brightness. The follows describe the evaluation condition for examining the applicability of our algorithm while the images are captured with decreasing the luminance of the test environment by controlling the ambient light. As shown in Fig. 20, the omnidirectional images and their corresponding Extended-HCT codebooks are obtained with the luminance of 300– 400 lux, and decreased around 50, 70, 90% sequentially. We then evaluate the dissimilarity between the images captured with full illuminance and the rest cases. The dissimilarity score defined by the sum

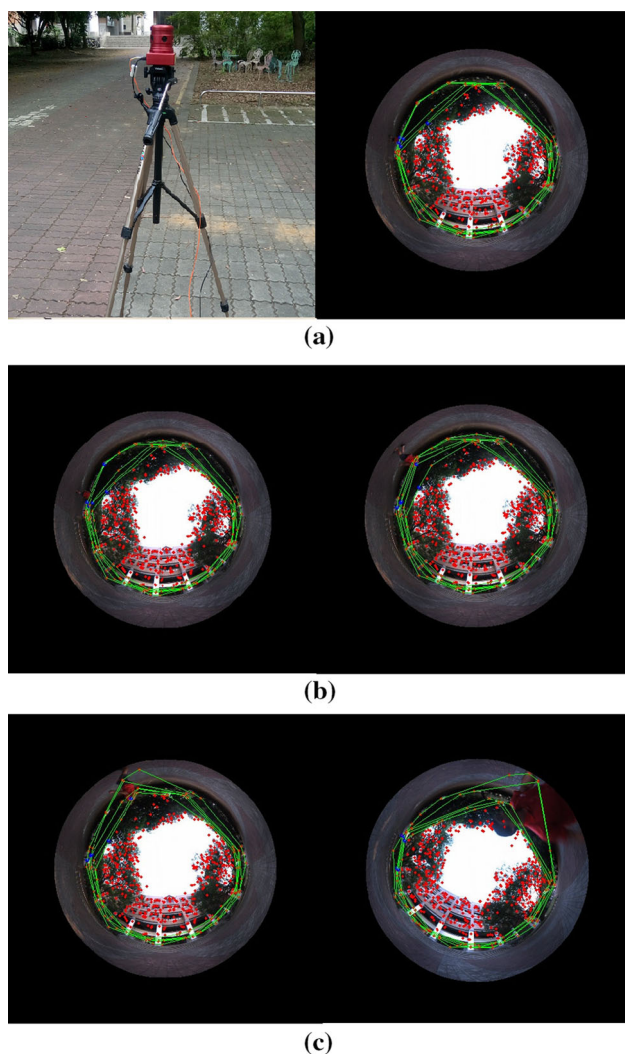


**Fig. 20** The dissimilarity value of the Extended-HCT codewords is evaluated according to the illuminance change of the test environment. The *left figures* are the side view of the environment and the *right figures* are the omnidirectional images with their Extended-HCT codes. **a** The Extended-HCT code and the image at the location are used as the reference (case 1) for evaluating the dissimilarity with different illuminance. **b** The second situation where the luminance is decreased around 50%. The sum of squared differences (SSD) of the Extended-HCT codebook between cases 1 and 2 is 68.46. **c** The third situation where the luminance decrease is around 70%. The SSD value of the Extended-HCT codebook between cases 1 and 3 is 92.32. **d** The fourth situation where the luminance is decreased around 90%. The SSD value of the Extended-HCT codebook between cases 1 and 4 is 5200.04



**Fig. 21** The similarity evaluation of the Extended-HCT codewords related to the tilt angle of the camera. The *left figures* are the side view of the environment and the *right figures* are the corresponding omnidirectional images with their Extended-HCT codes. **a** The Extended-HCT code from the image captured with a normal camera setting placed above the ground at about 140cm. It is used as the reference for evaluating the similarity. **b** In this case the camera is tilted for about 20°. The SSD value of the Extended-HCT codebook is 7.20. **c** In this case the camera is tilted for about 40°. The SSD value of the Extended-HCT codebook is 15.12. **d** In this case the camera is tilted for about 60°. The SSD value of the Extended-HCT codebook is 188.32





**Fig. 22** The Extended-HCT codewords dissimilarity values evaluation according to the changing environment with adding a dynamic object. The *first figure* the view of the capture environment and the *next figures* are the corresponding omnidirectional images with their Extended-HCT codes. **a** The Extended-HCT code and the image at the origin location placed above the ground 140cm, are used as the reference for evaluating the dissimilarity. **b** The changing environment by appearing one person at location 10 and 6m from the origin location. The sum of square different (SSD) value of the Extended-HCT codebook is 12.73 and 20.16, respectively. **c** The changing environment by appearing one person at location 4 and 1m from the origin location. The SSD value of the Extended-HCT codebook is 242.07 and 470.19, respectively

of squared differences is close for the cases in Fig. 20b, c, but it is large for the dark image as shown in Fig. 20b. It is observed that a high dissimilarity value leads to a poor recognition result.

In addition, the number of convex hulls should be assigned in advance. If there are too few interested points detected from the omnidirectional image, even increasing the number of convex hulls will not affect the result. This said, some hulls make zero vectors in the codeword. In our work, we have

successfully combined the image features and color cues for describing the visual places. However, it is hard to say which factor has the highest influence. For example, the color cues might have impacts associated with the weather and features provide the high reliable value while the vehicle navigates in outdoor.

#### 5.4.2 Evaluation with camera balance change

This section discusses the applicability of the proposed technique while the camera's view point is changed. More specifically, we examine the toleration of the algorithm when the camera is not balanced with respect to the ground. As shown in Fig. 21, the omnidirectional images and the corresponding Extended-HCT codebooks are obtained with different camera tilt angles. The similarity between the images captured at the normal orientation and the rest cases is calculated for comparison. Generally speaking, a low similarity score will lead to a poor recognition rate. For the results shown in Fig. 21, the similarity value declines proportional to the tilt angle of the camera. However, it does not affect the stability of the Extended-HCT features for the angle up to  $40^\circ$ .

#### 5.4.3 Evaluation with dynamic object

In this section, the evaluation with a dynamic object in the scene is presented. As shown in Fig. 22, a person appeared at 10, 6, 4 and 1 m away from the camera is used to evaluate the dynamic change of the environment. For these cases, the SSD value of the Extended-HCT codebook is 12.73, 20.16, 242.07 and 470.19, respectively. The result indicates that the algorithm is robust for the environment change with long distance objects, but the close range object affects the stability of the extended-HCT features. In other words, the proposed technique can be used for the applications without dynamic objects within the surrounding 4-meter area.

## 6 Conclusion

In this paper, we present a technique to detect the node information and construct the topological map based on omni-directional image sequences. We apply the combination of CBIR and FBIR methods, and perform the vehicle localization using a recorded image dataset. The Extended-HCT technique is adopted to build the omni-directional image descriptors and detect the scene change nodes. The images captured from different viewpoints on the same location are then mapped to a 3D spherical surface to get approximate perspective images for image retrieval and localization. In the image retrieval and localization stage, we take the fast CBIR method as a filter for picking the most similar candidate images. Thus, the significant image retrieval

time can be reduced. In this experiment, our work can assist to the vehicle navigation based on topological map construction and scene recognition.

**Acknowledgements** Funding was provided by National Science Council Taiwan (Grant No. NSC-99-2221-E-194-005-MY3).

## References

- Cantone, D., Ferro, A., Pulvirenti, A., Recupero, D., & Shasha, D. (2005). Antipole tree indexing to support range search and k-nearest neighbor search in metric spaces. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 535–550. doi:10.1109/TKDE.2005.53.
- Chatzichristofis, S. A., & Boutalis, Y. S. (2008). Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Proceedings of the 6th International Conference on Computer Vision Systems, ICVS'08* (pp. 312–322). Berlin, Heidelberg: Springer-Verlag. <http://dl.acm.org/citation.cfm?id=1788524.1788559>.
- Chatzichristofis, S. A., & Boutalis, Y. S. (2008). Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '08* (pp. 191–196). Washington, DC, USA: IEEE Computer Society. doi:10.1109/WIAMIS.2008.24.
- Chatzichristofis, S. A., & Boutalis, Y. S. (2010). Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. *Multimedia Tools and Applications*, 46(2–3), 493–519. doi:10.1007/s11042-009-0349-x.
- Chi, Z., Yan, H., & Pham, T. (1996). *Fuzzy algorithms: With applications to image processing and pattern recognition*. River Edge, NJ, USA: World Scientific Publishing Co., Inc.
- Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3), 462–467. doi:10.1109/TIT.1968.1054142.
- Conte, G., & Doherty, P. (2009). Vision-based unmanned aerial vehicle navigation using geo-referenced information. *EURASIP Journal on Advances in Signal Processing* 2009(1), 387,308. doi:10.1155/2009/387308. <http://asp.eurasipjournals.com/content/2009/1/387308>.
- Cummins, M., & Newman, P. (2008). Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6), 647–665. doi:10.1177/0278364908090961.
- Eakins, J., Graham, M., Eakins, J., Graham, M., & Franklin, T. (1999). Content-based image retrieval. *Library and Information Briefings*, 85, 1–15.
- Huang, C. L., & Liao, B. Y. (2001). A robust scene-change detection method for video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(12), 1281–1288. doi:10.1109/76.974682.
- Larnaout, D., Gay-Bellile, V., Bourgeois, S., & Dhome, M. (2014). Vision-based differential gps: Improving vslam / gps fusion in urban environment with 3d building models. In *2014 2nd International Conference on 3D Vision (3DV)*, (Vol. 1, pp. 432–439). doi:10.1109/3DV.2014.73.
- Lee, S. W., Kim, Y. M., & Choi, S. W. (2000). Fast scene change detection using direct feature extraction from mpeg compressed videos. *IEEE Transactions on Multimedia*, 2(4), 240–254. doi:10.1109/6046.890059.
- Lin, H., Lin, Y., & Yao, J. (2013). Scene change detection and topological map construction using omnidirectional image sequences. In *Proceedings of the 13. IAPR International Conference on Machine Vision Applications (MVA 2013)* (pp. 57–60). Kyoto, Japan, 20–23 May 2013. <http://www.mva-org.jp/Proceedings/2013USB/papers/04-06.pdf>.
- Liu, M., Alper, B., & Siegwart, R. (2013). An adaptive descriptor for uncalibrated omnidirectional images - towards scene reconstruction by trifocal tensor. In *2013 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 558–563). doi:10.1109/ICRA.2013.6630629.
- Liu, M., & Siegwart, R. (2012). DP-FACT: Towards topological mapping and scene recognition with color for omnidirectional camera. In *2012 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3503–3508). doi:10.1109/ICRA.2012.6225040.
- Liu, M., & Siegwart, R. (2014). Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera. *IEEE Transactions on Robotics*, 30(2), 310–324. doi:10.1109/TRO.2013.2272250.
- Maier, D., & Kleiner, A. (2010). Improved gps sensor model for mobile robots in urban terrain. In *2010 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4385–4390). doi:10.1109/ROBOT.2010.5509895.
- Milford, M., & Wyeth, G. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1643–1649). doi:10.1109/ICRA.2012.6224623.
- Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). doi:10.1109/CVPR.2007.383172.
- Pronobis, A., & Caputo, B. (2009). Cold: The cosy localization database. *The International Journal of Robotics Research*, 28(5), 588–594. doi:10.1177/0278364909103912.
- Radioman. (2013). Gmap.net - great maps for windows forms & presentation, <http://greatmaps.codeplex.com>.
- Ranganathan, A. (2012). Pliss: Labeling places using online change-point detection. *Autonomous Robots*, 32(4), 351–368. doi:10.1007/s10514-012-9273-4.
- Rituerto, A., Murillo, A., & Guerrero, J. (2014). Semantic labeling for indoor topological mapping using a wearable catadioptric system. *Robotics and Autonomous Systems* 62(5), 685 – 695. doi:10.1016/j.robot.2012.10.002. <http://www.sciencedirect.com/science/article/pii/S0921889012001856>. (Special issue semantic perception, mapping and exploration).
- Shahraray, B. (1995). Scene change detection and content-based sampling of video sequences. *Proceedings of SPIE*, 2419, 2–13. doi:10.1117/12.206348.
- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12), 1349–1380. doi:10.1109/34.895972.
- Stone, D., Differt, Milford, M., & Webb, B. (2016). Skyline-based Localisation for Aggressively Manoeuvring Robots using UV sensors and Spherical Harmonics. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016*. Stockholm, Sweden, May 16–21. doi:10.1109/ICRA.2016.7487780.
- Toledo-Moreo, R., Zamora-Izquierdo, M., Ubeda-Miarro, B., & Gmez-Skarmeta, A. (2007). High-integrity imm-ekf-based road vehicle navigation with low-cost gps/sbas/ins. *IEEE Transactions on Intelligent Transportation Systems*, 8(3), 491–511. doi:10.1109/TITS.2007.902642.
- Wang, M. L., & Lin, H. Y. (2010). A hull census transform for scene change detection and recognition towards topological map building. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 548 –553). doi:10.1109/IROS.2010.5652817.



Wang, M. L., & Lin, H. Y. (2011). An extended-hct semantic description for visual place recognition. *The International Journal of Robotics Research*, 30(11), 1403–1420. doi:[10.1177/0278364911406760](https://doi.org/10.1177/0278364911406760).



**Hwei-Yung Lin** is a professor and the director of Robot Vision Lab at National Chung Cheng University, Taiwan. He received his Ph.D. degree in electrical and computer engineering from the State University of New York at Stony Brook, United States. In 2002 he joined the Department of Electrical Engineering, National Chung Cheng University, as an assistant professor and currently a full professor. He is the author of more than 110 journal and conference articles, and has written

ten two book chapters. He also holds 10 US patents and 8 Taiwan patents. He serves as an organizing committee member and a program committee member of more than 50 international conferences. His research interests include computer vision, robotics, machine learning, and image processing. He is a senior member of OSA and a senior member of IEEE.



**Chia-Wei Yao** received his B.S. degree in computer science and information engineering from National Taiwan University of Science and Technology, Taiwan, and M.S. degree in electrical engineering from National Chung Cheng University, Taiwan. He has been engaging in the research areas of computer vision and image processing. He is currently with Moxa Inc.



**Kai-Sheng Cheng** received his B.Eng from National Dong Hwa University in 2014, and M.Eng from National Chung Cheng University in 2016. His research interests focus on computer vision, in particular include 2D/3D recognition system. He is now with Cadence Design Systems, Inc.



**Van Luan Tran** is a Ph.D. student at Department of Electrical Engineering, National Cheng Cheng University, Taiwan. Before that, he received a MS degree in electrical and electronic engineering from the HCMC University of Technology and Education in 2015. His research interests are in the areas of robot vision. Specifically, he focuses on localization, developing RGB-D camera, object recognition and 6D pose estimation.