Fusion of Visual and Compass Sensors for Location Recognition

Xuejie Zhang, Leng Phuan Tay
Information Systems Division
Temasek Laboratories@NTU
Singapore
zhangxuejie@ntu.edu.sg, aslptay@ntu.edu.sg,

Brian Ji Hua Ang, Gee-Wah Ng
Information Division
DSO National Laboratories
Singapore
ajihua@dso.org.sg, ngeewah@dso.org.sg

Abstract - **This paper presents a fusing methodology through an illustration of how visual and compass sensors for location and route recognition can be combined. The data acquisition platform consists of two cameras, a compass sensor and several peripheral inputs such as a GPS and an accelerometer. With just the visual features and compass directions, a Fast Learning Artificial Neural Network (FLANN) is used to learn and recognize locations. Through video sequence training and corresponding compass information, the system was able to automatically learn key visual and compass information from key locations and routes. This combination of visual and compass information provides the possibility for visual navigation. While there are only two sensor-type dimensions, this serves as a blue print for extensions onto other sensor fusing capabilities.**

Keywords-visual recognition; compass; neural networks

I. INTRODUCTION

There are methodologies for location recognition in the literature. Range sensors such as laser and sonar have been used for robot localization [1-3]. Vision based location recognition has attracted much attention since visual sensors are low-cost and easily available [4-8]. Ulrich and Nourbakhsh presented a topological localization model using color histograms and a voting method [9] where the color histograms were generated from images captured by a 360 degrees panoramic camera. Wu and Rehg proposed PACT representations that encode local and global shape information for location recognition [8]. Siagian and Itti used bottom-up vision attention based early-visual features to capture the 'gist' of the scene into a low-dimensional signature vector and used it for location recognition [6]. The Bag-of-Words model (BOW) in the literature of image classification has also inspired the use of it in vision based localization [7, 10, 11]. The BOW models initially retrieve preliminary visual descriptors such as SIFT and SURF descriptors from an input image. After this, the formation of the image representation is produced as a visual word frequency histogram which is generated by classifying the visual descriptors according to a pre-learned [12] or incremental [11] visual word dictionary (a set of representative descriptors). Angeli et al. proposed a model embedding metrical information from robot odometry to BOW based location recognition [13]. In the area of sensor fusion, Pronobis et al. presented a multi-modal place classification system for indoor environment by fusing multiple visual cues and laser range data [14]. A Support Vector Machine (SVM) was used in this work to learn the optimal combination of each cue. Oskiper et al. presented a multi-model system using an error state Kalman filter algorithm for camera tracking based on fusing visual odometry, visual landmark matching and range measurements from radio frequency (RF) ranging radios [15]. Fusing different types of sensors or localization cues has been shown to provide more accurate localization performance [14, 15].

This paper presents a location recognition system by learning the fused visual and compass information. The visual descriptors were generated using a BOW-based method. The visual descriptors combining with compass reading provide a possibility for recognition and thus navigation. A neural network model using the Fast Learning Artificial Neural Network (FLANN) [16] was used to fuse the visual and compass data. The BOW-Compass clusters were generated by FLANN and used for recognition in the testing phase. The unique feature of the FLANN is its ability to make allowance for incremental learning of BOW-Compass clusters thus providing a very flexible way of training and recognition.

The organization of this paper is as follows. Section II presents the details of the proposed model. Hardware descriptions, fusion of visual and compass descriptors, learning and recognition mechanisms are presented. Section III presents experimental results including key location and route learning and recognition. Section IV concludes the paper.

II. THE SYSTEM

The system used entails the combination of hardware and software modules. This segment introduces the commercially-off-the-shelf (COTS) hardware and describes the algorithm used.

A. Sensors

The hardware platform includes forward and rear facing cameras, a compass, a GPS sensor and an accelerometer. All the sensors were mounted on the flat-top surface of the test vehicle and connected to a laptop through USB (Universal Serial Bus) and RS232 serial port for data acquisition. Results of the initial study to combine image and compass data for location recognition are discussed in a later section. Data from

other sensors were recorded but not currently used in the fusion exercise. This is because our current aim is to establish a viable methodology that allows for incremental combination of sensor inputs so that each sensor contributes significantly towards localization. The emphasis is thus a proof-of-concept rather than a brute forced illustration of multiple sensor fusion. The platform, assembled from commercially-off-the-shelf (COTS) components, was able to capture visual and compass data at a speed of 10 fps. Fig. 1 shows the assembled structure of our data acquisition system. One advantage of using forward and rear-facing cameras is the ease of reversal image and compass swaps to increase virtual data points. The forward and rear-facing cameras provide stable and robust visual signatures and shares similar considerations for the color histogram generated from panoramic cameras [9].

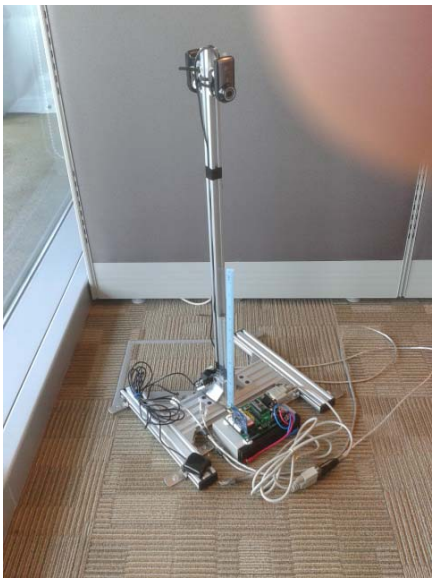

Figure 1. The cameras and sensors on mounting frames

B. SURF Detection and Descriptor

Due to the need for expedient image information extraction, the Speeded-Up Robust Features (SURF) [17] and the Bag-Of-Words (BOW) [12] model were used to create the visual descriptors.

The SURF model includes a Fast-Hessian interest point detector which relies on integral images to improve the detection speed [17]. Fig. 2 illustrates the interest point detection by the SURF detector. It was able to detect fine and distinct features in the image. After detecting interest points, a 64-dimensional SURF descriptor is retrieved for each interest point by modeling the distribution of Haar-wavelet responses within the interest point neighborhood [17].

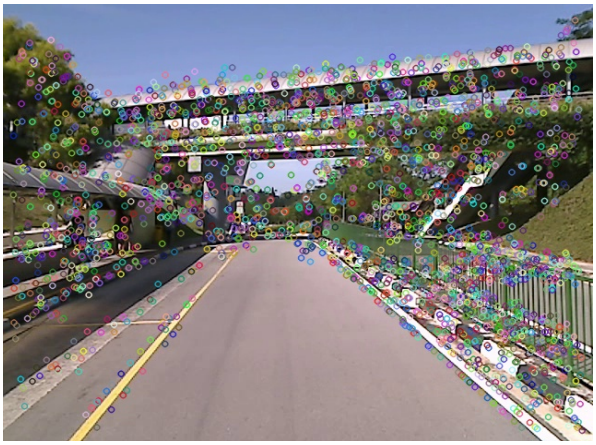

Figure 2. SURF feature detection

C. Bag- of-Words Descriptor

The Bag-of-Words (BOW) model that was inspired from natural language processing, describes an image using a visual dictionary. The BOW categorized documents based on its words and occurring frequencies. 'Bag' was referenced to signify a word count rather than a sequence and thus the order of the words is ignored in the representation. The visual dictionary is generated using visual descriptors, such as SURF, retrieved from training images. The visual descriptors for BOW representation can either be generated from regular grid [18, 19], interest point detectors [12] or randomly selected points [20]. The SURF feature detector in this case was used to generate interest point for SURF descriptor retrieval. Fig. 3 shows a typical BOW model where the visual dictionary is generated using K-means clustering and the BOW descriptors are generated as a visual word frequency histogram.

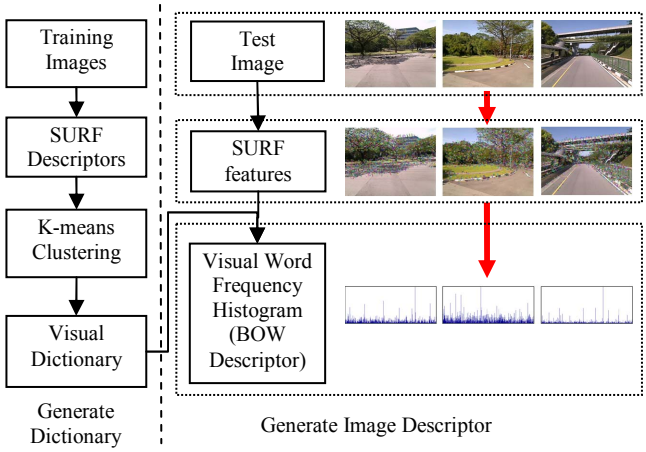

Figure 3. Flowchart of dictionary learning and BOW feature retrieval

In our system, to increase the learning speed, 100 images from a training video are randomly selected to retrieve about 300,000 SURF descriptors. From these descriptors only 20,000 were randomly selected as the training set. The training set was clustered by a K-means clustering model to generate 1000 clusters or 'visual words'. These 1000 visual words form the

visual dictionary and serve as the primitives for classifying the SURF descriptors generated from images in the location recognition phase.

The process of generating BOW descriptor for an input image is as follows. SURF descriptors are extracted and classified using the visual dictionary and visual word frequency histogram is generated based on the classification result. This visual word frequency histogram is called the Bag-of-Word (BOW) descriptor of the input image and it is used as the raw feature for location recognition.

A combination of the BOW descriptors from the front and the rear cameras was used. The back BOW descriptor was directly appended to the front BOW descriptor to form the visual descriptor at a point in the environment. This causes the visual descriptors in the system to become 2000 dimensions. To make full use of the training data the front and rear camera images are systematically swapped and the corresponding inversed compass directions to make a 'virtual' captured data.

Visual descriptors and compass readings were fused into a data structure using a neural network to enable training and recognition.

D. Fast Learning Artificial Neural Network

The Fast Learning Artificial Neural Network (FLANN) is a network that can perform online clustering [16, 21]. It has been used in applications such as image segmentation [16, 22], robot localization [2] and object recognition [23]. Fig. 4 shows the flow chart of FLANN. Basically the input data can be sequentially fed to the network. For each data, a vigilance test is used to choose nodes that are similar to this data. If there is some similar nodes, the closest similar node is considered as the representative node for the input data. Otherwise, the input data pattern is used to create a new node. This is an incremental process where the node created covers a subspace in the feature space and any input data not falling into the coverage of existing nodes will create a new node.

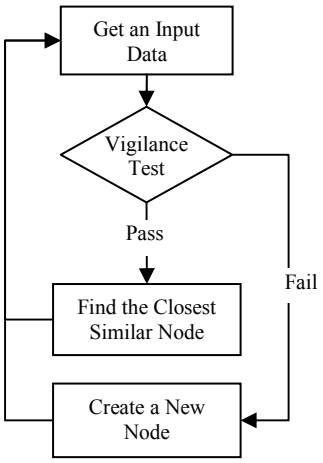

Figure 4. The flow chart of FLANN

The detailed algorithm of FLANN is shown below.

Step 1 Network Initialization:
The input data are d-dimensional vectors.
Set vigilance ρ and the d-dimensional tolerance δ.

Step 2 Present an input data pattern x to the network.
If the network have no output node
GOTO Step 6.

Step 3 Vigilance Test:
Determine matching output nodes by the vigilance test:

$$V = \frac{\sum_{i=0}^{d} D\left[\delta_i - \left|w_{ji} - x_i\right|\right]}{d} \geq \rho \qquad (1)$$

where $D[a] = \begin{Bmatrix} 1, a > 0 \\ 0, a \leq 0 \end{Bmatrix}$

V is the vigilance score and D is the discriminating function imposed on the difference between the tolerance and the absolute difference between the j^{th} node w_j and the input data x along each dimension i.

Step 4 Maintain a list of all matching nodes that fulfilled Step 3.
Determine a single winner cluster from the matching list using the winner-take-all criteria below

$$winner = \min_j \left[\sum_{i=0}^{d} \left(w_{ji} - x_i\right)^2 \right]$$

Step 5 If winner is found
GOTO Step 2.
Else
GOTO Step 6.

Step 6 Create a new output node o and assign vector $w_o = x$. Record the label of the node if applicable.
GOTO Step 2.

The tolerance, δ, in the algorithm presented controls the extent at which the individual elements of the vector can vary. This value can at times be heuristically determined by examining the standard deviations of the training data for each dimension [21]. During the initial stages of the algorithm, the FLANN will create the very first new cluster. In subsequent data samples arriving, the FLANN uses a vigilance test to find all the matching clusters from existing clusters. The Vigilance value sets the threshold for determining which cluster a data belongs to. In the case where no suitable clusters are found, the data in question naturally becomes a novel set and thus a new discovery and can form a new cluster or it can be classified as unknown in visual recognition. In the training phase, the vigilance is set to 0.7, implying that two data patterns are considered visually similar if 70% of all dimensions match in their given tolerances.

Each FLANN generated node is labeled. During the testing phase, the trained FLANN network is used as the means to classify the input visual scene. At Step 3 of the algorithm, if it generates no matching node for an input data pattern, the input data is classified as unknown because the existing FLANN network does not contain any nodes similar to it. If Step 3 generates some matching nodes, Step 4 will be used to determine the label of the input data pattern.

E. Fusion of Visual and Compass Information

The combined information combining visual and compass sensors provide possibilities for better recognition and navigation. FLANN is used for the fusion of visual and compass information. This fusion is done in the vigilance test step. Since it is possible to obtain a vigilance scores for both the visual and compass data, the key is to allocate the relevant significant contributions from each entity. The vigilance scores of visual and compass can thus be combined using a weight. The final vigilance test in (1) becomes

$$V = aV_v + bV_c \geq \rho \quad (2)$$

Where V_v is the visual vigilance score and V_c is the compass vigilance score. Since the compass is only 1-dimension, V_c is derived directly from the difference between the compass direction of the input data and that of the nodes. In the specific model used was set to 5/6 and b was set to 1/6 which implies that visual consideration is the main contributor when recognizing a location. However, while the visuals provided a major contribution, the compass direction also provides for a higher precision to the recognition process. As a result, not only does a place need to look similar, the compass direction must also be similar in order for a correct match to occur. Fig. 5 illustrates the training process. The result of FLANN processing is a set of BOW-Compass clusters.

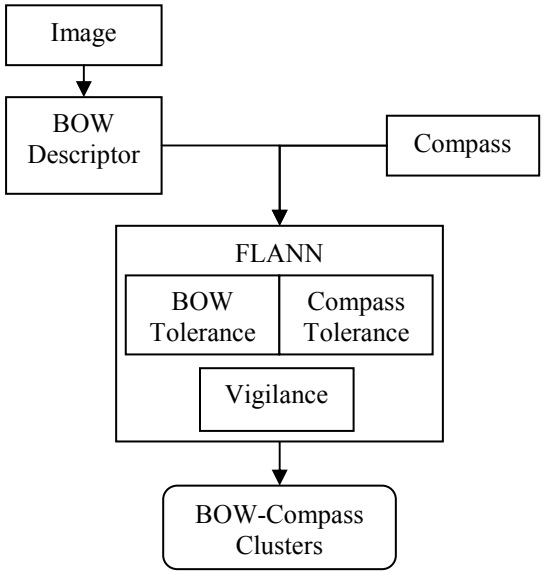

Figure 5. Fusion of visual and compass information

F. Key Location Representation

We define a key location as a place with a neighborhood of approximately 20 meters. Fig. 6 shows an aerial map of the university campus with the three key locations selected for our experiments. Each segment is stored as a route between two locations and there is one route between every two key locations in this map. Data was recorded using the sensors shown in Fig. 1.

Each key location is represented by a set of visual descriptors with corresponding compass data and its information is encapsulated in a FLANN network for storing key location descriptors (BOW and compass). Video sections of key locations and the corresponding compass readings were fed into corresponding FLANN network for the generation of representative BOW-Compass clusters. As a result of the multiple perspectives available in each 20 meters region, each key location can contain several BOW-Compass cluster representations.

G. Route Representation

A route is defined as the path between two key locations and it does not include visual and compass information of its two ending key locations. This is illustrated in Fig. 6. The route is also learned by using video sections and corresponding compass readings. One difference between route learning and key location learning is that the BOW-Compass clusters of the two ending locations of the route are included in the FLANN network at the beginning of route learning. This means that the route learning only memorizes those descriptors that are different from its two ending locations. Each segment sandwiched by key locations is stored in separate FLANN networks.

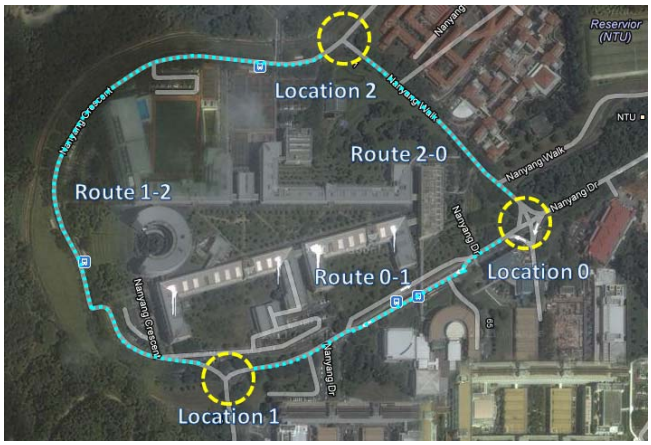

Figure 6. The testing environment in NTU campus. Image was extracted from Google Maps.

III. Experimental Results

The key location and route recognition results are presented in this section. Two cycles of data acquisition were conducted. One round is used for key location and route learning and the other round is used for recognition testing.

A. Experimental Setup

In the learning stage, the vigilance of FLANN was set to 0.7 to provide for a strict learning experience. The BOW tolerance was set to 1 times the standard deviation of each dimension of the BOW descriptors of 1000 images randomly selected from the training video. The compass tolerance was set to 10 degrees.

During the testing stage, the vigilance was lowered slightly to cater for the possibility of slight changes in the scene or view perspectives.

A higher vigilance value is often used in the training stage so that the clustering is more rigid and covers a more complete set of information. The relaxation of vigilance in the testing phase allows for a more flexible matching. When the vigilance is set to a sufficiently low value (0 is the extreme case), the FLANN model works similar to a nearest neighbor classifier because all clusters will pass the vigilance test and go to the winner-take-all stage.

B. Key Location Training and Recognition

Video sections and corresponding compass readings of a training data set were used for training. The number of clusters of trained key locations is shown in Table I. The three key locations generated an equivalent number of clusters.

Video sections and accompanying compass readings were used to test the performance of key location recognition. The number of video frames for testing is shown in the fourth column of Table I. To examine the effectiveness of the Vigilance testing as a first stage filter, a closer scrutiny was done on the results immediately after the Vigilance stage. It was found that there were fewer passes for the combined BOW-compass (~10) combination as compared to the discrimination with only the BOW (~25 – 30) vector. This implies that the BOW-compass fusion provides an efficient pre-selection of matching nodes which greatly reduces the second discrimination computation during the winner-take-all stage of the FLANN algorithm. On the other hand, the individual tolerance setting for each sensor addresses the problem of uncertainty in sensor data such as low precision data. The FLANN is able to capture the similarity of the inaccurate input data and existing nodes as long as the error or uncertainty in sensor falls in the tolerance range. The fusion of sensors at the vigilance stage is linear fusion. However, the tolerance stage resembles a non-linear model since the vigilance ratio reflects the percentage of dimensions matching instead of the direct difference between vectors. With the tolerance and vigilance components, the FLANN is able to learn and recall data patterns fast and efficiently.

TABLE I. TRAINING AND TESTING OF KEY LOCATION RECOGNITION, COMPASS TOLERANCE=10, VIGILANCE=0.7, VIGILANCE WEIGHT = 5/6 AND 1/6 FOR BOW AND COMPASS RESPECTIVELY.

Key Location	*Number of BOW-Compass Clusters in Training*	*Number of Video Frames for Training*	*Number of Video Frames for Testing*
0	17	83	106
1	16	67	91
2	18	57	67

Table II shows the result of testing with vigilance 0.7 where some unknown locations were encountered and no matching nodes are generated in the vigilance test for these locations. However, if the vigilance was lowered to 0.6, as shown in Table III, all places were recognized successfully and there were no occurrences of unknown matches.

TABLE II. TESTING ACCURACY FOR KEY LOCATION RECOGNITION. COMPASS TOLERANCE = 10, VIGILANCE=0.7, VIGILANCE WEIGHT = 5/6 AND 1/6 FOR BOW AND COMPASS RESPECTIVELY.

True\Recognized	*0*	*1*	*2*	*Unknown*
0	100%	0	0	0
1	0	94.51%	0	5.49%
2	0	0	97.01%	2.99%

TABLE III. TESTING ACCURACY FOR KEY LOCATION RECOGNITION. COMPASS TOLERANCE = 10, VIGILANCE=0.6, VIGILANCE WEIGHT = 5/6 AND 1/6 FOR BOW AND COMPASS RESPECTIVELY.

True\Recognized	*0*	*1*	*2*	*Unknown*
0	100%	0	0	0
1	0	100%	0	0
2	0	0	100%	0

C. Route Learning and Recognition

Some sample images corresponding to the generated BOW-Compass clusters for each route is shown in Fig. 7. The generated nodes are all very useful representatives of the routes.

Together with Fig. 7, Table IV shows that a more complex route generates more BOW-Compass clusters. Route 0-1 is the most complex route with more buildings and so it generated the largest number of clusters (156). Route 1-2 is a road with fewer buildings but more trees along the road. Although the road appearances are similar it still generated 79 clusters since it was the longest route in the map. Route 2-0 is the shortest route with many trees along the road and it generated only 24 clusters.

Table IV summarizes the results described in this section.

TABLE IV. TRAINING OF ROUTES, COMPASS TOLERANCE=10, VIGILANCE=0.7, VIGILANCE WEIGHT = 5/6 AND 1/6 FOR BOW AND COMPASS RESPECTIVELY.

Route	*Number of BOW-Compass Clusters*
0-1	156
1-2	79
2-0	24

Figure 7. Generated key images for each route

The route recognition results are shown in Fig. 8-10. The vigilance used in this testing is 0.6.

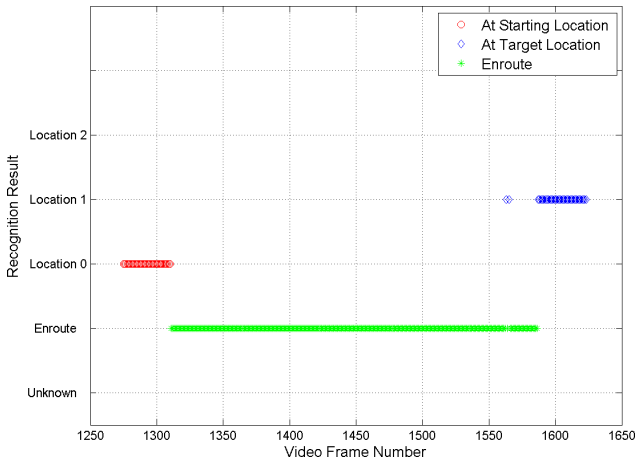

Figure 8. Location recognition chart showing route recogition for route 0-1.

Route 0-1 was recognized very accurately, as shown in the location recognition chart in Fig. 8. The starting and ending locations were also recognized at the start and end of testing. Only a few data points were recognized wrongly when reaching location 2, which is acceptable.

Fig. 9 shows the location recognition chart for route 1-2. At a very short section in the middle of the route there are some points recognized wrongly. Overall speaking, the system was still able to effectively recognize the route and the locations.

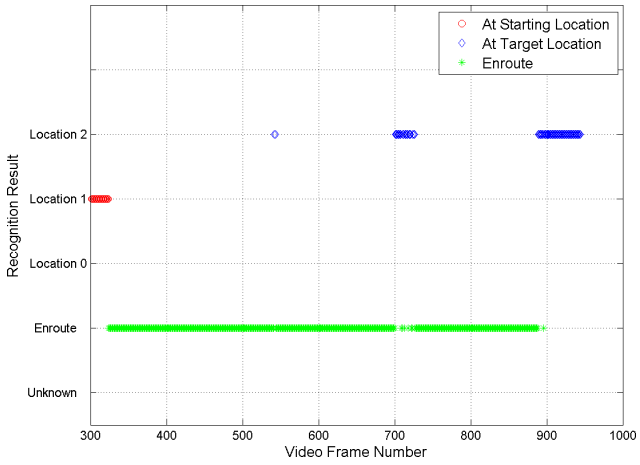

Figure 9. Location recognition chart showing route recogition for route 1-2.

Fig. 10 shows the location recognition chart for route 2-0. There are a number of samples recognized as location 2 when reaching location 0. When the vigilance was lowered to 0 the nearest neighbor performance of FLANN was similar. This indicates that location 2 and location 0 share some common features and the BOW-Compass clusters are similar. It was also found that there is no node generated for the second part of the route nearer to the location 0. This means the FLANN nodes generated for the first half of route are similar to the data points in the second half, thus many data points in the second half of the route was recognized as location 2. To resolve this, a more rigid learning of the route can be made by setting to a slightly higher vigilance. However, to make the experimental results more comparable, the vigilance setting was kept the same for the three route testing cases. In fact, for different input data, the vigilance may not be fixed. The vigilance should be set so that sufficient FLANN nodes are generated to fully represent a route. Further investigations on how to effectively set a flexible vigilance setting for the training of different input data are in the pipeline.

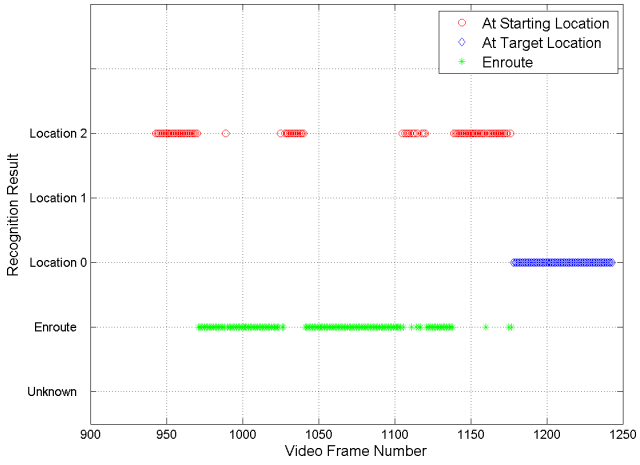

Figure 10. Location recognition chart showing route recogition for route 2-0.

IV. CONCLUSIONS

A location recognition system using a Fast Learning Artificial Neural Network for sensor fusion is presented. Representative nodes are generated automatically for key locations and routes. The learned FLANN network was used for key location and route recognition. The system developed was able to classify key locations at a high accuracy.

Combining different types of sensors provides improved recognition capability and speed. For example, visual descriptors are excellent features for localization. However, visual descriptor is usually used to construct a topological map. This is the furthest it can perform unless an accurate visual odometry model can be developed. Topological maps alone cannot provide sufficient information for navigation. In order for navigation to become possible, it will be necessary to combine compass sensors or other information about the vehicle's position so that a direction for navigation to the target location can be derived. In this paper a simple fusion of visual and compass sensors is presented. However, the framework enables utilization of more sensors for fusion. The advantage of the proposed framework is that a sensor can incrementally be included in the fusion by a suitable tolerance setting and vigilance test. The parameter settings for each sensor can be tuned separately to provide optimal performance. Simply embedding anther attribute for vigilance test enables fusion of more types of sensors.

In future investigations on more flexible FLANN configurations for learning of different routes will be conducted. A more complex testing location and map are also necessary for further evaluation of the proposed algorithm.

REFERENCES

[1] S. Thrun, *et al.*, "Robust monte carlo localization for mobile robots," *Artificial Intelligence,* vol. 128, pp. 99-141, May 2001.

[2] A. L. P. Tay, *et al.*, "Biologically inspired KFLANN place fields for robot localization," in *International Joint Conference on Neural Network*, 2006, pp. 4201-4208.

[3] J. J. Leonard and H. F. Durrantwhyte, "Mobile robot localization by tracking geometric beacons," *IEEE Transactions on Robotics and Automation,* vol. 7, pp. 376-382, Jun 1991.

[4] A. Torralba, *et al.*, "Context-based vision system for place and object recognition," in *International Conference on Computer Vision*, 2003, pp. 273-280.

[5] J. Q. Wang, *et al.*, "Coarse-to-fine vision-based localization by indexing scale-invariant features," *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics,* vol. 36, pp. 413-422, Apr 2006.

[6] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 300-312, Feb 2007.

[7] G. Schindler, *et al.*, "City-scale location recognition," in *Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1378-1384.

[8] J. N. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," *2008 Ieee Conference on Computer Vision and Pattern Recognition, Vols 1-12,* pp. 2221-2228, 2008.

[9] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *IEEE International Conference on Robotics and Automation*, San Francisco, 2000, pp. 1023-1029.

[10] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research,* vol. 27, pp. 647-665, Jun 2008.

[11] D. Filliat, "Interactive learning of visual topological navigation," in *International Conference on Robots and Intelligent Systems*, 2008, pp. 248-254.

[12] G. Csurka, *et al.*, "Visual categorization with bags of keypoints " in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1-22.

[13] A. Angeli, *et al.*, "Visual topological SLAM and global localization," *Icra: 2009 Ieee International Conference on Robotics and Automation, Vols 1-7,* pp. 2029-2034, 2009.

[14] A. Pronobis, *et al.*, "Multi-modal Semantic Place Classification," *The International Journal of Robotics Research,* vol. 29, pp. 298-319, 2010.

[15] T. Oskiper, *et al.*, "Multi-modal Sensor Fusion Algorithm for Ubiquitous Infrastructure-free Localization in Vision-impaired Environments," in *International Conference on Intelligent Robots and Systems*, Taipei, 2010.

[16] X. Zhang and A. L. P. Tay, "Fast learning artificial neural network (FLANN) based color image segmentation in R-G-B-S-V cluster space," in *International Joint Conference on Neural Networks*, 2007, pp. 563-568.

[17] H. Bay, *et al.*, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding,* vol. 110, pp. 346-359, 2008.

[18] J. Sivic, *et al.*, "Unsupervised discovery of visual object class hierarchies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 2182-2189.

[19] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524-531.

[20] E. Nowak, *et al.*, "Sampling strategies for bag-of-features image classification," *Computer Vision - Eccv 2006, Pt 4, Proceedings,* vol. 3954, pp. 490-503, 2006.

[21] A. L. P. Tay, *et al.*, "The hierarchical fast learning artificial neural network (HieFLANN) - An autonomous platform for hierarchical neural network construction," *IEEE Transactions on Neural Networks,* vol. 18, pp. 1645-1657, Nov 2007.

[22] X. J. Zhang and A. L. P. Tay, "A binocular vision system with attentive saccade and spatial variant vergence control," *Cybernetics and Systems,* vol. 42, pp. 45-63, 2011.

[23] X. J. Zhang, *et al.*, "Neural classification of objects based on Gabor signature," in *International Joint Conference on Neural Networks*, 2008, pp. 893-900.