Journal of Science and Technology ***49****, Number 5 (2011) 57-67*

TRAFFIC SIGN DETECTION USING LOCAL FEATURES

Khanh Nguyen Duy[1], Duy Le Dinh[2], Duc Duong Anh[2]

[1]*Faculty of Electronic and Informatics, Cao Thang Technical College*

[2]*Faculty of Information Technology, University of Science Ho Chi Minh City*

ABSTRACT

Automatic traffic sign detection and recognition are very important for GPS-based navigation systems; however, it also raises many challenges in research and practice. Our work solves some of these difficulties: First, we have analyzed traffic sign system in real conditions in Vietnam. Besides, we have also proposed high-diversity datasets including 160 types of road signs under real-world conditions; Secondly, on using these datasets, we have done experiments based on local features and "Bag of Words" model (BoW) – which are the state-of-the-art approach in image classification and object class detection. The results are very encouraging to develop this approach in later works. Our experiments also clarify the effect of codebook size in BoW model and the drawbacks of local features.

Keywords. object detection; traffic signs detection; traffic signs dataset; bag of words; local features; SIFT

1. INTRODUCTION

An essential requirement ensuring GPS-based navigation system efficiency is that we have to update constantly the road sign system, especially for areas where road sign system changes frequently, i.e. Vietnam. To accomplish this task efficiently, we need to develop a system which to detects and to recognizes traffic signs automatically.

However, we face many challenges to apply a such system in practice. The most difficult problem is that the influence of outdoor conditions such as: light condition; occlusion caused by trees, vehicles, pedestrians, and other road signs; cluttered background; bad weather conditions i.e. fog, rain, shadows and clouds. Besides, traffic signs may be damaged or faded over long time, or images from camera have motion blur. For this reason, many proposals have been presented [4, 7, 10, 11, 12], but they are unconvincing under the real-world conditions.

On the other hand, there is not a popular traffic sign dataset that is widely used over the research community [14]. Although a few datasets were proposed along with papers, they are collected and used separately. Most of these datasets contain a very few images with low diversity, and consist of a few road signs of only one specific country.

Therefore, the goals of our work are to develop a good dataset of traffic signs sampled in

Vietnam, and to experiment on this dataset. Recently, the algorithm based on local features is one of the most advanced approaches in image classification and object class detection [13], [20]. The experiments on images and video datasets such as VOC PASCAL, Caltech and TRECVID show surprising effectiveness of local features and "bag of words" model (BoW) under challenging real-world conditions [2, 3, 5, 9]. However, how effective they are on traffic signs, especially under outdoor conditions in Vietnam, remains unanswered. Thus, in this paper we focus on local features and BoW method to evaluate their effectiveness.

2. RELATED WORKS

In recent years, traffic sign detection and recognition have attracted much attention as the researches of Piccioli [7], Yuille [4], Fang [10], Barnes [11], and Ruta [12]. Almost of these studies use global features (i.e. edge images) and shape detectors to detect traffic signs, such as the works of Piccioli, Yuille, Barnes, Ruta. In addition, colors and promising sub regions where a target object often appears (such as at the right of images) are also used as additional information to narrow down the search space [7, 4, 12]. Fang [10] also used edge map with color information, but instead of using the shape detectors, neural network and connected components model are used to detect signs. Unfortunately, these approaches are not really effective in practical conditions because of the following causes:

- Firstly, road signs consist of many different shapes such as circles, triangles, rectangles, etc. In fact, implementing an algorithm which can detect all shapes is too difficult and costs many calculations. As well as the confusion with other objects (especially which have quite similar shapes) will increase. So, most studies are limited to a few certain shapes. For example, Piccioli proposes an algorithm using for triangular and circle signs; Yuille and Barnes detect signs that are polygons only (triangles, rectangles, etc). The Figure 1 illustrates some road sign's shapes.

  In addition, this approach faces some challenges:

  - Road signs are distorted and their shapes are changed (shown on Figure 3).
  - Road signs are occluded partly (shown on Figure 4).
  - Road signs are captured sidelong.
  - Cluttered background.

- Secondly, road signs also have many different colors such as: red, blue, yellow, white, green, etc. In fact, the colors may fade day by day. Lighting conditions, weather conditions can also affect the color of collected images (shown on Figure 2). Therefore, the above approaches only use color as additional information to narrow the search space. Nevertheless, the limitation of color scales is required in some studies, as Piccioli's algorithm is able to detect blue and red signs, Yuille only experiments on STOP signs (red), Ruta's proposal can detect signs with three colors: red, yellow, green. Although this approach works effectively on a number of signs, it's hard to provide good performance with various color scales of the actual road signs.

- In addition, we believe that using shape detectors will miss some context information which is useful for road signs detecting, for example, the ability to detect road signs in images is great when there are piles which are used to attach road signs.

*Figure 4.*Traffic signs are occluded under different situations

Figure 1. Traffic signs have different shapes

Figure 2. Faded traffic signs

Figure 3. Damaged traffic signs

Besides, recently many researchers have successfully employed a new object detection and recognition approach which inspiring from the state-of-the-art image classification method [13]. The main idea of this approach includes two proposals: first, objects are represented in the form of sparse parts; second, detection and recognition are based on classifying by machine learning algorithms (AdaBoost, Neural network, pLSA, SVM). These ideas were originally suggested by C. Papageorgiou [15] and A. Mohan [16], and successfully experimented on face detection and pedestrian detection. Outstanding contribution of this approach is its effectiveness for occlusion, cluttered background, and damaged objects. It inspires for a lot of other further researches [17], [18, 19]. Remarkable contributions are studies of G. Csurka and J. Sivic [19, 20]. Inspired from [15, 16] and a model in statistical natural language processing, they propose "Bag of Words" (BoW) method for image classification and object detection. On the other hand, some other researchers also focus on methods to detect and describe local features such as D. G. Lowe [1], Mikolajczyk [6, 21]. Zhang experiments BoW and local features for classification of texture and object images in several datasets including PASCAL-VOC and Caltech [9]. The results show surprising effectiveness of local features and BoW under challenging real-world conditions, including substantial cluttered background. Other works such as Lazebnik [2], Van Gemert [3], Van de Sande [5], and Varma [8] continue to improve approaches based on BoW and prove their real effectiveness. The effectiveness of this approach, especially under real-world conditions inspires for our traffic signs detection approach.

3. OUR DATASET AND APPROACH

3.1. Vietnam traffic sign system

Traffic sign system in Vietnam consists of 241 signs divided into 5 groups in priority order: regulatory signs (51), warning signs (64), indicator signs (18), guide signs (88), supporting signs (20) (shown on Figure 5).

Figure 5. 5 groups of Vietnamese traffic signs

Figure 6. Example of inconsistent traffic signs

Figure 7. Example of combination form of traffic signs

However, we can also see some inconsistent signs, or the combination form of different road signs. Thus, traffic sign system is more complex, in practice (shown on Figure 6 and 7).

3.2. Our datasets

Still image dataset. The dataset includes training set and testing set separately from each other. Each set consists of 600 images (300 positive images and 300 negative images). Each image is resized to 640 pixels x 480 pixels. The images are collected in Ho Chi Minh City and Tuy Hoa City (Vietnam), and are taken at different times of day (morning, noon, afternoon, evening). Images are captured from distance 3 meters – 50 meters with various angles (frontal, skew), high diversity backgrounds (highway, alley, crowded roads...). They are also occluded and shaded.

Our dataset includes 160 types of road signs (including combination forms).

Video dataset. The dataset includes 93 video files. Each file has 20 seconds to 3 minutes in length. Video resolution is 640 pixels x 480 pixels. The videos are recorded from motors in several districts of the city, speed from 20 km/h - 40 km/h. Each road is recorded at different times: morning, noon, afternoon, evening. The videos capture many different situations (few vehicles on roads, many vehicles on roads, crowded roads...).

3.4. Our approach

Local features extraction. The feature extraction is done through two operations: a keypoint detector is used to select patches of images, after that patches are described by local descriptors. Keypoint detector's goal is to find rich information areas which determine the presence of objects in images. In ideal case, they are parts of objects. To find the parts of an object effectively, we need to select the appropriate keypoint detector because each keypoint

detector will select different regions of the image according to different criteria such as: image patches that contain corners (Harris), blobs (Hessian), etc [6]. Therefore, different classes of objects are suitable with different keypoint detectors. In object detection, it is very undesirable that we get numerous unrelated image patches that make noise, confusion and slow down the algorithm. In other words, a keypoint detector will be effective if it discovers many image patches of the target object parts and limits the other unrelated patches.

In this study, we experiment and evaluate the effectiveness of some state-of-the-art keypoint detectors with traffic signs. The keypoint detectors are: Harris - Laplace, Hessian – Laplace [21], Difference of Gaussian (DoG) [1], and dense sampling [22].

Figure 8. Example of images in dataset at different distances.

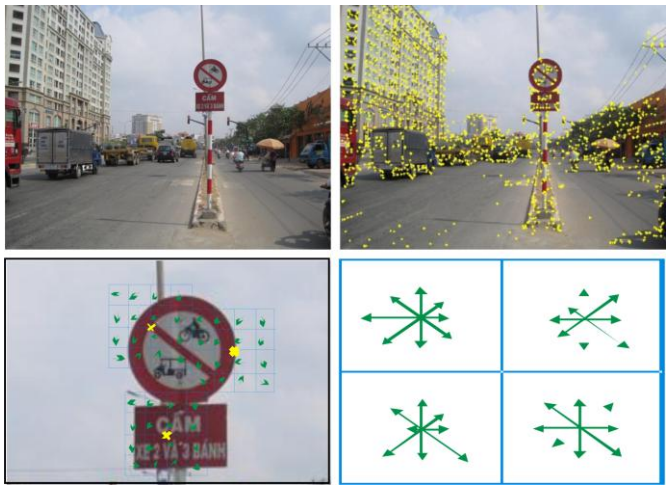

*Figure 9.*Illustration of local features extraction

After selecting images patches, the next task is to describe these patches by local descriptors. Local descriptors are used in order to increase discrimination between different class of patches while decrease discrimination between patches of the same class i.e. road signs.

Effects of local descriptors has been confirmed by many studies [1, 6, 17, 18, 19], in which the SIFT has been proved to archive the best performance [1, 6]. Besides, color SIFT [5] is appropriate for traffic signs because colors are also an important information of signs.

Creating Wordbook. Input of this stage is the local descriptors which are extracted from the train set in the previous step. The algorithm clusters these descriptors to form a "wordbook" or "codebook" that is used to represent images. This idea is originally from a natural language processing method, in which a sentence would be represented as a set of words. Similarly, clustering local descriptors is in order to create a set of "words" which correspond to parts of objects (traffic signs). Then objects will be represented as a collection of the "words" certainly.

Besides, the local descriptors are extracted from various patches of image in the training set, of which many patches do not relate to target objects. Because they are also clustered, they create useless "words" and make noise. The problem of clustering is to construct a set of "words" which can better corresponds to different parts of objects, and reduce useless "words". To solve this problem well, apart from a good clustering algorithm, we must choose a reasonable number of clusters. If the number of cluster is too small, many different patches may be clustered into a group. Obviously it cannot form a good "vocabulary". However, if the number of cluster is large, many "words" which are really the same are created, and noise caused by unrelated descriptors will increase (in traffic sign detection problem, unrelated descriptors dominate because traffic signs occupy only a small portion of images).

In this study, we use kmeans++ algorithm to cluster the local descriptors which have been extracted from the training set into N clusters (N is the parameter of the system). Distances between descriptors are L2.

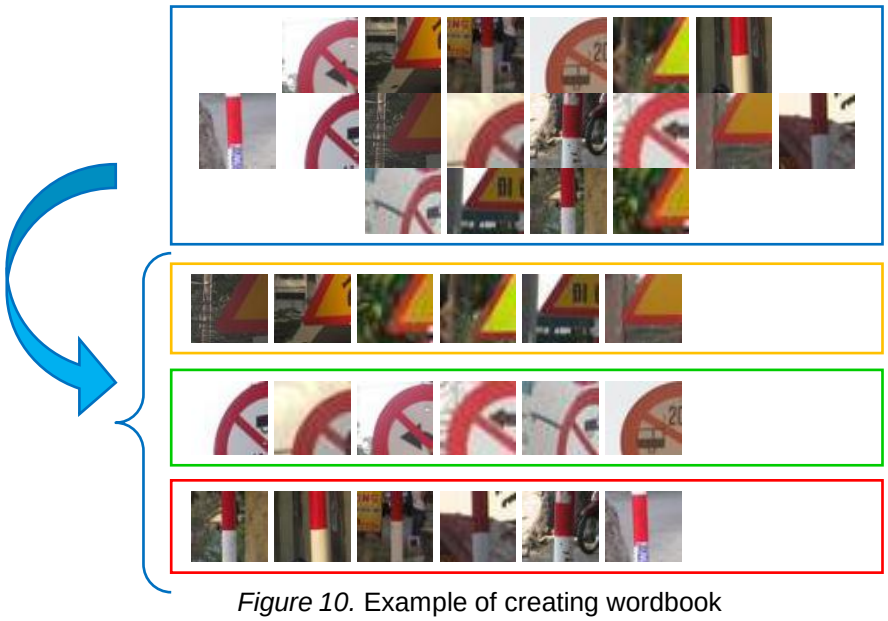

Figure 10. Example of creating wordbook

Image Representation. The main idea of representing images is reconstructing images from image patches which are extracted and clustered from the training set. More specifically, images are represented as histogram vectors of words appearing. Distances between local descriptors and words in wordbook are computed based on L2 distance, then descriptors are

assigned to one word which is the nearest.

Furthermore, to add spatial information of words, S. Lazebnik [2] uses spatial pyramid with the idea of dividing images into a grid (2×2, 3×3, etc.), and then computes histograms of words for each cell. But this method is not effective for traffic sign detection because signs are only occupy a small portion of images. Consequently, all parts of signs fall into the same cell in the grid. Moreover, the signs may be located in many different positions in images, so information about the spatial location of signs is not necessary.

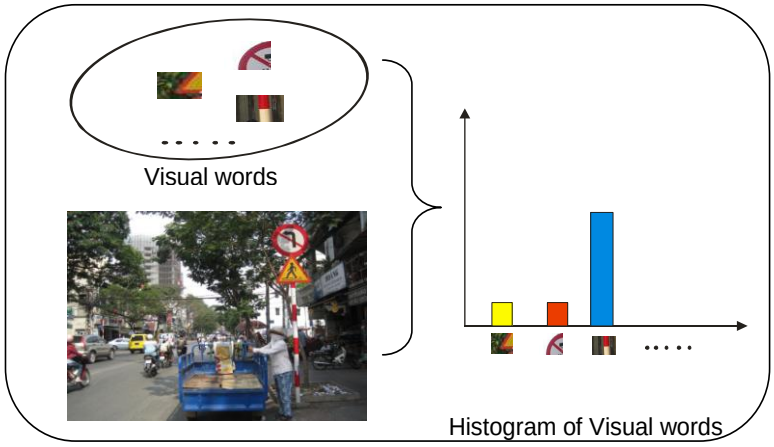

Figure 11. Illustration of Histogram of Visual words

Classification. The classification stage determines whether if traffic signs are detected or not. Train set is divided into two subsets: positive (containing signs) and negative (containing no sign). After that, they are learned by SVM classifier. In addition, we use grid search tool to optimize the remaining parameters.

4. EXPERIMENTS AND RESULTS

4.1. The effect of codebook size

Experiment 1. We experiment on the still images dataset presented in Section 3. The number of descriptors used for clustering is 100,000 (randomly selected from over 1,000,000 descriptors extracted). Intensity based SIFT is used for descripting local features. Keypoint detectors include: DoG, Harris-Laplace (HL), Dense sampling (DS), and Hessian-Laplace-Affine (HS). For dense sampling detector, keypoints are taken on each 6 pixels, scale factor is 1.2. SVM classifier is used with RBF kernel and grid search tool. Configuration parameters for grid search are: logC = [0, 6] step 2, logG = [-12, 4] step 2.

Results in Table 1 show how codebook size affects the performance. For most of the detectors, performance improves persistently as codebook size increases from 10 to 2000 (1000 for DoG), and turns downward after that. It can be explained that images are represented better when the number of "words" increases, but after archiving the peak, images cannot be represented better any more. In addition, noises begin to dominate and reduce performance. Results also show that the best range of number of "words" is between 1000 and 3000 for our traffic signs dataset.

On the other hand, the effectiveness of keypoint detectors can be compared here (as shown on Figure 12). Harris – Laplace outperforms the other detectors because traffic signs have many corners which can be detected efficiently. As well as dense sampling detector also archives very good performance in spite of their simple approach. An important advantage of dense sampling is that it collects image patches at a regular grid, so it is more stable than other detector on the difficult data.

Table 1. Detector and codebook size evaluation on images dataset using Average precision

Keypoint detector	Codebook size								
	10	20	50	100	300	500	1000	2000	3000
HL	78.87	80.35	81.61	83.56	86.87	85.70	85.91	**87.03**	86.85
HSL	62.80	68.45	74.06	74.95	79.60	79.23	79.12	**82.43**	81.34
DoG	63.84	65.53	67.96	69.96	71.33	72.90	**73.23**	70.63	72.55
DS	71.89	72.78	76.81	80.15	82.89	84.53	84.89	**85.15**	84.63
MK	78.65	79.54	82.22	83.63	86.67	87.66	88.94	**89.38**	89.3

Table 2. Detector and codebook size evaluation on videos dataset using Average precision

Keypoint detector	Codebook size								
	100	300	500	1000	2000	3000	4000	5000	6000
HL	47.78	48.44	48.54	48.84	50.12	50.86	51.34	**52.17**	50.72
HSL	45.66	46.29	46.86	45.46	46.46	46.88	46.5	46.4	**47.22**
DoG	45.56	45.94	45.12	45.20	45.61	46.23	45.69	45.41	**47.04**
DS	48.50	47.62	47.86	48.43	48.65	48.30	48.83	**49.27**	48.81

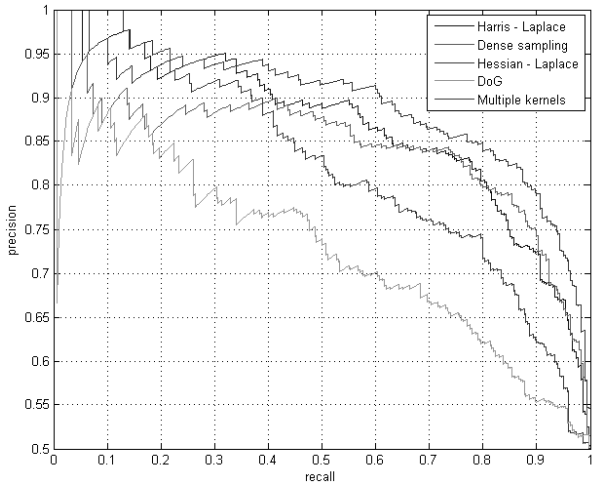

Figure 12. recall/precision curves for the best ap in table 1

Experiment 2. Similarly, the experiment is done on the video dataset with parameters as in Experiment 1. Videos are divided into 1 second segments and labeled positive/negative corresponding with the presence/absence of traffic signs. Training and testing processes are based on the single key frame which is selected as a middle frame of segments. Dataset includes 3201 negative segments and 2592 positive segments. 500 segments of each are used to training, the rest for testing.

The results confirm that Harris-Laplace and dense sampling detector are better than the remaining two detectors. The results also show that with video dataset, the codebook size must be at 5000 - 6000 for best performance. This can be explained that the diversity in video dataset is higher than still image dataset; therefore, we need more words to represent images. On the other hand, by comparing with the results obtained in experiment 1, the difficulty and diversity of this dataset are exposed strongly. This dataset will be a challenge for further researches.

4.1. Combine multiple kernels

Experiment 3. J. Zhang [9] proposes using simultaneously of multiple detectors and descriptors to increase the efficiency of image representation method. Each {detector, descriptor} forms a channel. Our experiment is conducted with the Harris-Laplace and dense sampling detectors. Local descriptors are used: SIFT and colorSIFT including rgsift, transformedcolorsift, and opponentsift [5].

The distance is used to calculate the difference between two channels:

$$D(S_1, S_2) = \frac{1}{2}\sum_{i=1}^{m}\frac{(u_i - w_i)^2}{u_i + w_i} \quad (1)$$

where:

$$S_1 = (u_1, u_2, ..., u_m) \text{ and } S_2 = (w_1, w_2, ..., w_m)$$

are histograms of words of the two channels.

The difference between two images is based on distances of channels:

$$K(S_i, S_j) = \exp(-\sum_{c \in C}\frac{1}{A_c} D_c(S_i, S_j)) \quad (2)$$

where C is the set of channels combined, A_c is the weight corresponding to each channel.

SVM is used with pre-computed kernel calculated from distances as above.

Result on dataset shows (Figure 12) that the combination method is always better than any single channel. The best performance is 89.38% (corresponding to codebook size 2000). Thus, we can confirm that the combination of multiple detectors and descriptors improves the algorithm efficiency. However, it can be seen that efficiency is not too significantly increasing.

5. CONCLUSION

In this paper, we have built a high diversity dataset of traffic signs of Vietnam, which is the important support for further researches. Besides, we have proposed a method using SIFT and

BoW for traffic sign detection, and experimented on this method using our dataset. Results show good performance of SIFT for this task, as well as Harris-Laplace and dense sampling detectors. Besides the archived advantages: have not to set constraints on the shape and color of road signs as some studies shown; works well with cluttered backgrounds, occlusion, damaged signs, our approach encounters some difficulties such as: still being relatively sensitive to light especially low light, low contrast. Similarly, the image size changes largely (from the train set), or road sign images are captured at too skew angle are also issues to overcome in the future.

Furthermore, according to the experimental results, we conclude that the best size of the codebook depends heavily on the diversity of datasets. It means that datasets having high variability require larger codebook sizes. In addition, our results have also confirmed that the algorithm using a combination of many channels would provide a better representation than another using a single channel.

REFERENCES

1. D. G. Lowe - Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision **60** (2004) 91-110.
2. S. Lazebnik, C. Schmid, J. Ponce - Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2** (2006) 2169-2178.
3. J. C. van Gemert, JM. Geusebroek, C. J. Veenman, A. W.M. Smeulders - Kernel codebooks for scene categorization, European Conference on Computer Vision **3** (2008) 696-709.
4. A. L. Yuille, D. Snow, and M. Nitzberg - Signfinder: Using Color to detect, localize and identify informational signs, Proceedings of the IEEE International Conference on Computer Vision, 1998.
5. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek - Evaluation of color descriptors for object and scene recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.
6. K. Mikolajczyk, C. Schmid - A performance evaluation of local descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1615-1630.
7. G. Piccioli, E. D. Micheli, M. Campani - A robust method for road sign detection and recognition, Proceedings of Third European Conference on Computer Vision **1** (1994) 495-500.
8. M. Varma and D. Ray - Learning the discriminative power-invariance trade-off, IEEE 11th International Conference on In Computer Vision, 2007, pp. 1-8.
9. J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid - Local features and kernels for classification of texture and object categories: A comprehensive study, International Journal of Computer Vision **73** (2) (2007) 213–238.
10. C. Y. Fang, C. S. Fuh, S. W. Chen, P. S. Yen - A Road Sign Recognition System Based on Dynamic Visual Model, IEEE Computer Society Conference on Computer Vision and Pattern Recognition **1** (2003) 750.

11. N. Barnes, G. Loy, D. Shaw, A. R. Kelly - Regular Polygon Detection, Proceedings of 10th IEEE International Conference on Computer Vision, 2005, pp. 778-785.
12. Ruta, Y. Li, and X. Liu - Towards real-time traffic sign recognition by class-specific discriminative features, British Machine Vision Conference, 2007.
13. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman - Multiple Kernels for Object Detection, Proceedings of IEEE 12th International Conference on Computer Vision, 2009, pp. 606-613.
14. Tam T. Le, Son T. Tran, Seichii Mita, Thuc D. Nguyen - Realtime Traffic Sign Detection Using Color and Shape-Based Features, The 2nd Asian Conference on Intelligent Information and Database Systems (ACIIDS), 2010.
15. Papageorgiou and T. Poggio - A Trainable Pedestrian Detection system, International Journal of Computer Vision (IJCV) 2000, pp. 15-33.
16. Mohan, C. Papageorgiou, and T. Poggio - Example-based object detection in images by components, PAMI, 2001, pp. 249-261.
17. S. Agarwal, A. Awan, D. Roth - Learning to detect objects in images via a sparse, part-based representation. IEEE TPAMI **26** (11) (2004) 1475-1490.
18. S. Agarwal and D. Roth - Learning a Sparse Representation for Object Detection, Proc. of the European Conference on Computer Vision, 2002, pp. 113-128.
19. J. Sivic, B. Russell, A. Efros, A. Zisserman, B. Freeman - Discovering Objects and their Location in Images, ICCV, 2005.
20. G. Csurka, C. Bray, C. Dance, and L. Fan - Visual categorization with bags of keypoints, In Workshop on Statistical Learning in Computer Vision, ECCV, 2004, pp. 1-22.
21. K. Mikolajczyk and C. Schmid - Scale and affine invariant interest point detectors, International Journal of Computer Vision **1** (60) (2004) 63-86.
22. F. Jurie and B. Triggs - Creating Efficient Codebooks for Visual Recognition, International Conference on Computer Vision, 2005.

Corresponding author:

Khanh Nguyen Duy,

Faculty of Electronic and Informatics, Cao Thang Technical College

Email: *khanhduy@gmail.com*