# Stereo Image Based Object Localization Framework for Visually Impaired People Using Edge Orientation Histogram and Co-occurrence Matrices

Supakit Fuangkaew, Karn Patanukhom
Visual Intelligence and Pattern Understanding Laboratory
Department of Computer Engineering
Chiang Mai University, Chiang Mai, Thailand
karn@eng.cmu.ac.th

***Abstract*—A new framework that uses internet-based images for detecting objects and estimating real world location of the objects via stereo images is proposed. This framework provides a self-learning ability for detecting desired objects in the scene without pre-prepared classifiers by harvesting sample images of the objects from the internet. Histogram and co-occurrence matrices of edge orientation are used as features. The objects are recognized based on likelihood scores and distance in the feature space between every window in the scene and k-nearest prototypes. A local feature matching is used to match the feature points in stereo pair. Disparities from stereo images are used to estimate real world distance and direction of the objects. The experiments on 120 pairs of stereo images from three object classes show the satisfying results in comparison to baseline methods.**

***Keywords; visually impaired; assistive device; object detection; stereo images; co-occurrence matrix***

## I. INTRODUCTION

One of the most important problems of visually impaired people is living or staying in unfamiliar places. Currently, there are many assistive devices that have been developed to help the daily life of visually impaired people such as navigation system, voice guidance system or obstacle detection where the object detection is one of the common tasks in the assistive tool development.

The navigation system for the visually impaired people is developed in a NAVIG project [1]. The NAVIG system provides both far-field navigation such as pedestrian navigation and near-field guidance for object identification and grasping guidance. A head mounted stereo camera is used to capture stereo images of surroundings. An object distance is computed by using stereoscopic disparity and the calibration matrix of the lenses. A Spike Net Framework [2] is applied in the NAVIG system for object recognition and detection. The prototype of NAVIG system has been tested on classifying European currency bills. 100% of the bills were able to identify while the average response time was slightly above ten seconds per bill.

Recently, there are many researches developed to detect object in the scene image. Laptev [3] developed the object detection framework by combining AdaBoost with local histogram features. The boosting framework was adapted to vector-valued histogram features and a Weighted Fischer Linear Discriminant (WFLD) is applied to improve the weak learners. Evaluation on VOC05 Challenge of four object classes shows the average precisions of 0.896 for motorbike detection, 0.370 for bicycle detection, 0.250 for people detection, and 0.663 for car detection.

Felzenszwalb, et. al. [4] developed the object detection system using mixtures of multi-scale deformable part models. A Latent Support Vector Machine (LSVM) is developed to train models using partially labeled data. Large training dataset is important to obtain high performance discriminative classifier. To solve problem of unbalance dataset, a methodology of data-mining for hard negative examples was applied. The system was tested on VOC06, VOC07 and VOC08 Challenges including 20 object classes. The results showed both efficient and accurate of the method.

In this paper, the self-learning framework using internet based images and new approaches for object detection and localization are developed for the navigation system of visually impaired people. The objective of this work is to develop the object detection method that can perform under a limited size of the internet based training images and requires a reasonable time in training and detection processes. The proposed object detector can be trained by using a small number of positive samples (object images) and without negative sample (the other images). Histogram and co-occurrence matrices of edge orientation are used as features for similarity measuring. A new scoring method is proposed to determine object location in the scene image. In the real world location estimation process, a local feature method is used to match points between stereo images. The stereoscopic disparity between two images are extracted by calculating the different in positions of the local feature points and used for estimating the distance between the camera and the target object. The rest of the paper as follows. In Sections II, details of proposed framework are described. Then, Section III provides the experimental results of object localization. Finally, the conclusion of the paper is in Section IV.

INISCom 2015, March 02-04, Tokyo, Japan

DOI 10.4108/icst.iniscom.2015.258352

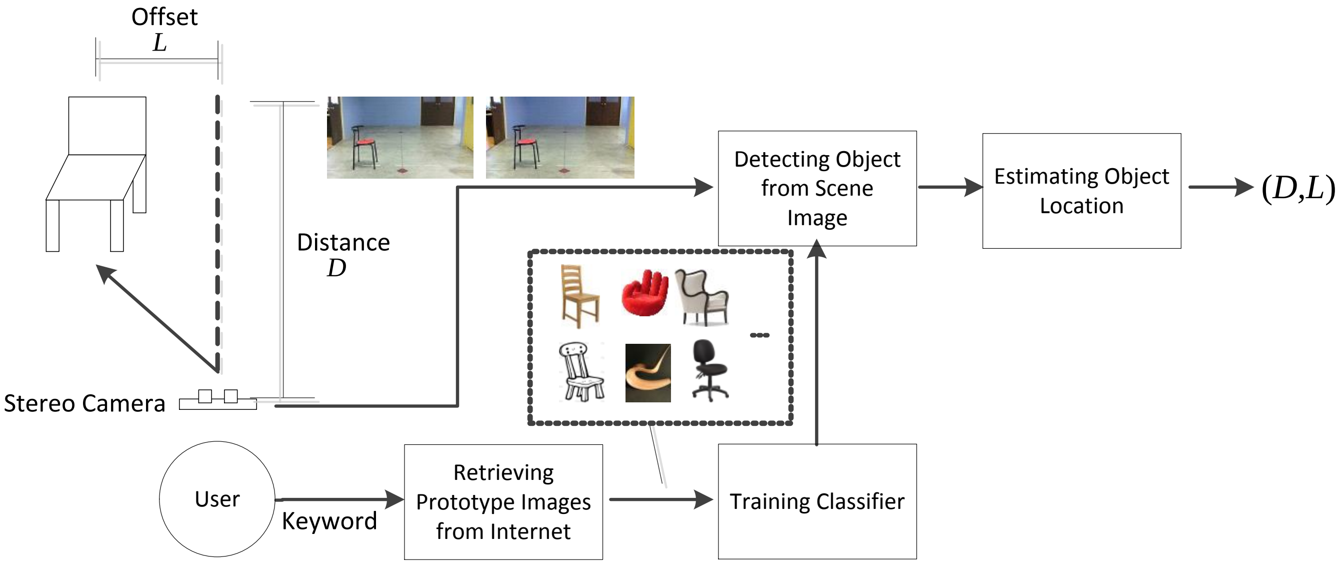

Offset
L
Distance
D
Stereo Camera
Detecting Object from Scene Image
Estimating Object Location
(D,L)
User
Keyword
Retrieving Prototype Images from Internet
Training Classifier

Figure 1. Overview of the proposed framework.

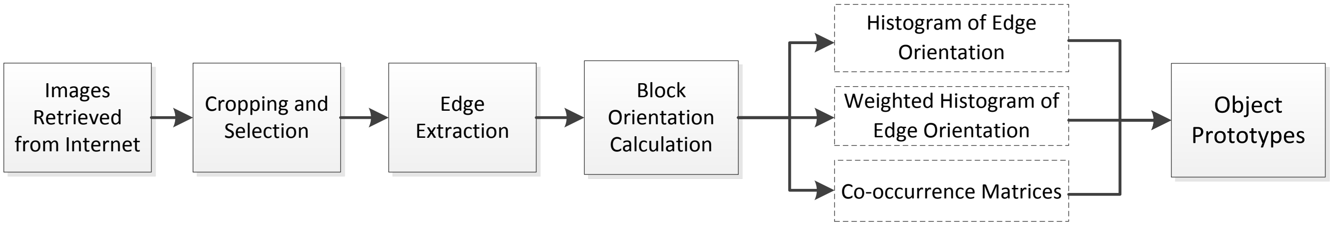

Images Retrieved from Internet
Cropping and Selection
Edge Extraction
Block Orientation Calculation
Histogram of Edge Orientation
Weighted Histogram of Edge Orientation
Co-occurrence Matrices
Object Prototypes

Figure 2. Feature extraction scheme of the prototype images

## II. The Proposed Framework

### A. System Overview

In this work, the assistive system for visually impaired people whose cannot see normally is developed. The objective is to find the real world location (distance and direction) of the requested object by using stereo images as the inputs. In the typical object recognition or detection system, the classifiers must be prepared in advance for each specific target object. In addition, to obtain the better result, the pre-prepared classifiers must be trained by using a large number of training samples and require long time consumption. However, in some situations, the requested object can be widely varied depending on user and environment. The requested objects are sometimes not general and may be unique for the specific users. Therefore, it is impossible to prepare the classifiers to cover every class of object.

To solve this problem, we develop the framework for object localization system that allows the system to automate the learning process via the internet-based images. The proposed framework provides an ability to detect the requested target objects; even though there are no pre-prepared classifiers for those object classes. For example, if the user wants to find location of a chair in the front of him, he has to point the camera to that direction and tells the system that he wants to find a chair via speech recognition system. In this example, the input requested keyword is “chair”. If the system has no pre-trained chair detector, then the prototype images of chair are automatically collected from internet via search engines and are used to train a chair detector. Then, the detector is used to detect chairs in the input scene images captured from the stereo camera. After the chairs are detected in the stereo images, the disparities of the pixels between two images are extracted from the object area and used for estimate a direction and distance between camera and the chair in real world. Finally, the system can return the distance and direction of the chair via speech synthesis. The proposed framework is expected to provide the better performance and require less training time than other conventional object detection approach for this limited situation.

An overview of the proposed framework is illustrated in Fig.1. This framework consists of three main processes that are training process, object detection process and object location estimation process. In the proposed framework, only positive images (images of target object) called as prototypes are retrieved from the internet by using search engines such as Google. Similar to the other object detection scheme, to detect the objects in the target scene image, sliding windows in multi-scale are applied to scan for the objects in the entire image. The feature distances between each window and all prototype images are computed. Histogram and co-occurrence matrices of edge orientation are used as features in this step. The detection result can be obtained by analyzing the distances to the prototypes and other statistical parameters extracted from prototypes. After the object can be detected in the scene image, the location of the object in real world is estimated by using the disparity of the pixels in the stereo images. The proposed framework can perform without iterative training process and with limited number of prototype images that can be retrieved from the internet and all processes can be performed within reasonable time.

### *B. Edge Orientation Histogram (EOH)*

Edge orientation histogram [6] is one of the shape descriptor which shows the distribution of directions of the edge image. In this work, to compute the EOH, firstly, edge image is extracted from the target image. The gradient directions are computed for every edges pixel. The edge image is divided into $n \times n$ blocks. The dominant direction of the edge pixels in each block is quantized into $N_B$ orientation bins. The dominant direction is defined as a mode of edge orientation of every pixel in that block which is weighted by the gradient magnitude. $\theta(i, j)$ denotes the dominant orientation bin in the block $(i, j)$ where $\theta(i, j) \in [0, N_B]$. Note that $\theta(i, j) = 0$ for the empty blocks with no edge pixel inside.

The edge orientation histograms used in this framework are divided into two types.

*1) Unweighted Edge Orientation Histogram (UEOH):* Unweighted edge orientation histogram $h_U(x)$ is defined as

$$h_U(x) = \frac{\#\text{Blocks with}\,\theta(i, j) = x}{n^2}. \quad (1)$$

where “$\#$" represents as “the number of”.

*2) Weighted Edge Orientation Histogram (WEOH):* In this work, we proposed a novel weight calcuation for collecting the number of occurrences in EOH. To compute WEOH, each block $(i, j)$ is weigthed by two parameters as follows.

*a) Non-empty Probability:* The system create the weight for every block $(i, j)$ by considering a probability that block $(i, j)$ is empty in the prototype images. The non-empty probability weigth $w_{NEP}(i, j)$ is defined as

$$w_{NEP}(i, j) = \frac{\#\text{Prototypes with}\,\theta(i, j) \neq 0}{N_P}. \quad (2)$$

where $N_P$ represents the number of prototypes. Since the blocks in the target image that are frequently empty in the prototype images tend to belong to the background area rather than the object area, the blocks that are frequently empty in the prototype images will provide less confidence (weight) for counting the occurrences in WEOH. The occurrences of the blocks that are always empty in the prototype images will not be counted in WEOH.

*b) Uniqueness:* The second parameter to determine the weights is uniqueness measure. The uniqueness weight $w_{UNQ}(i, j)$ is defined as

$$w_{UNQ}(i, j) = \frac{\max\limits_{1\le k\le N_B}(N(i, j, k)) - \min\limits_{1\le k\le N_B}(N(i, j, k))}{N_P}. \quad (3)$$

where $N(i, j, k)$ is the number of prototypes that have $\theta(i, j) = k$. According to (3), the uniqueness weight for the block $(i, j)$ is determined from the difference between the numbers of occurrences from prototypes of the most frequent orientation bin and the least frequent orientation bin in the block $(i, j)$. The uniqueness weight $w_{UNQ}(i, j)$ is maximized as $w_{UNQ}(i, j) = 1$ when the orientations of all prototypes images in that block are located in the same bin and it is minimized as $w_{UNQ}(i, j) = 0$ when the orientations of all prototypes in that block distribute uniformly. The blocks with higher uniqueness have more confidence for counting the occurrences in WEOH since the feature is not much varied among the prototypes.

By applying both weights, WEOH can be computed from

$$h_W(x) = \sum_{\theta(i, j)=x} w_{BG}(i, j) \times w_U(i, j). \quad (4)$$

where $h_W(x)$ represents weighted edge orientation in $x$-bin.

### *C. Edge Orientation Co-occurrence Matrices (EOCM)*

The co-occurrence matrix is a two dimensional statistics that show a distribution of co-occurring values. The co-occurrence matrices such as a Gray level Co-occurrence Matrix (GLCM) [7] are widely used as the feature extraction in the object recognition problem. In this work, we proposed the co-occurrence matrices of edge orientation as the features for recognizing the object. The Edge Orientation Co-occurrence Matrices (EOCMs) are extracted from the block orientation matrix $\theta(i, j)$ by considering the orientation bins of every pairs of adjacent blocks along two directions (horizontal and vertical) as

$$C_H(p, q) = \frac{\#\text{Blocks that}\,\theta(i, j) = p\,\text{and}\,\theta(i, j+1) = q}{n^2}, \quad (5)$$

$$C_V(p, q) = \frac{\#\text{Blocks that}\,\theta(i, j) = p\,\text{and}\,\theta(i+1, j) = q}{n^2}, \quad (6)$$

where $p, q \in [1, N_B]$. $C_H$ and $C_V$ represent horizontal and vertical EOCMs, respectively.

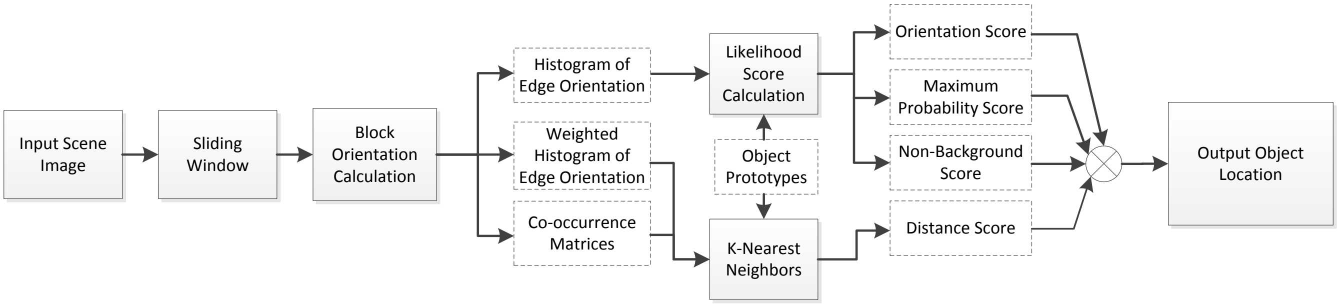


Figure 3. Object detection scheme from the scene image.

### D. Training Process

The training process starts when the user inputs the keyword for the object that he wants to find its location. The system will retrieve the prototype images of the target object from the internet via the search engines. However, a precision of an automatically harvesting process for a particular image class from the internet is still limited [8]. Some filtering techniques [8] can be applied in this step to remove the false images. In this paper, we limited the number of prototype images to improve the precision of harvesting process and reduce the distance calculation cost in detection process. In order to remove the background area so that the prototype images are fitting to the target object, the images are pre-processed by cropping based on the edge projection. Since, in this work, we focus on locating the object in upright position, an aspect ratio can be used to filter out the outliers. The UEOHs, WEOHs and EOCMs are extracted for every prototype images. $h_U^{(p)}, h_W^{(p)}, C_H^{(p)}, C_V^{(p)}$ denote UEOH, WEOH, Horizontal EOCM and Vertical EOCM of the $p$ -th prototype image, respectively. Fig. 2 shows a summary of the features extraction process from internet based prototype images.

In addition, the system also computes the following statistical properties that will be used in the detection process.

- $\mu_M$ : Mean of maximum UEOH.

$$\mu_M = \frac{1}{N_P}\sum_{p=1}^{N_P}\max_{x\in[1,N_B]}(h_U^{(p)}(x)), \tag{7}$$

- $\sigma_M$ : Standard deviation of maximum UEOH.

$$\sigma_M = \sqrt{\frac{1}{N_P}\sum_{p=1}^{N_P}\left[\max_{x\in[1,N_B]}(h_U^{(p)}(x)) - \mu_M\right]^2}\ . \tag{8}$$

- $\mu_E$ : Mean of empty block UEOH.

$$\mu_E = \frac{1}{N_P}\sum_{p=1}^{N_P} h_U^{(p)}(0)\ . \tag{9}$$

- $\sigma_E$ : Standard deviation of empty block UEOH.

$$\sigma_E = \sqrt{\frac{1}{N_P}\sum_{p=1}^{N_P}\left[h_U^{(p)}(0) - \mu_E\right]^2}\ . \tag{10}$$

- $\mu_N$ : Mean of UEOH for non-maximum bins.

$$\mu_N = \frac{1}{N_P}\sum_{p=1}^{N_P} g^{(p)}\text{ , where} \tag{11}$$

$$g^{(p)} = \left(\sum_{x=1}^{N_B} h_U^{(p)}(x)\right) - \max_{x\in[1,N_B]}(h_U^{(p)}(x))\ . \tag{12}$$

- $\sigma_N$ : Standard deviation of UEOH in non-maximum bins.

$$\sigma_N = \sqrt{\frac{1}{N_P}\sum_{p=1}^{N_P}\left[g^{(p)} - \mu_N\right]^2}\ . \tag{13}$$

- $P_D(i)$ : Distribution of dominant directions in the entire set of prototype images.

$$P_D(i) = \frac{\#\text{Prototypes with } \arg\max(h_U^{(p)}(x)) = i}{N_p}\ . \tag{14}$$

### E. Object Detection

To detect the target object, only one image from the stereo pair is used here. The scheme of object detection process is shown in Fig. 3. The sliding windows in the multi-scale image pyramid are applied to scan the entire scene image. Each window is segmented into $n \times n$ blocks as in the prototype images. The UEOHs, WEOHs and EOCMs are extracted for every window. Then, the system calculates two types of score called distance score and likelihood score.

The distance score denoted by $D$ is a parameter that measures the similarity of the target window to the object prototypes in the feature space. The distance score is composed of two components as follows.

- EOH Distance Score ( $D_H$ ) represents a distance score of each window that is obtained by measuring the difference between WEOH of the target window and the prototypes. $D_H$ is extracted by finding the $K$ nearest prototypes of the target window in WEOH feature space. The distance function is defined by using Euclidean distance as $\left\| h_w - h_w^{(p)} \right\|$ where $h_w$ is WEOH of the target window and $h_w^{(p)}$ is WEOH of the prototype. $D_H$ is calculated from the summation of WEOH distance between the target window and the $K$ nearest prototypes as

$$D_H = \frac{1}{1 + \sum_{p=1}^{K} \left\| h_w - h_w^{(p)} \right\|} . \qquad (15)$$

- EOCM Distance Score ( $D_C$ ) represents a distance score of each window that is obtained by measuring the difference between EOCM of the target window and the prototypes. Similar to $D_H$ , the $K$ nearest prototypes of the target window in EOCM space are firstly determined. The distance function is defined as $\left\| c - c^{(p)} \right\|$ where $c$ is a vector representation of the values in $C_H, C_V$ that are extracted from the target window and $c^{(p)}$ is a vector representation of $C_H, C_V$ extracted from the prototype images. Note that the dimension of $c$ becomes $2N_B^2$. $D_C$ can be defined based on the summation of EOCM distance between the target window and the $K$ nearest prototypes as

$$D_C = \frac{1}{1 + \sum_{p=1}^{K} \left\| c - c_w^{(p)} \right\|} \qquad (16)$$

Then, the final distance score $D$ can be obtained by adding two components together as

$$D = D_H + D_C \qquad (17)$$

On the other hand, the likelihood score $P$ is a parameter that measures the likelihood probability $P(x | \text{object class})$ of the given observation parameters $x$ that are extracted from every window in the scene image. The observation parameters used in the system are maximum UEOH, empty block UEOH and UEOH for non-maximum bins. The distribution of each parameter is estimated from the prototype images in training process as mentioned in Section II-C. The likelihood score can be separated in to four terms as

$$P = P_M \cdot P_E \cdot P_N \cdot P_D . \qquad (18)$$

Each term has the definition as follows.

- Maximum EOH Likelihood $P_M$ is defined as

$$P_M(x) = \frac{1}{\sqrt{2\pi\sigma_M^2}} e^{-\frac{(x-\mu_M)^2}{2\sigma_M^2}} , \qquad (19)$$

  where $x = \max_{i \in [1, N_B]} (h_U(i))$ which is a maximum value of UEOH extracted from the corresponding window.

- Empty Block EOH Likelihood $P_E$ is defined as

$$P_E(x) = \frac{1}{\sqrt{2\pi\sigma_E^2}} e^{-\frac{(x-\mu_E)^2}{2\sigma_E^2}} , \qquad (20)$$

  where $x = h_U(0)$ which is a value of empty block UEOH counted from the corresponding window.

- Non-maximum Bin EOH Likelihood $P_N$ is defined as

$$P_N(x) = \frac{1}{\sqrt{2\pi\sigma_N^2}} e^{-\frac{(x-\mu_N)^2}{2\sigma_N^2}} , \qquad (21)$$

  where $x = \left( \sum_{i=1}^{N_B} h_U(i) \right) - \max_{i \in [1, N_B]} (h_U(i))$ which is a summation of UEOH values in non-maximum bins.

- Dominant Direction Likelihood $P_D(x)$ can be obtained by using the distribution in (14) which is trained from the prototype images where $x$ is an index of the orientation bin that provides the maximum UEOH in the corresponding window.

In the final step, after distance score $D$ and likelihood score $P$ have been determined for every window in every image scale, the final score map can be extracted by

$$S(i, j) = \max_{\alpha} \left( D(i, j, \alpha) \cdot P(i, j, \alpha) \right) \qquad (22)$$

(a) (b) (c)

Figure 4. Examples of prototype images retrieved from Google, (a) bicycle, (b) chair and (c) standfan.

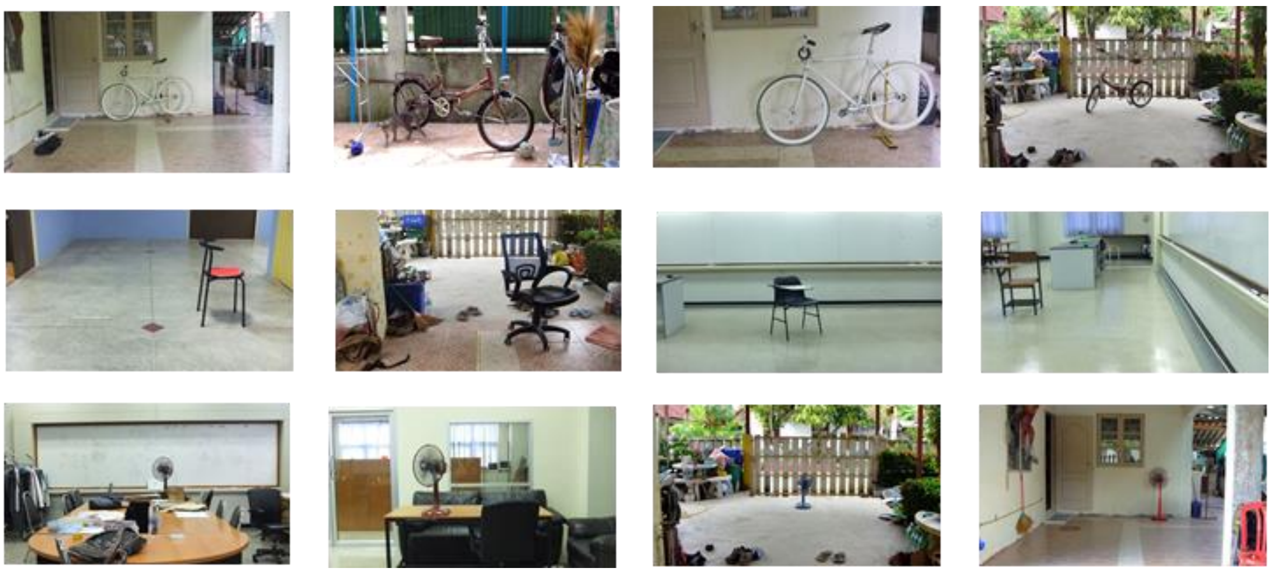


Figure 5. Examples of test images from stereo camera, (top) bicycle, (middle) chair and (bottom) stand fan.

where $S$ represents the final score map, $(i, j)$ is a coordinate of the scene image in the original scale, and $\sigma$ is a scaling parameter. Boundaries of the object can be determined based on the local maxima positions and thresholding method.

### *F. Object Location Estimation Using Stereo Images*

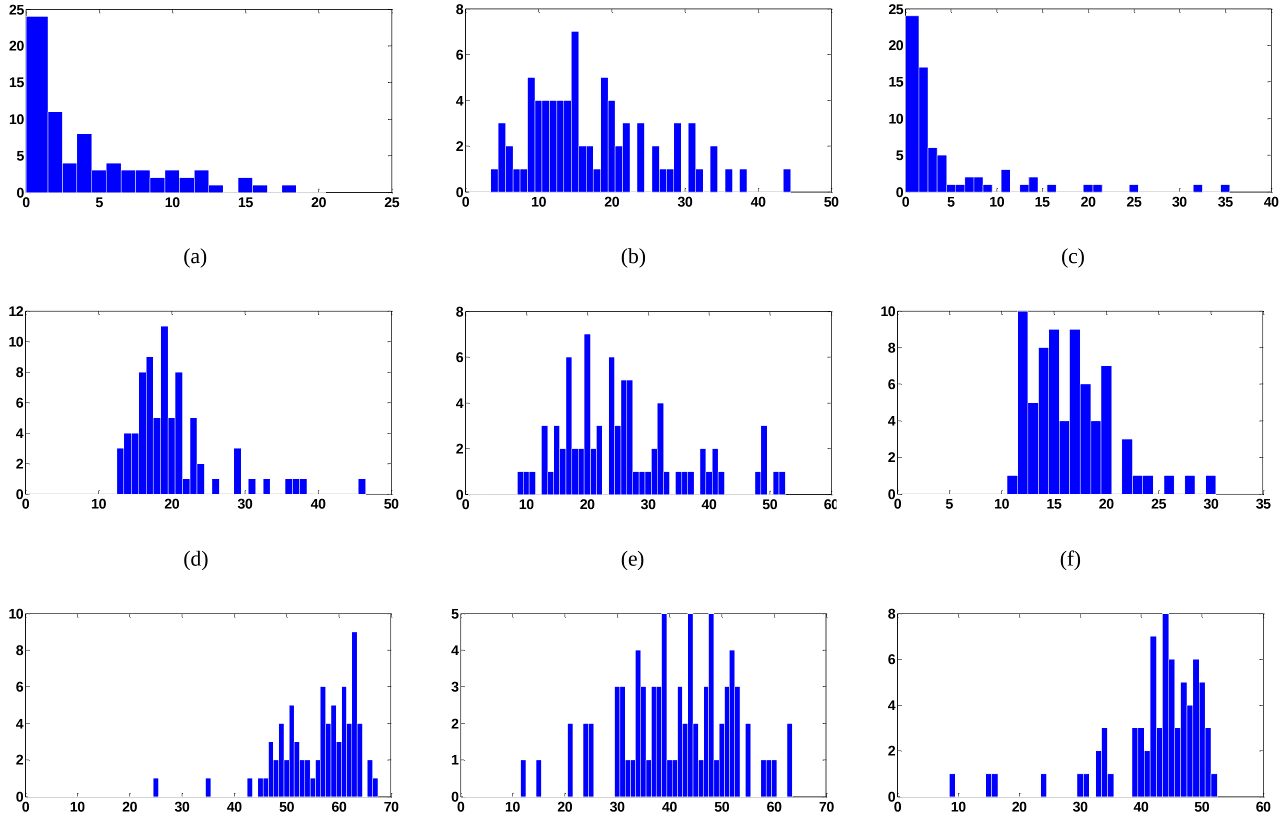
After the boundaries of the object are found in the scene image, a local feature matching such as Scale Invariant Feature Transform (SIFT) [9], Speed-Up Robust Feature (SURF) [10] is applied to extract the keypoints inside the object boundaries. Then, the keypoints from the first image are matched to keypoints in the other image. Let $(x_i^{(1)}, y_i^{(1)})$ and $(x_i^{(2)}, y_i^{(2)})$ denote coordinates of the matched pair of keypoints in left and right images, respectively. There are three constraints applied to filter out the wrong matched keypoints. The system removes the pairs of keypoints that

- Dissimilarities between descriptors of two keypoints are over the threshold,
- Disparities in vertical direction $\Delta x_i = x_i^{(2)} - x_i^{(1)}$ are not approximately zero, and
- Disparities in horizontal direction $\Delta y_i = y_i^{(2)} - y_i^{(1)}$ are outliers.

The disparity in horizontal direction of the stereo images can be transformed into the real world distance (depth) between the camera and the object for each keypoint by using triangular property [11]. The relation between the disparity and the depth in stereo images based on pinhole camera model can be written as

$$d_i = \frac{Bf}{\Delta y_i} \tag{23}$$

where $d_i$ represents the estimated depth of the $i$-th keypoint, $f$ is a focal length of the camera and $B$ is a distance



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 6. Examples of distibution of EOH parameters, (a) Empty Block EOH of bicycle, (b) Empty Block EOH of chair, (c) Empty Block EOH of stand fan, (d) Maximum EOH of bicycle, (e) Maximum EOH of chair, (f) Maximum EOH of stand fan, (g) Non-maximum Bin EOH of bicycle, (h) Non-maximum Bin EOH of chair, (i) Non-maximum Bin EOH of stand fan.

between two lens of the stereo camera. Finally, the depth of the target object can be estimated from a median of $d_i$ obtained from every matched keypoint in each object boundary.

After the depth have been estimated, the offset distance $L$ which is defined as the distance between the object and center axis of the stereo camera as shown in Fig. 1 can be estimated as a function of the depth and the pixel position as

$$L_i = (C_1 \cdot d_i + C_2) \cdot \Delta y_i + (C_3 \cdot d_i + C_4) \qquad (24)$$

where $C_i$ are constant values depending on the camera and the image size.

## III. Experimental Results

In this experiment, we tested our proposed framework by using three object classes that are bicycle, chair and stand fan. Since the bicycle is a widely used class of the object that tested in many object recognition schemes, in order to compare the result with the other baseline methods, the bicycle is chosen in this experiment although it is very rare case that the visually impaired people want to find the bicycle. On the other hand, chair and stand fan are typical objects that may be searched by the visually impaired people. The details of the experiment are described in the following sections.

### A. *Prototypes*

The prototype images are retrieved from Google image. The examples of prototype images are shown in Fig. 4. There are 84 images of bicycle class, 86 images of chair class and 90 image of stand fan class. Then, ratio selection process is used to cutoff some images with abnormal ratio. After selection process, there are 75 images of bicycle, 78 images of chair and 71 images of stand fan left to train the object detector.

### B. *Test Sets*

In this work, the test images are collected from the stereo camera model FinePix REAL 3D W3. We set up the objects in different environments and location. In this test set, the depths $d$ are varying from two to eight meters and the offsets $L$ are varying -2.50 to 2.50 meters. This test set contains 40 images of bicycle class, 40 images of chair class and 40 images of stand fan class. The examples of the test images are shown in Fig. 5.

### C. Features Extraction

In this experiment, we extracted the edge images by using Canny edge detection and divided the edge images into $5\times5$, $10\times10$ , $15\times15$ , and $20\times20$ blocks. Nine bins plus additional bin for empty bin are used in EOH and EOCM extraction. The distribution of empty block EOH, maximum EOH, non-maximum bin EOH and dominant direction are extracted from the prototype images. Fig. 6 shows the difference in distribution of three object classes. For example, the distribution of empty block EOH of the chair shifts to the higher value comparing to other objects because most of chair images have a large area of homogenous texture.

### D. Baseline Methods

We compared our proposed method with two baseline methods. I. Laptav's method [2] and P. Felzenszwalb's method [3]. I. Laptav provides the bicycle detector that is trained by the Pascal Visual Object Classes (VOC) dataset on the website (http://www.di.ens.fr/~laptev/download.html). On the other hand, P. Felzenszwalb provides the bicycle detector and chair detector that is is trained by VOC on the website (http://www.cs.berkeley.edu/~rbg/latent/index.html). Since the models were trained by a large number of positive images and backgrounds, we use the results form VOC trainned model as reference to measure the difficulty of the test images and object classes. In order to compare the result based on same training set that is retrieved from Google, we trained the P. Felzenszwalb's detector by using same positive samples with additional 300 background images ramdomly selected from website and VOC database.

### E. Results

In this experiment, we varied the threshold levels of our proposed method and the baseline methods and measured the precision, recall and F1 score. The bounding boxes of the object are extracted by using different detection methods. The depth and offset are estimated for each bounding box. The bounding boxes are considered to be correct if the depths and offsets are located in the object area within error margin of 50 cm. The object detection results of the proposed method are shown in Fig. 7. The final score maps are represented by heat map overlaid image to demonstrate the confidence level of that area to be the target object. The best results in term of F1 score for every method are illustrated in Table I. The results show that, in case of stand fan, the proposed scheme outperforms P. Felzenszwalb's method by using the same positive training samples and without negative samples from background. In case of chair, our method can provide the better result than P. Felzenszwalb's detector trained by both Google and VOC training sets. Finally, in case of bicycle, the proposed scheme also provides the better F1-score than P. Felzenszwalb's method by using the same Google training samples. The experiment shows that not only the accuracy of object detection from limited training samples can be improved by the proposed scheme but total time consumption for training and detection process is also dramatically reduced.

## IV. Conclusions

In this work, we proposed an internet based approach to detect and localize object for the visually impaired people. For object detection, the pre-prepared classifier is not needed. The framework requires only object keyword that user want to detect and localize. The scoring method based on EOH and EOCM features are introduced to obtain the fast training process. The experimental results show that the proposed scheme can provide the better performance than baseline method when a small number of internet based images are used as training samples. The proposed scheme can significantly reduce time consumption of training process in comparison with baseline method.

## Ackknowledgements

This research was supported by Graduate School, Chiang Mai University, Thailand.

## References

[1] Katz, Brian FG, et al. "NAVIG: guidance system for the visually impaired using virtual augmented reality." Technology and Disability 24.2 (2012): 163-178.

[2] SpikeNet Technology;. Available from: www.spikenettechnology.com.

[3] I. Laptev. "Improving object detection with boosted histograms." Image and Vision Computing 27.5 (2009): 535-544.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. "Object Detection with Discriminatively Trained Part Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep. 2010

[5] Robert M Haralick, K Shanmugam, Its'hak Dinstein (1 973). "Textural Features for Image Classification". IEEE Transactions on Systems, Man, and Cybernetics. SMC-3 (6): 610–621.

[6] Alefs, Bram, et al. "Road sign detection from edge orientation histograms." Intelligent Vehicles Symposium, 2007 IEEE. IEEE, 2007.

[7] Robert M Haralick, K Shanmugam, Its'hak Dinstein (1973). "Textural Features for Image Classification". IEEE Transactions on Systems, Man, and Cybernetics. SMC-3 (6): 610–621.

[8] Schroff, Florian, Antonio Criminisi, and Andrew Zisserman. "Harvesting image databases from the web." Pattern Analysis and Machine Intelligence, IEEE Transactions on 33.4 (2011): 754-766.

[9] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91 - 110.

[10] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006. 404-417.

[11] Sakuragi, Kei, and Akira Kawanaka. "Depth estimation from stereo images using sparsity." Signal Processing (ICSP), 2010 IEEE 10th International Conference on. IEEE, 2010.

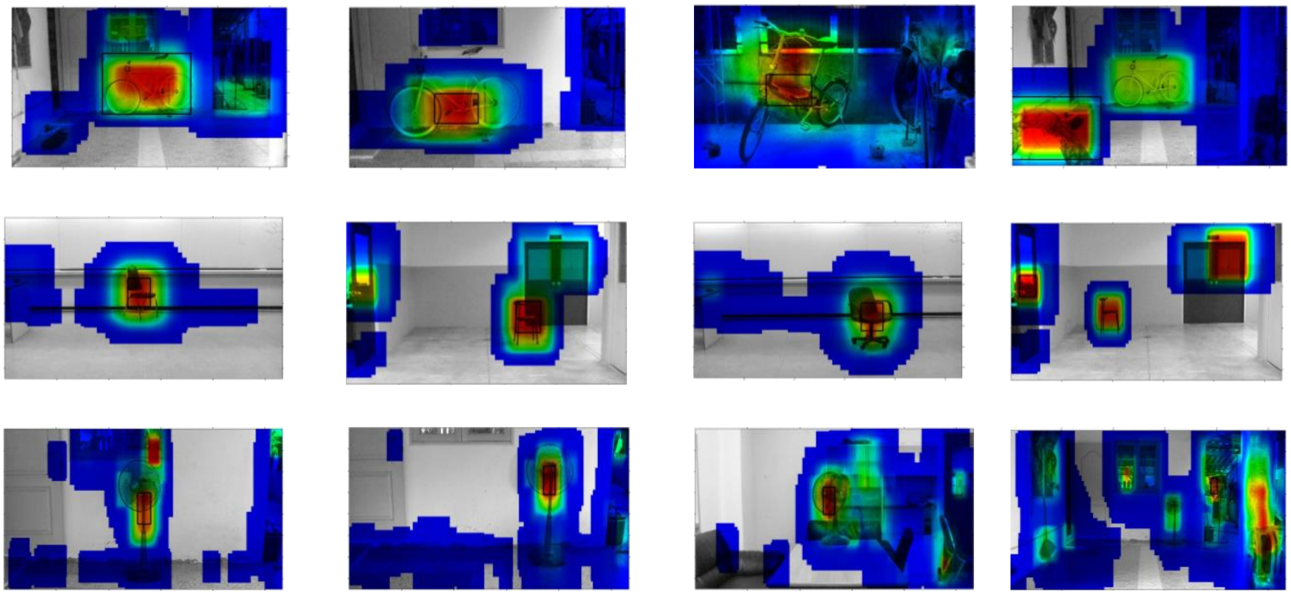


Figure 7. Examples of score map and detected bounding box of the proposed detector, (top) bicycle detector, (middle) chair detector and (bottom) stand fan detector.

TABLE I. PRECISION AND RECALL OF THE OBJECT LOCALIZATION USING DIFFERENT METHODS

| Class | #images | Proposed | | | P. Felzenszwalb, et. al. [3] | | | | | | I. Laptev [2] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Training Images retrieved from Google | | | | | | Training Images from VOC Dataset | | | | | |
| | | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| **Bicycle** | 40 | 0.660 | 0.875 | 0.753 | 0.714 | 0.500 | 0.588 | 0.947 | 0.900 | 0.923 | 0.973 | 0.900 | 0.935 |
| **Chair** | 40 | 0.535 | 0.775 | 0.633 | 0.371 | 0.575 | 0.451 | 0.444 | 0.800 | 0.571 | - | - | - |
| **Stand fan** | 40 | 0.263 | 0.525 | 0.350 | 0.074 | 0.050 | 0.060 | - | - | - | - | - | - |