Indonesian Journal of Electrical Engineering and Computer Science
Vol. 4, No. 2, November 2016, pp. 447 ~ 458
DOI: 10.11591/ijeecs.v4.i2.pp447-458

Crowd Detection in Still Images Using Combined HOG and SIFT Feature

Machbah Uddin*[1], HiraLal Gope[2], SayeedIftekhar Yousuf[3], Dilshad Islam[4], Mohammad Khairul Islam[5]
[1,3]Department of CSM, Bangladesh Agricultural University, Mymensingh-2202, Bangladesh
[2]Department of CSE, Sylhet Agricultural University, Sylhet-3100, Bangladesh
[4]Dept. of PMS, Chittagong Veterinary and Animal Sciences University, Bangladesh
[5]Department of CSE, University of Chittagong, Chttagong-4331, Bangladesh
*Corresponding author, e-mail:machbah.csm@bau.edu.bd

Abstract

Person detection and tracking in crowd is a challenging task. We detect the head region and based on this head region we can detect people from crowd. Individual object detection has been improved significantly in recent times but the crowd detection and tracking contains some challenges. Crowd analysis is a highly focused area for law enforcement, urban engineering and traffic management.There are a lot of incident occurred in crowd area during some fabulous event. In this research low resolution and verities of image orientation is a key factor as well as overlapping person images in crowd misguided the system. An enhanced system of interest point detection based on gradient orientation information as well as improved feature extraction HOG is used for identifying the human head or face from crowd. We have analyzed different types of images in different varieties and found accuracy 88-90%. In a number of applications, such as document analysis and some industrial machine vision tasks, binary images can be used as the input to algorithms that perform useful tasks. These algorithms can handle tasks ranging from very simple counting tasks to much more complex recognition, localization, and inspection tasks. Thus by studying binary image analysis before going on to gray-tone and color images, one can gain insight into the entire image analysis process.

Keywords: *crowd image, crowd detection, leveling image, connected component, hog, manual annotation, interest point.*

1. Introduction

Crowd detection is a highly focused area for law enforcement, urban engineering and traffic management. Public places such as shopping centers and airports are monitored using closed circuit television in order to ensure normal operating conditions. Automated analysis of crowd activities using surveillance videos is an important issue for communal security during violence, strike, heavy gathering allows detection of dangerous crowds and where they are headed. In case of surveillance, group behavior modeling and crowd disaster prevention people detection and tracking in crowd is a crucial component for wide range of application. Due to heavy occlusions, view variations and varying density of people as well as the ambiguous appearance of body the reliable person detection and tracking in crowd becomes a challenging task. Computer vision based crowd analysis algorithm can be divided into some groups; people counting, people tracking and crowd behavior analysis, movement analysis.

Person detection and tracking in crowd is a challenging task. Individual object detection has been improved significantly in recent times but the crowd detection and tracking contains some challenges. The head density of one person could be similar to another person density. The density of pedestrians significantly impacts their appearance in a video. For instance, in the videos with high density of crowds, people often occlude each other and usually few parts of the body of each individual are visible. On the other hand, the full body or a significant portion of the body of each pedestrian is visible in videos with low crowd-density. These different appearance characteristics require tracking methods which suite the density of the crowd [1].

Received September 3, 2016; Revised September 26, 2016; Accepted October 11, 2016

This research proposed a system that detect the head region and based on this head region that can detect people from crowd. The more accurate head detection can lead a good result for detecting a person in crowd domain. This research focused on gradient feature based image analysis and found a good accuracy rate of head detection described based on below Figure 1 and Figure 2 as sample.

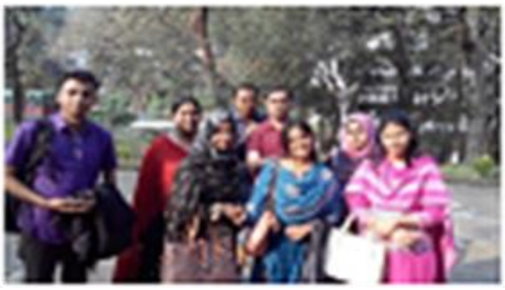


Figure 1. Sample Input Image to Detect Crowd

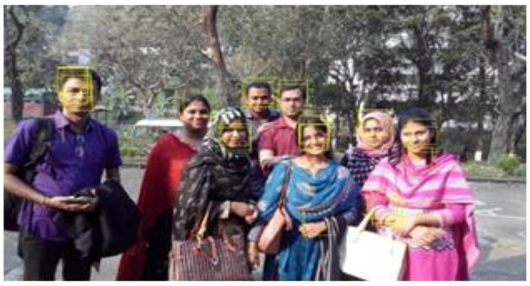


Figure 2. Detected Crowd Result

Concentrated on image gradient based people detection. Image gradient basically contains the directional changes information. This information can be used to track different objects or regions as well as boundary shape, getting a rough idea of an object location and other information. By analyzing the regions an assumption of item can be found that it is human or not in another word we can say that this step is important part selection or interest point detection. Natural images contain a lot of changes in orientation. So the number of important part may be large and huge as it is counting based on orientation information. There is some types of methods is needed to reduce this important part such as Adaboost and others. We applied different feature extraction technique to detect human on that region or from crowd place. We analyzed with HOG, SIFT and SURF feature. We used HOG and SIFT combined feature to test the result.

Assume a section that is a strong candidate for head region means an interest point that's may be head or not, is compared with trained support vector machine. Applying manual annotation technique we have prepared two classes of data, one is positive dataset another is negative dataset. During dataset preparation we have developed dynamic patch selection and its size. Supervised SVM is used to train with two dataset. All the candidate regions are tested with SVM. This test said which one is head or not. We got a marked output that processed with proposed method.

Next sections are organized as follow. Details of implementation in section three, preparing dataset with manual annotation in section four, experimental results are shown and compared with different methods and the next section contains the conclusion.

2. Literature Review

There are different approach to estimating the people count from an image and videos. Some of the approaches are density based; some of them are background subtraction based and some of them are corner based people detection [2]. Density based energy minimization framework which combines crowd density estimates with the strength of individual person detection [3] and tracking. Background subtraction based people detection which has a lower accuracy but having a good response with movement tracking [4-5]. Parts based object and crowd detection and tracking in still images have some research [6]. Scene priming to constrain locations of objects in the image [7]. Constant scale model, sliding window pyramid and scale window word density based people and parts of body detection as well as head detection is described in [6]. This paper describes a pedestrian detection system that integrates image intensity information with motion information. We use a detection style algorithm that scans a detector over two consecutive frames of a video sequence. The detector is trained (using AdaBoost) to take advantage of both motion and appearance information to detect a walking person. Past approaches have built detectors based on motion information or detectors based on appearance information [8] also development of a representation of image motion which is extremely efficient and implementation of a state of The art pedestrian detection system which operates on low resolution images under difficult conditions have analyzed [8]. Background subtraction results based estimation the number of people in a complicated scene which includes people who are moving only slightly an Expectation Maximization (EM)-based method has been developed to locate individuals in a low resolution scene [2]. A global data association method based on Generalized Graphs for tracking each individual in the whole video. In videos with high crowd-density, Track individuals Using a scene structured force model and crowd flow modeling [1]. Contextual information without the need to learn the structure of the scene based people detection [1]. Integral channel features for image classification tasks, focusing in particular on pedestrian detection [9] that focuses on feature extraction and feature description from RGB images. Integral channel features is that multiple registered image channels are computed using linear and non-linear transformations of the input image, and then features such as local sums, histograms, and Haar features and their various generalizations are efficiently computed using integral images [9]. Compute crowd density maps in order to estimate the spatial distribution of people in the scene also the self adaptive dynamic parameterization based people detection [10] basically that works based on human aspect ratio.

There are some approaches based on density, background subtraction that we mentioned in chapter two. There are different approaches of feature extraction for human, pedestrian or crowd detection. Some of the approach is integral channel feature, corner feature, are measurement, energy function behavior. Our approach is orientation of gradients based object area locating by defining the interest point, based on these points we have removed some of the points that are overlapping or not import. For selecting or detecting the interest points we have applied an RGB image to Gray image conversion. From gray image we have calculated the orientation and gradients in X, Y direction. Based on that orientation value we have defined a range maximum and minimum to get a binary image. From this binary image we can decide separate object region by applying connected component algorithm [11]. Before this step we have marked some of the area that is not important for our computation, as example the region that consisted with 20 pixels only we have discarded these regions. Removing this regions we get another binary image, in this image for finding a separate region, applied some connected component algorithm [11] here 8-connectivity is used. This step provides a binary image with the region information.

Now it’s turn to get some concrete value from each of the region, we have selected centroid value as concrete value of each region. Based on this centroid value we have selected a patch area of 201x201 from original RGB image.

Next steps is feature extraction, there was some previous approach like density, integrated channel feature and histogram etc. We have selected the Histogram Oriented Gradient (HOG) as a feature extraction technique. HOG is very much responsive to head or face detection because its computes the orientation changes based on color information changes that described more on chapter four. We have trained the SVM by our manual annotated dataset positive and negative both of the set at a time. Tested each of the HOG value with SVMStruct that means classify that point with trained SVM. Get the result from it. This result indicate the crowd face is detected. In case of non face and overlapping face this system also showing marked area as crowd people. Hence gradient information based binary image creation and the HOG feature based crowd face gives us a good result that’s around 90% accurate detection.

3. Proposed Method

This in this research we have formulated our orientation in below way that represents the overall architecture of our system. That’s started from taking an input image and ends with crowd people detection. Overall workflow in Figure 3.

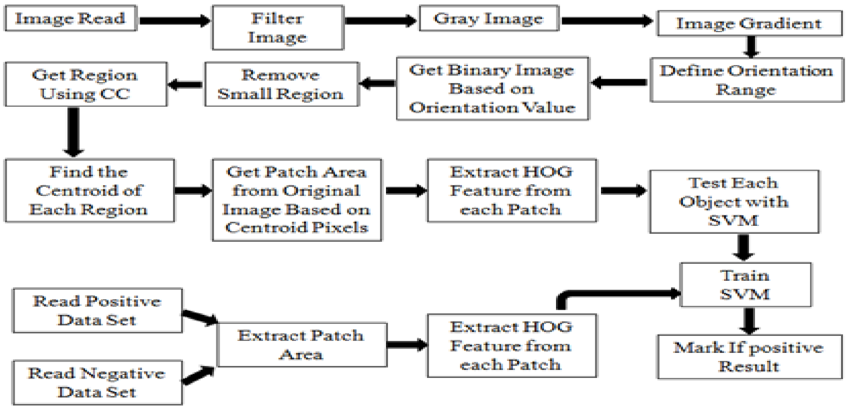


Image Read
Filter Image
Gray Image
Image Gradient
Define Orientation Range
Get Binary Image Based on Orientation Value
Remove Small Region
Get Region Using CC
Find the Centroid of Each Region
Get Patch Area from Original Image Based on Centroid Pixels
Extract HOG Feature from each Patch
Test Each Object with SVM
Train SVM
Mark If positive Result
Read Positive Data Set
Read Negative Data Set
Extract Patch Area
Extract HOG Feature from each Patch

Figure 3. Workflow of Crowd Detection

3.1. Preprocessing

Preprocessing starts with the smoothing or sharpening an image by any technical terminology like filtering. The Gaussian smoothing operator is a 2-D convolution operator that is used to blur images and remove detail and noise it uses a different kernel that represents the shape of a Gaussian (bell-shaped). The Gaussian distribution in 1-D has the form

$$\mathrm{G(x)} = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{\mathrm{x}^2}{2\sigma^2}} \tag{3.1}$$

and in 2-D

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2+\mathrm{y}^2}{2\sigma^2}} \tag{3.2}$$

where σ is the standard deviation. X, Y is the image pixel location.

This Gaussian filter applies in our input image Figure 1. And the resultant filtered image is in Figure 4.

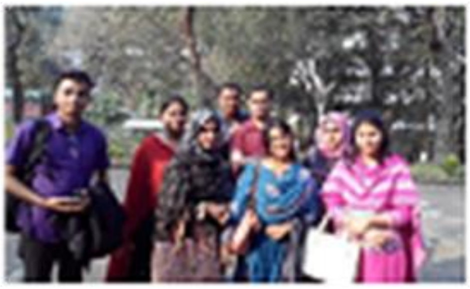


Figure 4. After Applying Gaussian in Figure 1

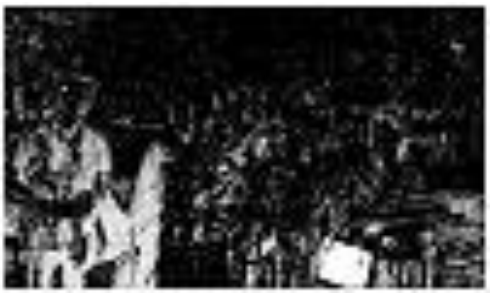


Figure 5. Binary After Removing Small Regions

Image Gradient

An image gradient is a directional change in the intensity or color in an image. Image gradients may be used to extract information from images. Let an image is I and the gradient in x is $xG(x)$ and in y is $yG(y)$ direction and that is computed by below equation

$$g_x(i,j) = I(i,j-1) - I(i,j+1) \tag{3.3}$$

$$g_y(i,j) = I(i-1,j) - I(i+1,j) \quad (3.4)$$

Where i and j represents the image index hence the magnitude is M and Orientation is O

$$M(i,j) = \sqrt{g_x(i,j)^2 + g_y(i,j)^2} \quad (3.5)$$

$$O(i,j) = \tan^{-1}\frac{g_y(i,j)}{g_x(i,j)} \quad (3.6)$$

Here are the steps how image gradient and magnitude is calculated and the results of every steps [12].

3.2. Interest Point Detection

3.2.1. Binary Image

By analyzing the data of orientation in details we found the maximum value is **179.85** and minimum value is **-180**. We have to define a range value. Based on this range value the binary image is formed [13]. Formula of Binary image is:

$$B(i,j) = \begin{cases} 1 \; if O(i,j) \in [\pi_l, \pi_u] \\ 0 \; otherwise \end{cases} \quad (3.7)$$

Where π_l is the minimum allowed value and π_u is the maximum allowed value. In our experimental setup the $\pi_l = -10$ and$\pi_u = 10$. If the pixel orientation value is in range value then the binary value is 1 otherwise the value is zero [13-14]. Above binary image containing all the region with different size [14]. After removing small area with 20 pixels binary image result is in Figure 5. Its reduces the small blobs and others sections.

3.2.2. Connected Component

Binary image that found on previous steps contains an image which is not connected or group of objects. This steps we are applying some connected component mechanism to find which one is connected and which of the groups worked together to make a region. In 8-connectivity pixels are connected if their edges or corners touch. This means that if two adjoining pixels are on, they are part of the same object, regardless of whether they are connected along the horizontal, vertical, or diagonal direction. Below is the pattern of 4 and 8 connected component. Using the 8 connectivity in Figure 4 we find that there are **5360** individual objects [11-12]. Each of the objects contains the different numbers of pixels that is covered by that region.

Now in all of the regions we have we have to summarize it. For that we are taking the centroid value of that region. Different centroid is in Figure 6.

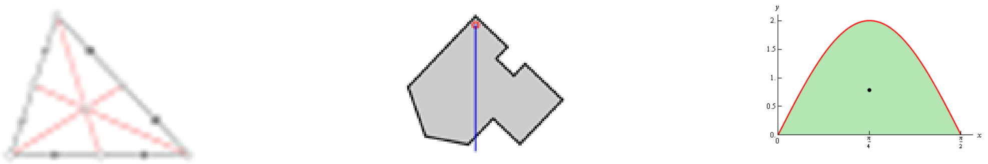


Figure 6. Different Centroid Structure

3.2.3. Centroid of a Finite Set of Points

The centroid of a finite set of k points$x_1, x_2, x_3 \ldots \ldots .. x_k$ in R^n is

$$C = \frac{x_1+x_2+x_3 \ldots\ldots +x_k}{k} \quad (3.8)$$

This point minimizes the sum of squared Euclidean distances between itself and each point in the set.

This centroid is our interest point centre. After taking the cell values we have reduced some of the overlapping points. Hence its reduces the size from **5360** to **3500** points.

Now consider this region as interest point. From this centroid value take a patch size that is 201x201 sizes. Some example patch size is give in Table 1.

All of the patch area is considered as an important point. Descriptor is generated from this patch area. Basically HOG and SIFT and SURF is produced.

Table 1. Some Patch Area from Sample Input

3.2.4. Feature Extraction

Patches are that we found in interest point detection section. In this steps needed to extract some features from that patch area. Our patch size is 201x201, first apply HOG features to get patch.

Histogram of Oriented Gradients descriptors or HOG descriptors, are feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. This feature will be used in object detection and classification.

HOG feature have been introduced by NavneedDaala and Bill Triggs who have developed and tested several variants of HOG descriptors, with differing spatial organization, gradient computation and normalization methods.

3.2.5. Algorithm of HOG implementation

a) Gradient computation

The first step of calculation in many feature detectors in image preprocessing is to ensure normalized color and gamma values. The most common method is to simply apply the 1-D centered, point discrete derivative task in one or both of the horizontal and vertical directions. Specifically, this method requires filtering the color or intensity data of the image with the following filter kernel $[1,0,1]$and$[-1,0,1]^T$. Here 3 X 3 Sobel masks and other masks like Gaussian mask are used.

b) Orientation binning

The second step of calculation involves creating the cell histograms in Figure 5. Each pixel within the cell casts a weighted vote for an orientation-based histogram [13] channel based on the values found in the gradient computation.

c) Descriptor blocks

In order to account for changes in illumination and contrast, the gradient strengths must be locally normalized; In addition, these C-HOG blocks can be described with four parameters: the number of angular and radial bins, the radius of the center bin, and the expansion factor for the radius of additional radial bins.

d) Block normalization

Dalal and Triggs explore four different methods for block normalization [13]. Let be the non-normalized vector containing all histograms in a given block, $\|v\|_k$ be its k-norm for k=1, 2, 3 and e be some small constant (the exact value, hopefully, is unimportant). Then the normalization factor can be one of the following:

$$\text{L2-norm: } f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \qquad (3.9)$$

L2-hys: L2-norm followed by clipping (limiting the maximum values of v to 0.2) andrenormalizing,

$$\text{L1-norm: } f = \frac{v}{(\|v\|_1 + e)} \qquad (3.10)$$

In addition, the scheme L2-Hys can be computed by first taking the L2-norm, clipping the result, and then renormalizing.

After applying the above steps to get HOG value from any of the samples we have created our descriptor from here. We found a vector of 1x20736 sizes for a single patch area. There are some sample values in Table 2.

Table 2. Sample Patch HOG Values

1st	2nd	3rd	4th	..	..	20735th	207356th
0	0.010255	0.014258	.144286	..	..	0.002242	0.001514

Manual annotation dataset preparation is described in next section. By this time our Classifier Support Vector Machine is trained from Dataset. There are two sets of data. Positive and negative data set. These HOG value from image is send to classifier to test the result. Based on this result if it is positive we have marked as final head or crowd detected output that is mention in Figure 2.

3.2.6. SIFT - Scale Invariant Feature Transforms

The SIFT approach, for image feature generation, takes an image and transforms it into a "large collection of local feature vectors each of these feature vectors is invariant to any scaling, rotation or translation of the image. This approach shares many features with neuron responses in primate vision. To aid the extraction of these features the SIFT algorithm applies a 4 stage filtering approach: Steps of SIFT

a) Scale-Space Extrema Detection

This stage of the filtering attempts to identify those locations and scales that are identifiable from different views of the same object. This can be efficiently achieved using a "scale space" function. Further it has been shown under reasonable assumptions it must be based on the Gaussian function

b) KeypointLocalistaion

This stage attempts to eliminate more points from the list of keypoints by finding those that have low contrast or are poorly localised on an edge

c) Orientation Assignment

This step aims to assign a consistent orientation to the keypoints based on local image properties. The keypoint descriptor, described below, can then be represented relative to this orientation, achieving invariance to rotation

d) Keypoint Descriptor

The local gradient data, used above, is also used to create key point descriptors. The gradient information is rotated to line up with the orientation of the key point and then weighted by a Gaussian with variance of 1.5 * key point scale

In our method we have used combined SIFT and HOG feature as a feature extractor.

4. Experemental Results and Discussion

4.1. Preparing Dataset with Manual Annotation

Manual annotation for preparing data set is a process in which we can select and crop specific portion of an image that will used to evaluate our interest point is head or not. Here we have developed two programs one is responsible for preparing positive dataset and another one is preparing negative dataset. It produces dataset with annotation positive and negative.

Steps to create positive dataset:

This program will read the entire image from source data. It will display an image and will ask to select some positive data means some of human face from this image just in Figure 7.

Figure 7. Crop Positive Data Manual Annotation

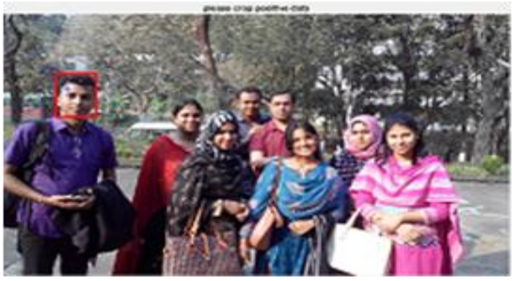


Figure 8. Positive Face is selected

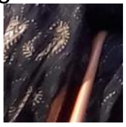
Then select a portion by double clicking on image. It will select a rectangle shape face in Figure 8. This way from all the images we have selected faces area and produced dataset that in Table 3.

Table 3. Positive Dataset Example

Steps to create negative dataset:

Start the program for negative dataset cropping. Follow the steps like positive dataset cropping. Choose point. Double click on image. Negative data will store in folder. Below is the some negative data set snapshots in Table 4.

Table 4. Negative Dataset Example

4.2. Results

In our research there are 2 classes of image. At first step we have applied folding method on two classes of images that classes are Inside Room and Outside Room class and we see that our proposed method based crowd detection gives a good result. Each of the images containing different number of people with different verities. Inside room contains 4/5 people in images. Outside room contains 8/10 people per snap shot. Now the number of identified people in all images that we test using our positive and negative dataset result is in Table 5.

Table 5. Number of Class Accuracy

Class Name	Number of People	Successfully identified	Not detect
Inside Room	4	4	0
Outside Room	10	9	1

Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

K-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter.

4.3. K-Fold Cross-Validation

In stratified k-fold cross-validation [15-16], the folds are selected so that the mean response value is approximately equal in all the folds. In the case of a dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels.

2 fold-cross validation, this is the simplest variation of k-fold cross-validation. For each fold, we randomly assign data points to two sets d0 and d1, so that both sets are equal size (this is usually implemented as shuffling the data array and then splitting in two). We then train on d0 and test on d1, followed by training on d1 and testing on d0. This has the advantage that our training and test sets are both large, and each data point is used for both training and validation on each fold.

Now 2 folding method applied on ten classes of image, that means 50 percent image of a class are in training set and 50 percent images of that class are on test, that started 2 class,3 class,4 class,5 class,6 class. The result given on Table 6.

Table 6. Results Per Class Success

Number of image	Classes Name	Number of Person per image	Total Number of person	Person Correctly Identified	Person Wrong Identified	Accuracy
10	Inside Room	4/5	43	38	5	88.80%
10	Outside Room	8/10	91	82	9	90.03%

Per class success and failed rate in Figure 9. The next section that describes the accuracy rate of two class and the expected class in following chart in Figure 10.

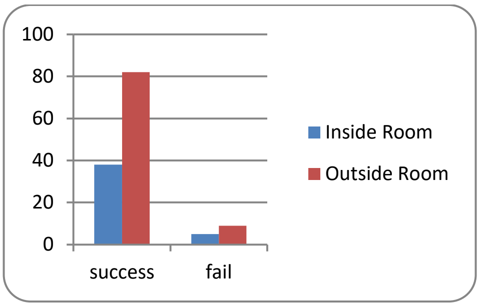


Figure 9. Success and Failure per Class

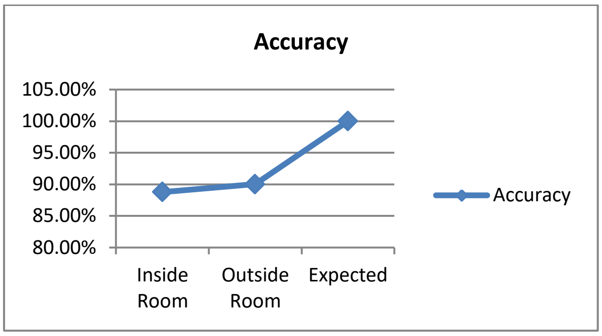


Figure 10. Accuracy Graph of Different Class

4.4. Recall and Precision Graph

In pattern recognition precision [8][15] is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. Suppose a program for recognizing dogs in scenes identifies 7 dogs in a scene containing 9 dogs and some cats. If 4 of the identifications are correct, but 3 are actually cats, the program's precision is 4/7 while its recall is 4/9. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3.

The confusion matrix

When referring to the performance of a classification model, we are interested in the model's ability to correctly predict or separate the classes. When looking at the errors made by a classification model, the confusion matrix gives the full picture. Consider e.g. a three class problem with the classes A, B, and C. A predictive model may result in the following confusion matrix when tested on independent data (Table 7).

Table 7. Confusion Matrix

		Predicted class		
		A	B	C
Known class (class label in data)	**A**	25	5	2
	B	3	32	4
	C	1	0	15

Table 8. Confusion Matrix with Notation

		Predicted class		
		A	B	C
Known class (class label in data)	**A**	tp_A	e_{AB}	e_{AC}
	B	e_{BA}	tp_B	e_{BC}
	C	e_{CA}	e_{CB}	tp_C

The confusion matrix shows how the predictions are made by the model. The rows correspond to the known class of the data, i.e. the labels in the data. The columns correspond to the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column for examples with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made. In the calculations in Table 8, we will also use this abstract confusion matrix for notation.

Precision:

Precision is a measure of the accuracy provided that a specific class has been predicted.

It is defined by:

$$Precision = tp/(tp + fp) \tag{4.1}$$

where tp and fp are the numbers of true positive and false positive predictions for the considered class. In the confusion matrix above, the precision for the class A would be calculated,

$$Precision_A = \frac{tp_A}{tp_A + e_{BA} + e_{CA}} = 25/(25 + 3 + 1) \approx 0.86 \tag{4.2}$$

The number is reported by RDS as a value between 0 and 1.

Recall:

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly also called sensitivity, and corresponds to the true positive rate. It is defined by the formula:

$$Recall = Sensitivity = tp/(tp + fn) \tag{4.3}$$

where tp and fn are the numbers of true positive *and* false negative predictions for the considered class. $tp + fn$ is the total number of test examples of the considered class. For class A in the matrix above, the recall would be:

$$\begin{aligned} Recall_A = Sensitivity_A &= tp_A/(tp_A + e_{AB} + e_{AC}) \\ &= 25/(25+5+2) \approx 0.78 \end{aligned} \tag{4.4}$$

The sensitivity of recall and precision of that output is given in below Figure 11.

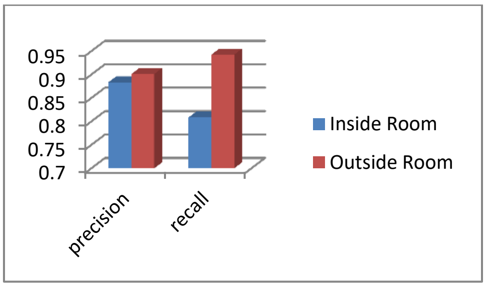


Figure 11. Recall and Precision Graph

5. Conclusion

In this dissertation achieved a good performance that is shown on previous graph and its accuracy rate is high that is above 85 percent. Selft database contains images in variety of format on same class. Images taken from different lighting condition different orientation and different position.Specially gradient orientation information based candidate selection and finally recognizing with HOG works fine for human head detection and crowd detection. This gives a better result in outside area. That's the prominent side of this research.

6. Future Directions

In future we will concentrate on increasing accuracy and reducing processing time. We will try to obtain processing time 0.05s per image. Will also work with low resolution and real time camera data from any source. In another part we will concentrate on huge gathering peoples from different angles with back side also.

Acknowledgment

In this work I am grateful to **Dr. Mohammad Khairul Islam**, Professor Department of Computer Science & Engineering, University of Chittagong, Bangladesh for his idea. I have had inspired by him as well as the necessity of current world.

References

[1] A Dehghan, H Idrees, M Shah. Automatic Detection and Tracking of Pedestrians in Videos with Various Crowd Densities. *Springer*. 2014.

[2] Hou YL, Pang GKH. People counting and human detection in a challenging situation. *IEEE Transactions on Systems*. Man, Cybernetics-Part A: Systems and Humans. 2011: 41 (1).

[3] R Rodriguez, I Laptev, J Sivic. Density-aware person detection and tracking in crowds. *Computer Vision (ICCV).* 2011.

[4] YJ, VS, A Davies. Image processing techniques for crowd density estimation using a reference image. *In ACCV*, 1995: 3: 6–10.

[5] R. Ma, LL, HW, TQ. *On Pixel Count Based Crowd Density Estimation For Visual Surveillance.* In IEEE Conf. Cybernet. Intell. Syst. 2004: 1.

[6] OgnjenArandjelovic. Crowd Detection from Still Images. *BMVC.* 2008 doi:10.5244/C.22.53.

[7] A Torralba. Contextual priming for object detection. IJCV. 2003: 53 (2).

[8] Powers, David MW. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*. 2011; 2 (1): 37-63.

[9] P Dollar, Z Tu, P Perona, S Belongie. Counting Integral Channel Features. *IEE Explorer.* 2012.

[10] VEiselein, H Fradi, I Keller, JDugelay, *Enhancing Human Detection using Crowd Density Measures and an Adaptive Correction Filter*. 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance.

[11] http://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm.

[12] https://en.wikipedia.org/wiki/Connected-component_labeling.

[13] V B Subburaman, A Descamps, C Carincotte. Counting people in the crowd using a generic head detector.*In IEE AVSS*. 2012.

[14] http://stackoverflow.com/questions/19815732/what-is-gradient-orientation-and-gradient-magnitude.

[15] http://www.cs.cmu.edu/~schneide/tut5/node42.html.

[16] Kohavi, Ron. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. 1995; 2(12). 1137-1143.