Comparative Feature Extraction and Analysis for Abandoned Object Classification

Ahmed Fawzi Otoom, Hatice Gunes, Massimo Piccardi
Faculty of Information Technology
University of Technology, Sydney (UTS)
Sydney, Australia
{afaotoom, haticeg, massimo}@it.uts.edu.au

***Abstract* - We address the problem of abandoned object classification in video surveillance. Our aim is to determine (i) which feature extraction technique proves more useful for accurate object classification in a video surveillance context (scale invariant image transform (SIFT) keypoints vs. geometric primitive features), and (ii) how the resulting features affect classification accuracy and false positive rates for different classification schemes used. Objects are classified into four different categories: bag (s), person (s), trolley (s), and group (s) of people. Our experimental results show that the highest recognition accuracy and the lowest false alarm rate are achieved by building a classifier based on our proposed set of statistics of geometric primitives' features. This set of features maximizes inter-class separation and simplifies the classification process. Classification based on this set of features thus outperforms the second best approach based on SIFT keypoint histograms by providing on average 22% higher recognition accuracy and 7% lower false alarm rate.**

***Keywords* — Abandoned object classification, video surveillance, statistics of geometric primitives, SIFT keypoints.**

I. Introduction

Automatic recognition, description, classification and grouping of patterns have been identified as significant problems within the computer vision research community and have been tackled for decades. In recent years, there has been growing interest and effort in developing research approaches for recognizing objects in still images. The majority of these approaches focus on extracting local regions such as Difference of Gaussian (DoG) regions [7], saliency regions [6], or other types of local patches. A discriminative model for recognition is then built based on these features such as: constellation models [5], "bag of words" models [12], and others. Results of these approaches are promising for objects categorization. However, the extracted features depend largely on local regions, such as corners and textured patches, therefore recognize objects only from one viewpoint and might not be accurate for recognizing objects when the viewpoint changes (e.g. [5]).

Object classification in video surveillance has also gained more attention recently. It aims to classify objects of interests into a number of predefined categories. Object categories are defined in advance depending on the environment where these objects are likely to be detected in the scene. Images of objects of interest are first analyzed in order to choose features that are simple yet efficient to discriminate between the predetermined classes. Extracted features should be robust to various challenging conditions such as occlusion and change in viewpoint and illumination. In general, moving object recognition has gained more attention than abandoned object recognition [3, 11]. However, abandoned objects need to be detected and classified in an accurate way due to the fact that such objects may represent a high security threat. Efficient and accurate classification is needed in order to assess the potential danger they might cause prior to taking appropriate actions. Existing approaches for abandoned object recognition mainly depend on extracting a limited number of shape or appearance features [2, 8], resulting in a classifier that may not be capable of addressing the various challenges faced in a surveillance environment (e.g. [8]).

Within the rich body of literature on object and/or object class recognition, it is often stated that great attention should be paid to the definition of a discriminative feature set. There exist previous works for evaluating the performance of feature extraction techniques based on different local region descriptors and across a number of classifiers (e.g. [9]). However, there has been no attempt to compare local region features with statistics of geometric primitives' features in a visual surveillance context. Accordingly, in this paper, we aim to determine (i) which feature extraction technique proves more useful for accurate object classification in a video surveillance context (scale invariant image transform (SIFT) keypoints vs. geometric primitive features), and (ii) how the resulting features affect classification accuracy and false positive rates for different classification schemes used.

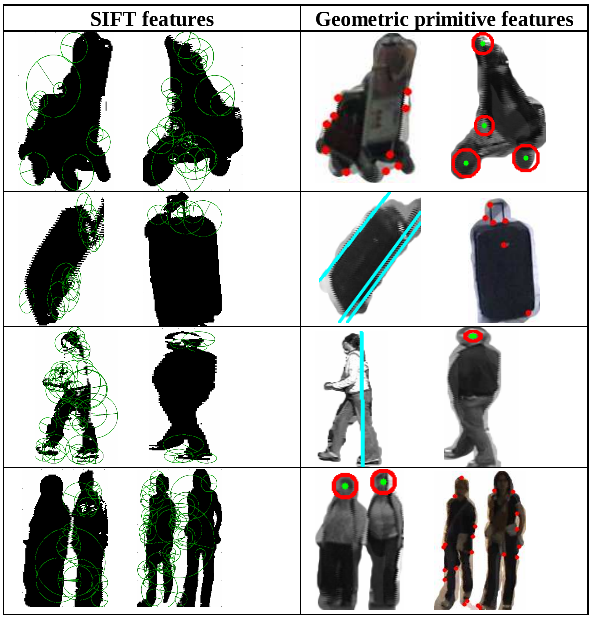

Figure 1. Examples of features detected in a number of images: trolley (1st row), bag (2nd), person (3rd) and group of people (4th).

The work presented in this paper aims to become an integral part of a video surveillance system framework that is able to track multiple people and automatically detect abandoned objects for security of crowded areas such as a railway station or an airport terminal. Our work is based on the assumption that the abandoned object is already detected by a detector of "new stationary objects" in the scene; its location and size are also made available. A commercial off-the-shelf technology product (e.g., [15]) can be used for this task. We also assume that the area of interest is located within an airport or train station, and the objects of interest consist of trolley(s), bag(s), single person and group(s) of people. The problem at stake should not be confused with generic object classification, for which several methods exist suited to variable number and type of object classes ([5-7] and others), instead, given the high cost associated with misclassification errors in a surveillance context, we aim to devise the most accurate feature extraction procedure possible given the categories of interest. The remainder of this paper is organized as follows: in Section 2 we introduce the feature extraction techniques. Classification learning methods and performance evaluation are described in Section 3. Experimental results and analysis are presented in Section 4. Finally, we draw our conclusions in Section 5.

II. Feature Extraction

The first step in any classification problem is feature extraction where features are extracted from images based on different image information. We apply three different approaches for extracting features. These approaches are based on SIFT keypoints and statistics of geometric primitives.

A. *SIFT keypoints*

SIFT (Scale-Invariant Feature Transform) keypoints are known to be invariant to rotation, scale, and translation, and are used to detect distinctive edges and textures in an image. Moreover, SIFT has empirically outperformed many other descriptors [9]. Because of the aforementioned reasons we choose to apply SIFT for the detection and description of local features (keypoints). Each keypoint is described with a 132-dimension vector: 128 spatial orientations, plus coordinates, scale, and rotation. After extracting SIFT keypoints from all images, we first apply dimensionality reduction and then we apply two different approaches for the final description of the features as illustrated in the following subsections. Fig. 1 (left column) shows examples of SIFT keypoints detected in a number of images.

1) Dimensionality reduction: After extracting SIFT keypoints, it is necessary to reduce the dimensionality in order to extract significant information and be capable of training classifiers. We apply two popular dimensionality reduction techniques: principle component analysis (PCA) and linear discriminant analysis (LDA). From the initial analysis of the results, both techniques seem similar in their performance for the final classification results, with PCA slightly outperforming LDA. Therefore, we present PCA-based results. PCA is an orthogonal transformation of the coordinate system that describes the data. Given a set of M centered observations $x_i \in R^N$, $i = 1,....,m$, $\sum_{i=1}^{m} x_i = 0$, PCA finds the principle axes by diagonalizing the covariance matrix

$$C = \frac{1}{m}\sum_{i=1}^{m} x_j x_j^T \qquad (1)$$

To provide the diagonalization, the Eigenvalue equation $\lambda v = Cv$ has to be solved where v is the Eigenvector matrix. The first few Eigenvectors are used as the basis vectors for the lower dimensional space. PCA aligns the data along the directions of the greatest variance. We keep only the eigenvectors corresponding to the highest eigenvalues, capturing 90% the variance within the data set. We thus reduce the dimensionality of the keypoint vectors down from 132 to 3. After applying PCA, we apply two approaches for the final description of the SIFT keypoints: majority rule approach and keypoint histograms approach.

2) Approach 1: SIFT keypoints and majority rule: In this method, each keypoint in an image is classified independently and the final decision for the image class is the same class assigned to the majority of its keypoints. Let x be the class assigned to keypoint i in an image M and

$d(x \mid f_i)$ be the binary decision $(0|1)$ for a keypoint i given feature vector f_i. Since x is one of four classes (person, group, bag, trolley), then $d(x \mid f_i)=1$ for only one class and 0 for all the others. For each image M, using the number of keypoints denoted as T, the multiple decisions are added up, for each class separately, as: $D(x \mid f_1 .... f_T) = \sum_{1}^{T} d(x \mid f_i)$. The final class assigned to M will then be

$$x^* = \arg\max_x \left(D(x \mid f_1 .. f_T)\right) \qquad (2)$$

3) Approach 2: SIFT keypoint histograms: As our main goal is that of comparing feature extraction techniques, this approach was inspired by [1], except that we apply PCA instead of LDA for the feature reduction. We create a keypoint histogram for each image allowing the relationships between numbers and types of keypoints to be extrapolated and the information on the actual location discarded. Following this rationale, we first apply PCA to each keypoint, as explained before. Secondly, we choose a number of bins for each feature to be approximately proportional to the data variance within that feature. Eventually we use a histogram with 6, 4 and 2 bins for 1-3 features obtained from PCA. The resulting histograms are then fed into the classifiers for object classification.

B. Approach 3: Statistics of geometric primitives

In [10], we analyzed a number of images for the four objects of interest (bags, trolleys, persons, and groups of people), and propose an effective feature set capable for discriminating the four classes with a high detection rate and a low false alarm rate. The features in the set represent the main statistics of geometric primitives for an object such as: corners, lines, circles, and other related statistics [10].

We follow the same approach and extract these features with the addition of the fitting ellipsis aspect ratio and the dispersion of the object. The fitting ellipse aspect ratio is calculated as the ratio between the length of minor axes and the length of major axes of the fitting ellipse. We further calculate the perimeter (the length of the external contour) and the area *(*the area under the external contour). The *dispersion* of an object is calculated as the ratio between the square of the perimeter and the area of the object. A full list of the features is illustrated in Table 1 and further described in [10]. Moreover, Fig. 1 (right column) shows such features as extracted in a number of images.

TABLE 1. LIST OF STATISTICS OF GEOMETRIC PRIMITIVES' FEATURES.

Corners	Circles	Lines	Other features
- No. of corners. - The ratios and percentages between corners. - Horizontal and vertical StDev.	- No. of circles. - The ratios and percentages between circles. - Horizontal and vertical StDev.	- No. of lines (strong, intermediate, and weak). - No. of horizontal, vertical, diagonal lines, and ratios between them.	- Bounding box dispersion & Height/Width ratio - Fitting ellipse aspect ratio - Object dispersion

III. Classification

The classifiers that have been used for the classification experiments in our system are the Bayesian-based classifier BayesNet, C4.5 or Decision Trees, Sequential Minimal Optimization (SMO) algorithm [14] , and MultiBoostAB (a variant of AdaBoost combining wagging and boosting) [13]. The performance of the classifier is evaluated in terms of classification accuracy (or detection rate for each class) and false positive rate (FPR). Classification accuracy is calculated as the proportion of the number of objects correctly detected against the total number of objects. The false positive rate is calculated as the proportion false positives against the sum of true negatives and false positives.

IV. Experimental Results and Analysis

Experiments are conducted in order to compare different feature extraction techniques and evaluate their performance across a number of classifiers. For this purpose, we collected 600 images of trolleys, bags, single persons, and groups of people. These images were collected from video footage provided by our industrial partner and were taken in a number of airports around the world. Objects of interest in these images appear from different viewpoints, under different illumination conditions and in varying size and scale. We carried out two experiments in order to validate the chosen approaches with the holdout method and k-fold cross-validation. For the first validation method, we partitioned the 600 images into two independent data sets, a training data set of 400 images and a test data set of 200 images, with equal number of images for each class. For the second validation method, we used all 600 images with 10-fold cross-validation. In this validation method the original sample is partitioned into 10 subsamples, of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data.

TABLE 2. CLASSIFICATION RESULTS AS A RANGE ACROSS MULTIPLE CLASSIFIERS FOR THE THREE APPROACHES USING HOLDOUT VALIDATION.

	Classification Accuracy	False Positive Rate
1 - SIFT keyp.	38% - 44.5%	20.6% – 22.8%
2 - SIFT hist [1]	44.5%-57.5%	14.2 %-18.5%
3 – Our approach	72% - 79.5%	6.8% - 9.3%

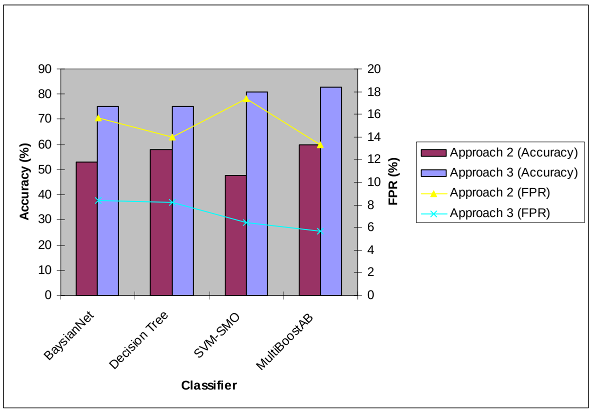

Figure 2. Classification results across multiple classifiers for 600 images using 10-fold cross-validation for SIFT keypoint histograms (approach 2) and statistics of geometric features (approach 3).

For approach 1 and approach 2, we first extract SIFT keypoints and then apply PCA in order to reduce the dimensionality. In approach 1, we apply the majority rule described in Section 2 and then feed the results to the four different classifiers mentioned in previous section. For approach 2, a histogram is built for the reduced dimensions and the results are also fed to the multiple classifiers. Finally, for our approach (approach 3), we extract lines, circles, corners, and all other related statistical features and also feed them to the same classifiers. The results of classification based on these approaches are presented in Table 2, where classification accuracy and FPR are presented as a range across multiple classifiers, from the minimum to the maximum percentages. It is clear from Table 2 that building a histogram for the SIFT keypoints outperforms the majority rule approach. The integral and non-local nature of the histogram as a feature results in a higher performance. However, by looking at Table 2, we observe that the highest performance is achieved by our approach (approach 3), which is based on statistics of geometric primitives.

Estimating the classification accuracy and the false positive rate by using a 10-fold cross-validation in general provides a better estimate than the estimate obtained from one single holdout test [4]. We thus chose approach 2 and approach 3 for this part of the experiment as they provided the best results in the holdout validation test. Fig. 2 presents

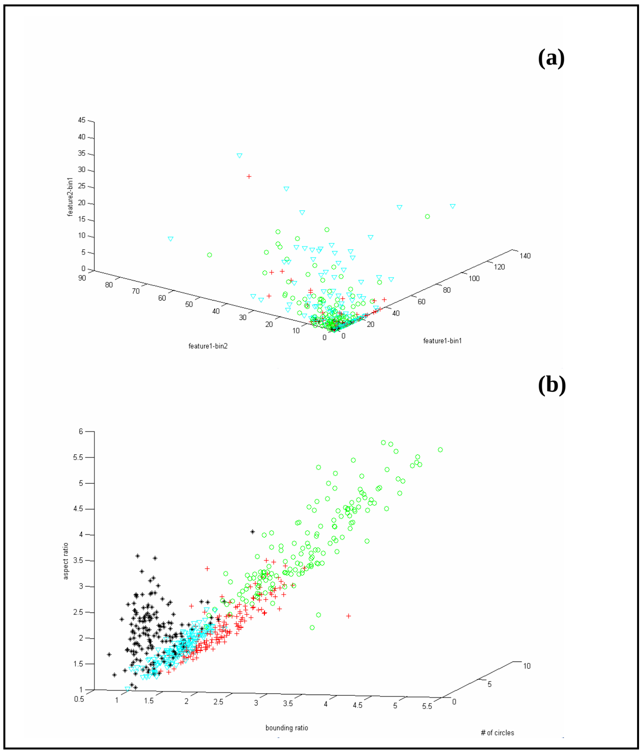

Figure 3. Visualization of the 600 instances by using the 3 best features (a) SIFT keypoint histograms and (b) statistics of geometric primitives.

the results across multiple classifiers for 600 images using 10-fold cross-validation for SIFT keypoint histograms and statistics of geometric primitives' features. Fig. 2 confirms that using a different evaluation criterion the highest performance is achieved again by our approach based on statistics of geometric primitives. The results obtained can be explained with the fact that in wide-area video surveillance, objects are often limited in size, and most often are low in texture and appear under different viewpoints. This results in a low number of detected SIFT keypoints and inconsistency of these keypoints across each class, leading to a lower classification performance compared to a classifier that is based on statistics of geometric primitives features.

In order to obtain a better understanding of the classification results we further analyze the classifier that provided the best classification accuracy and the lowest FPR, namely the MultiboostAB classification scheme for both approach 2 and approach 3. Using decision trees as the base learning algorithm, Multi-boosting has been demonstrated to produce decision committees with lower error than either AdaBoost or wagging. In our experiment MultiboostAB built 10 decision trees with different features and assigned them different weights. For statistics of geometric primitives the decision tree with the highest weight is based on the following 3 features as best features: aspect ratio, bounding box ratio and number of circles. For

SIFT keypoint histograms the decision tree with the highest weight is based on the following 3 features as best features: feature 2-bin 1, feature 1-bin 1, and feature 1-bin 2. In Fig. 3, we plot all 600 instances by using the aforementioned 3 features in different colors depending on their ground truth label. By looking at the figure, we are able to state that statistics of geometric primitives' features prove more discriminative by maximizing inter-class separation between the four classes of bag (s), person (s), trolley (s), and group (s).

Figures 4 and 5 illustrate how the 3 best features are able to discriminate between two specific classes, group vs. trolley and group vs. person, for the given 200 testing samples used in the holdout validation method. It should be noted that providing an insight for the SIFT keypoint histograms is not straightforward. However, by looking at Figures 4(a)-5(a), it is possible to state that the discriminative power of this feature set is rather limited. Instead, statistics of geometric primitives' features (see Figures 4(b)-5(b)) have sufficient discriminative power in order to provide a separation between classes.

When it comes to interpreting Figures 4(b)-5(b), in general, the bounding box ratio and the aspect ratio for trolleys are either similar to that of the group of people or lower depending on the shape of the group and this is clear in Fig. 4(b).

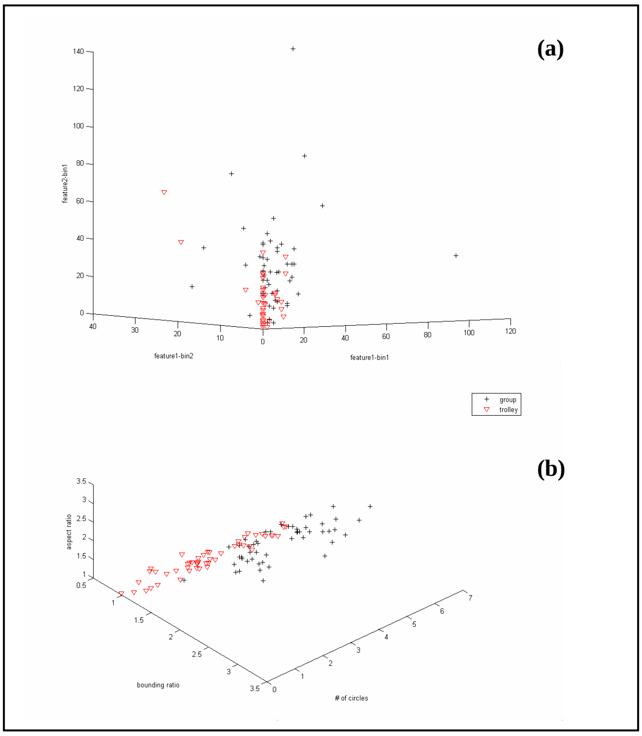

Figure 4. Visualization of how the 3 best features can discriminate between two specific classes, group (plus) vs. trolley (triangle down): (a) SIFT keypoint histograms and (b) statistics of geometric primitives.

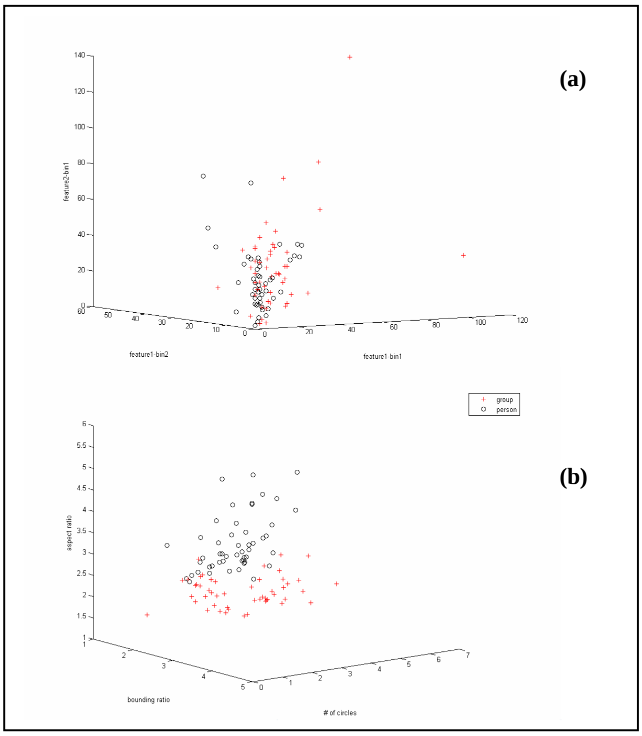

Figure 5. Visualization of how the 3 best features can discriminate between two specific classes, group (plus) vs. person (circle): (a) SIFT keypoint histograms and (b) statistics of geometric primitives.

In Fig. 5(b), it is obvious that a single person usually has higher aspect ratio compared to a group of people. Moreover, the number of circles for a group of people are either similar or higher compared to a person depending on the number of persons in each group and whether they occlude each other or not.

V. Conclusion

In this paper, we compared three different approaches for classification that use different techniques for feature extraction. Based on the experimental results obtained, we conclude that the results of our approach for classification based on statistics of geometric primitives outperforms the other two approaches that are based on SIFT keypoints using various classification and evaluation schemes. Classification based on statistics of geometric primitives with 10-fold cross-validation provides on average 22% higher recognition accuracy and 7% lower false alarm compared to the second best approach based on SIFT keypoint histograms. The illustrative analysis provided in this paper also demonstrates that statistics of geometric primitives maximize between-class separation and thus simplify the classification process.

The results of our approach are encouraging considering the challenges inherent to the intra-class shape variation, illumination changes, variable viewpoints, and clutter. We plan in the future to experiment with other feature reduction methods, possibly Kernel Principle Component Analysis (KPCA), to improve the classification performance even further.

ACKNOWLEDGMENTS

This research is supported by the Australian Research Council and iOmniscient Pty Ltd under the ARC Linkage Project Grant Scheme 2006 - LP0668325

REFERENCES

[1] B. Ayres and M. Boutell, “Home interior classification using SIFT keypoint histograms”, Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minnesota, USA, pp. 1-6, 2007.

[2] M.D. Beynon, D.J. Hook, M. Seibert, A. Peacock, D. Dudgeon, “Detecting abandoned packages in a multi-camera video surveillance system”, Proc. of the IEEE Advanced Video and Signal Based Surveillance Conference, Florida, USA, pp. 221-228, 2003.

[3] L.M. Brown, “View independent vehicle/person classification”, Proc. of ACM 2nd International Workshop on Video surveillance & Sensor Networks, New York, USA, pp. 114-123, 2004.

[4] T.G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms”, Neural Computation, vol. 10, no. 7, pp. 1895-1923, 1998.

[5] R. Fergus, P. Perona, A. Zisserman, “Object class recognition by unsupervised scale-invariant learning”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Wisconsin, USA , pp. 264-271, 2003.

[6] T. Kadir, A. Zisserman, and M. Brady, “An affine invariant salient region detector”, Proceedings of European Conference on Computer Vision, Prague, Czech Republic, pp. 228-241, 2004.

[7] D.G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[8] S. Lu, J. Zhang, D. Feng , “Detecting unattended packages through human activity recognition and object association”, Pattern Recognition, vol. 40, no. 8, pp. 2173-2184, 2007.

[9] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1615-1630, 2005.

[10] A.F. Otoom, H. Gunes, and M. Piccardi, “Towards automatic abandoned object classification in visual surveillance systems”, Proc. of Asia-Pacific Workshop on Visual Information Processing, Tainan, Taiwan, pp. 143-149, 2007.

[11] M. Tsuchiya, and H. Fujiyoshi, “Evaluating feature importance for object classification in visual surveillance”, Proc. of IEEE International Conference on Pattern Recognition, Hong Kong, pp. 978-981, 2006.

[12] G. Wang, Y. Zhang, and L. Fei-Fei, “Using dependent regions for object categorization in a generative framework”, Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, pp. 1597-1604, 2006.

[13] G.I. Webb, “MultiBoosting: a technique for combining boosting and wagging”, Machine Learning, vol. 40, no. 2, pp. 159-196, 2000.

[14] H. Witten and E. Frank, Data mining: Practical machine learning tools with java implementations, Morgan Kaufmann, San Francisco, CA, 2000.

[15] Non-Motion Detection Technology: http://www.iomniscient.com/products_iq140.htm (Access date: 14 March 2008).