2007

# A new image feature for fast detection of people in images

Son Lam Phung
*University of Wollongong*, phung@uow.edu.au

Abdesselam Bouzerdoum
*University of Wollongong*, bouzer@uow.edu.au

# A new image feature for fast detection of people in images


## Abstract

In this paper, we present a new method of detecting visual objects in digital images and video. The novelty of the proposed method is that it differentiates objects from non-objects using image edge characteristics. Our approach is based on a fast object detection method recently developed by Viola and Jones. While Viola and Jones use Harr-like features, we propose a new image feature called edge density that can be computed more efficiently. When applied to the problem of detecting people and pedestrians in images, the new feature shows very good discriminative capability compared to Harr-like features.



## Disciplines

Physical Sciences and Mathematics

## Publication Details

S. Phung & A. Bouzerdoum, "A new image feature for fast detection of people in images," International Journal of Information and Systems Sciences, vol. 3, (3) pp. 383-391, 2007.

INTERNATIONAL JOURNAL OF
INFORMATION AND SYSTEMS SCIENCES
Volume 3, Number 3, Pages 383–391



# A NEW IMAGE FEATURE FOR FAST DETECTION OF PEOPLE IN IMAGES

SON LAM PHUNG AND ABDESSELAM BOUZERDOUM

**Abstract.** In this paper, we present a new method of detecting visual objects in digital images and video. The novelty of the proposed method is that it differentiates objects from non-objects using image edge characteristics. Our approach is based on a fast object detection method recently developed by Viola and Jones. While Viola and Jones use Harr-like features, we propose a new image feature called edge density that can be computed more efficiently. When applied to the problem of detecting people and pedestrians in images, the new feature shows very good discriminative capability compared to Harr-like features.

**Key Words.** people detection, image edge analysis, object detection, video surveillance, pattern recognition.

## 1. Introduction

Detecting people and pedestrians in images and video has applications in video surveillance, road safety and many others. For example, Collins et al. [1] at CMU describe a multi-camera surveillance system that detects and tracks people over a wide area. Papageorgiou and Poggio [2] at MIT present a vision system that is used in Daimler-Chrysler Urban Traffic Assistant to detect pedestrians. Haritaoglu and Flickner [3] at IBM develop an intelligent billboard that uses a camera to detect and count the number of people in front of the billboard.

There are two major approaches to detecting people in images and video. The first approach finds people using heuristic visual cues such as motion, background scene or color. Using motion, the difference between consecutive video frames is calculated to identify image regions that contain moving objects [4]. Using background scene, a model is built to describe the statistical properties such as color, intensity, spatial and temporal variations of background pixels [3, 5, 6]; comparing this model and a new video frame will determine if a pixel belongs to the foreground or the background. The first approach can rapidly locate regions that likely contain people. However, these regions must be further processed using techniques such as face detection [4] or silhouette shape analysis [7]. Furthermore, this approach is of limited use when only a single input image is available.

The second approach scans the image window-by-window, a window is a fixed-size rectangular region of the image. Pattern classifiers are trained to determine if each window resembles the human body. This approach is computation-intensive but it can cope well with image variations. Papageorgiou and Poggio [2] proposed a pedestrian detection method that extracts Harr wavelet features from each 128-by-64 window and uses support vector machines to classify the features. Recently, Viola and Jones [8] developed a fast object detection method that relies on a cascade

Received by the editors October 3, 2006.

TABLE 1. People detection methods.

| **Author** | **Year** | **Is based on** |
|---|---|---|
| Papageorgiou and Poggio [2] | 1999 | Harr wavelets<br>support vector machines |
| Oliver et al. [9] | 2000 | eigen model of the background image |
| Haritaoglu et al. [7] | 2000 | model of background image<br>classification of shape features |
| Branca et al. [10] | 2002 | motion, Harr wavelets, 3-layer neural net |
| Rachlin et al. [6] | 2003 | color segmentation |
| Pantil et al. [4] | 2004 | motion, face detection |
| Yang et al. [11] | 2004 | depth, motion, color |
| Yoon and Kim [12] | 2004 | skin color, background subtraction,<br>Hausdorff-based shape comparison |
| Zang and Kodagoda [13] | 2005 | motion detection with laser range finder<br>edge-based template matching |
| Harasse and Bonnaud [5] | 2006 | statistical background model, skin color,<br>human model of head, skin and body regions |

of classifiers. Each classifier uses one or more Harr-like features and is trained using an adaptive boosting algorithm. Viola and Jones' method has been applied successfully to the face detection problem. A list of people detection methods is shown in Table 1.

This paper presents an object detection method that relies on object edge characteristics to differentiate objects and non-objects. We propose a new image feature called edge density that can be computed very fast, and apply it to detect people and pedestrians in images. This paper is organized as follows. Section 2 describes the proposed object detection method and the image feature. Section 3 focuses on an application of the proposed method in people detection and analyzes the discriminative power of the edge density feature. Section 4 is the conclusion.

## 2. Edge Density Approach

Our method is based on an object detection method that is proposed by Viola and Jones [8]. For a given input image, object regions are detected by scanning exhaustively windows of the image. Because there could be over $200,000$ windows in a typical image of size $640 \times 480$ pixels, a fast classification method is required to support real-time detection. Each window is processed by a cascade of strong classifiers to determine if it is an object or a non-object. If a strong classifier considers the window as a non-object, the window is immediately rejected; otherwise, the window is processed by the next strong classifier in the cascade. This means an object window must be processed by all strong classifiers, whereas a non-object window will be processed typically by a small number of strong classifiers. Because the majority of windows in an input image are non-object, the cascade structure reduces the average processing time per window.

A strong classifier is made up from one or more weak classifiers, and each weak classifier uses exactly one image feature extracted from the window. A strong classifier is so called because it has a lower error rate compared to a weak classifier. A strong classifier can be built from several weak classifiers using the AdaBoost algorithm [14]. The key idea of this algorithm is to force each weak classifier to focus on the training samples that the previous weak classifiers fail to process.

**2.1. New Image Feature based on Edge Density.** The system by Viola and Jones uses Harr-like feature that is defined as the difference in the pixel sums of two adjacent regions. If a Harr-like feature is greater than a threshold, the weak classifier considers the window as an object. Essentially, a salient Harr-like feature indicates a window as an object if region A appears significantly darker or brighter than region B, where regions A and B are to be found through training. This strategy works well for objects with a defined inner structure such as the human face. For example, it is a known fact that the eye region has a different brightness compared to its surrounding. However, for some objects such as the human body (in standing or walking pose) the dominant visual characteristics are the outer shape and edges. This observation motivates us to develop a new image feature that is based on edge density.

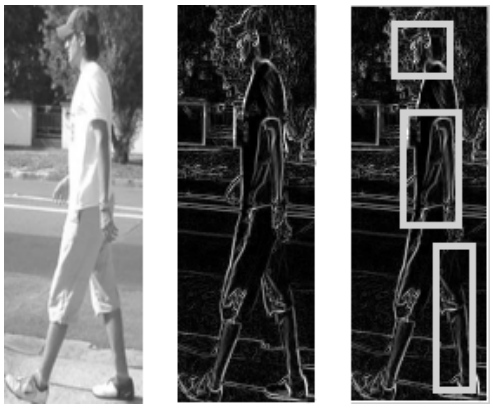


FIGURE 1. *Left*: an image window. *Middle*: the edge magnitude. *Right*: three edge density features where each feature is the average edge magnitude in a specific subregion.

For a given window, an edge density feature measures the average edge magnitudes in a subregion of the window (see Fig. 1). Let $i(x, y)$ be a window and $e(x, y)$ be the edge magnitude of the window. For a subregion $r$ with the left-top corner at $(x_1, y_1)$ and the right-bottom corner at $(x_2, y_2)$, the edge density feature is defined as

$$f = \frac{1}{a_r} \sum_{x=x_1}^{x_2} \sum_{y=y_1}^{y_2} e(x, y) \tag{1}$$

where $a_r$ is the region area, $a_r = (x_2 - x_1 + 1)(y_2 - y_1 + 1)$.

If the edge density feature is greater (or smaller) than a threshold, the weak classifier considers the window as an object. This is equivalent to saying that a strong (or weak) presence of image edges in a subregion will determine if the window is an object. In a window, there will be several thousands of subregions or features. The objective of system training is to identify the most salient subregions.

For the task of window scanning, there is a very efficient method to compute edge density features. Let $\mathbf{I} = \{I(x, y)\}$ be the input image of size $H \times W$. Let

$\mathbf{E} = \{E(x, y)\}$ be its edge magnitude; $E(x, y)$ is found by applying edge operators such as Sobel or Prewitt on the entire image [15]. The edge magnitude is a combination of the edge strengths along the horizontal and vertical directions:

$$E(x, y) = \sqrt{E_h^2(x, y) + E_v^2(x, y)} \tag{2}$$

From the edge magnitude image $\mathbf{E}$, we compute an edge integral image $\mathbf{S}$. The pixel value a location $(x, y)$ of $\mathbf{S}$ is defined as

$$S(x, y) = \sum_{x'=1}^{x} \sum_{y'=1}^{y} E(x', y') \tag{3}$$

That is, $S(x, y)$ is the sum of edge magnitudes in the rectangular region $\{(1, 1) - (x, y)\}$.

Given the edge integral image, the edge density feature of a subregion $r = \{(x_1, y_1), (x_2, y_2)\}$ can be computed using only a few arithmetic operations:

$$f = \frac{1}{a_r}\{S(x_2, y_2) + S(x_1 - 1, y_1 - 1) - S(x_2, y_1 - 1) - S(x_1 - 1, y_2)\} \tag{4}$$

Our approach requires computation of the edge magnitude image $\mathbf{E}$ before scanning occurs. Subsequently, each edge density feature involves only one subregion whereas each Harr-like feature involves at least two subregions. Hence, if the same number of features is used, the proposed approach can be expected to run faster compared to the Viola and Jones' system. In Section 3, we shall study the classification performance of the new image feature.

**2.2. Selecting the Most Salient Feature.** A weak classifier is built by selecting the best feature from a feature pool of several thousands. This section describes the feature selection technique.

In a given training set, let $w_1^+$, $w_2^+$,...,$w_M^+$ be the weights of $M$ training object patterns (i.e. positive patterns). Let $w_1^-$, $w_2^-$,...,$w_N^-$ be the weights of $N$ training non-object patterns (i.e. negative patterns). Let $w^+$ be the sum of all weights for object patterns, $w^+ = \sum_{i=1}^{M} w_i^+$. Let $w^-$ be the sum of all weights for non-object patterns, $w^- = \sum_{i=1}^{N} w_i^-$. During training, we can modify individual weights but the sum of $w^+$ and $w^-$ must be kept to 1.

Given an edge density feature $f$ that corresponds to a subregion $r$, we first compute the cumulative histograms $c^+(\theta)$ and $c^-(\theta)$ for the object and non-object patterns, taking into account pattern weights.

There are two possible decision rules: (1) object if $f > \theta$, and non-object otherwise; (2) object if $f \leq \theta$, and non-object otherwise. Here, $\theta$ is a threshold value. The error rate for the first decision rule is

$$e_1(\theta) = w^- + c^+(\theta) - c^-(\theta) \tag{5}$$

The error rate for the second decision rule is

$$e_2(\theta) = w^+ - c^+(\theta) + c^-(\theta) \tag{6}$$

Note that the sum of $e_1(\theta)$ and $e_2(\theta)$ is equal to 1. Among the two decision rules, we select the one that gives a smaller error, $e(\theta) = min[e_1(\theta), e_2(\theta)]$. The error rate using feature $f$ is the minimum value of $e(\theta)$ across the range of $\theta$. Finally, from the feature pool we choose the feature that gives the minimum error.

## 3. Experiments and Analysis

Having described the object detection approach, we now apply it to the problem of detecting people and pedestrians in images. The aim of this section is to study the process of building weak and strong classifiers, and the classification performance of the edge density feature.

**3.1. Experiment Data.** We collected a total of 622 images that contain people and pedestrians, and manually identified the coordinates of the people in these images. The images contain 817 people patterns. There are strong variations in the patterns: frontal view, side view, people in standing, bending, walking and running poses. Geometric transformations (image flipping and shifting) were applied to generate 2000 people patterns, of which 1000 patterns were used for training and 1000 patterns were used for testing.

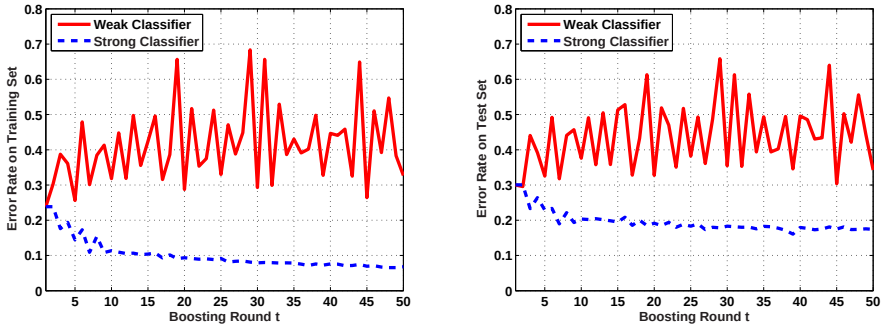
We also extracted 2000 non-people patterns from a set of landscape images, half of the non-people patterns were used for training and the other half for testing. Examples of the people and non-people patterns are shown in Fig. 2.

FIGURE 2. Examples of people and non-people patterns.

The average aspect ratio (height/width) of the people patterns in our dataset is $2.86 : 1$. Note that this aspect ratio covers children as well as people in running or striding pose. Based on this result, we selected a window size of $46 \times 16$ pixels for designing the classifiers. This window size is found to reduce the computation load while keeping sufficient visual details for classification.

**3.2. Analysis of Edge Density Features.** A strong classifier is trained in several rounds. In each round a weak classifier using exactly one edge density feature is formed. The weights of training patterns are modified according to the AdaBoost algorithm [14] to put emphasis on the patterns that the previous weak classifier incorrectly handles.

We trained a strong classifier for 50 iterations. The edge operator used is the difference operator. Figure 3a shows the error rates of the strong classifier and weak classifiers as training progresses. The results show that the training error of the strong classifier decreases steadily with respect to the number of the training rounds. However, the error rates of individual weak classifiers fluctuate with an




(a) (b)

FIGURE 3. Error rates of weak classifiers and a strong classifier on a) the training set, and b) the test set.

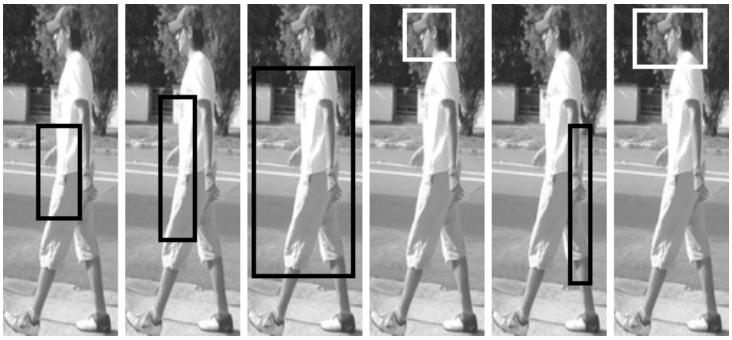


FIGURE 4. Examples of selected edge density features: at boosting round 1, 3, 5, 7, 13 and 23.

upward trend. This trend is explained by the fact that each new weak classifier in essence focuses on a small subset of the training set; this subset contains "difficult" patterns that previous weak classifiers cannot handle. After 30 training rounds, the strong classifier has an error rate of 0.079.

Some edge density features selected by the strong classifier are shown in Fig. 4. These features indicate that the strong classifier mostly picks up the edge difference between the human body and the surrounding. The feature selected at round 7 reflects the fact that there are strong edges in the human head region.

The performances of the strong classifier and individual weak classifiers on the test set are shown in Fig. 3b. The results show that even though the error rate of each weak classifier is high, the error rate of the strong classifier decreases steadily. In this case, there is little change in the error rate of the strong classifier after round 10. Using a validation set, we can detect when this occurs and stop training the strong classifier. At this point, we usually collect more data for training the next strong classifier and add it to the cascade.

Using the threshold found by the AdaBoost algorithm, the strong classifier with 10 features has an error rate of 0.2035, a false positive rate of 0.1570, and a false negative rate of 0.2500. The strong classifier with 50 features has an error rate of 0.1740, a false positive rate of 0.1420, and a false negative rate of 0.2060.

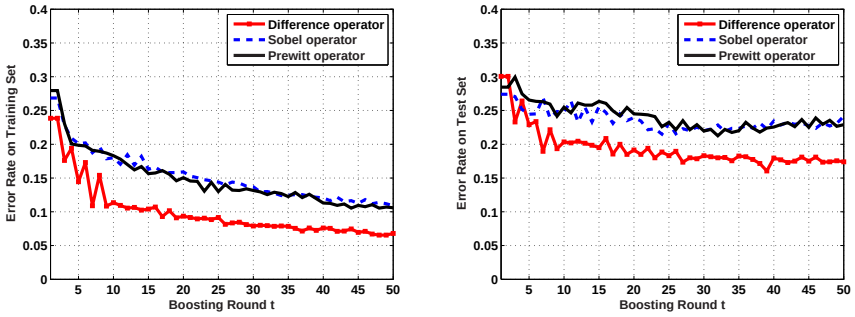
By reducing the threshold of a strong classifier, we can reduce its false negative rate to, say, $F_n = 0.0001$ at the cost of an increased false positive rate to $F_p$. If we put $n$ strong classifiers in series, the (expected) overall false negative rate will become $1 - (1 - F_n)^n$ whereas the overall false positive rate is $F_p^n$. Clearly, a large $n$ will give both a low false negative rate and a low false positive rate.

**3.3. Comparison of Edge Operators.** In this section, we compare the performance of different edge operators. Three edge operators were examined: the difference operator, the Sobel operator and the Prewitt operator. The convolution masks of these operators are shown in Table 2. The edge strengths along the horizontal and vertical directions are computed as

$$\mathbf{E}_h = \mathbf{I} \otimes h_h \text{ and } \mathbf{E}_v = \mathbf{I} \otimes h_v \tag{7}$$

TABLE 2. Edge operators used for feature extraction.

| Operator | Horizontal Mask $h_h$ | Vertical Mask $h_v$ |
|---|---|---|
| Difference | $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ | $\begin{bmatrix} 1 & -1 \end{bmatrix}$ |
| Sobel | $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ |
| Prewitt | $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$ | $\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$ |




(a) (b)

FIGURE 5. Error rates of strong classifiers using three edge operators on a) the training set, b) the test set.

The performances of the two strong classifiers that use different edge operators to extract the edge density features are shown in Fig. 5. This figure shows that compared with the other operators, the difference operator leads to faster training and a lower error rate on the test set. After 50 training rounds, the error rates on the test set for the difference, Sobel and Prewitt operators are 0.1740, 0.2415 and

0.2290, respectively. Note that the difference operator has a smaller size and hence can be applied very fast on the entire input image.

**3.4. Comparison of Edge Density and Harr-like Features.** For comparison purposes, we trained two strong classifiers: one using only edge density features, and the other using only Harr-like features [8]. The two types of image features are illustrated in Fig. 6. A Harr-like feature is the difference in the intensity sums of two adjacent rectangles. In comparison, an edge density feature is the average edge magnitude in a region.

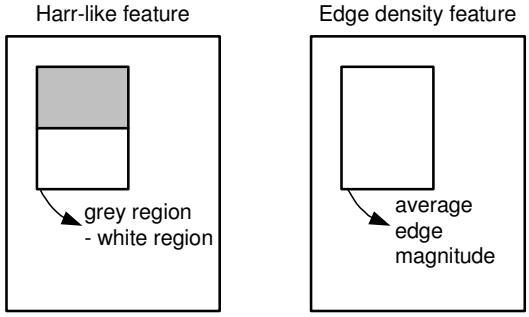
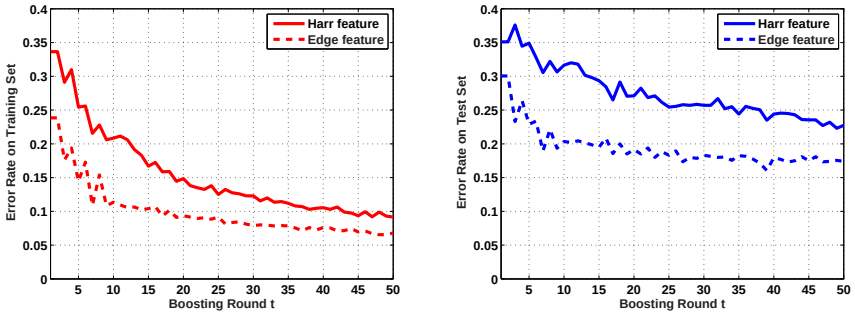



FIGURE 6. Harr-like features and edge density features.




(a) (b)

FIGURE 7. Error rates of strong classifiers that use Harr-like features and edge density features on (a) the training set, (b) the test set.

The performances of the two strong classifiers on the training set and the test set are shown in Fig. 7. The figure shows that the training error reduces faster using edge intensity features. For example, after 10 rounds the training error is 0.1135 for edge density feature, and 0.2085 for Harr feature. Furthermore, the test error is lower for the strong classifier that uses edge density features. After 50 training rounds the best test error is 0.1605 for edge density feature, and 0.2230 for Harr feature.

The above results for the people detection task demonstrate a clear improvement of the proposed image feature. We plan to study next the comparative performance of the full people detector and extend our approach to directional image features.

## 4. Conclusion

A new method for detecting objects in images that relies on object edge characteristics is presented. We propose a new image feature called edge density that can be computed very efficiently. The edge density feature is found to have better discriminative capability compared to the Harr-like feature for the task of detecting people in images. The difference operator is found to outperform the Sobel and Prewitt operators in terms of speed and classification accuracy.

## Acknowledgments

The authors thank Mr Markos Pratsas for taking part in data collection. This work is supported by the University of Wollongong Small Research Grant.

## References

[1] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1456–1477, 2001.

[2] C. Papageorgiou and T. Poggio, "Trainable pedestrian detection," in *International Conference on Image Processing*, 1999, vol. 4, pp. 35–39 vol.4.

[3] I. Haritaoglu and M. Flickner, "Attentive billboards," in *International Conference on Image Analysis and Processing*, 2001, pp. 162–167.

[4] R. Patil, P.E. Rybski, T. Kanade, and M.M. Veloso, "People detection and tracking in high resolution panoramic video mosaic," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, vol. 2, pp. 1323–1328 vol.2.

[5] S. Harasse, L. Bonnaud, and M. Desvignes, "Human model for people detection in dynamic scenes," in *International Conference on Pattern Recognition*, 2006, vol. 1, pp. 335–354.

[6] Y. Rachlin, J. Dolan, and P. Khosla, "Learning to detect partially labeled people," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, vol. 2, pp. 1536–1541.

[7] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[8] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[9] Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[10] A. Branca, M. Leo, G. Attolico, and A. Distante, "People detection in dynamic images," in *International Joint Conference on Neural Networks*, 2002, vol. 3, pp. 2428–2432.

[11] Mau-Tsuen Yang, Ya-Chun Shih, and Shih-Chun Wang, "People tracking by integrating multiple features," in *International Conference on Pattern Recognition*, 2004, vol. 4, pp. 929–932.

[12] Sang Min Yoon and Hyunwoo Kim, "Real-time multiple people detection using skin color, motion and appearance information," in *13th IEEE International Workshop on Robot and Human Interactive Communication*, 2004, pp. 331–334.

[13] Zhengzhi Zhang and K.R.S. Kodagoda, "Multi-sensor approach for people detection," in *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2005, pp. 355–360.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1995.

[15] Rafael C. Gonzalez and Richard E. Woods, *Digital image processing*, Prentice Hall, New York, 2002.

School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Northfields Av, Wollongong, NSW 2522, Australia.

*E-mail*: phung@uow.edu.au and a.bouzerdoum@ieee.org

*URL*: http://www.elec.uow.edu.au/staff/sphung/