*Special issue: 3D image and video technology*

Research Paper

# People detection based on co-occurrence of appearance and spatio-temporal features

Yuji YAMAUCHI[1], Hironobu FUJIYOSHI[2], Yuji IWAHORI[3], and Takeo KANADE[4]
[1,2,3]*Graduate School of Engineering, Chubu University*
[4]*Robotics Institute, Carnegie Mellon University*

**ABSTRACT**

**This paper presents a method for detecting people based on co-occurrence of appearance and spatio-temporal features. In our method, Histograms of Oriented Gradients (HOG) are used as appearance features, and the results of pixel state analysis are used as spatio-temporal features. The pixel state analysis classifies foreground pixels as either stationary or transient. The appearance and spatio-temporal features are projected into subspaces in order to reduce the dimension of feature vectors by principal component analysis (PCA). The cascade AdaBoost classifier is used to represent the co-occurrence of the appearance and spatio-temporal features. The use of feature co-occurrence, which captures the similarity of appearance, motion, and spatial information within the people class, makes it possible to construct an effective detector. Experimental results show that the performance of our method is about 29.0% better than that of the conventional method.**

**KEYWORDS**

People detection, histograms of oriented gradients, Pixel State Analysis, co-occurrence, AdaBoost

## 1 Introduction

Automatic people detection is a key enabler for applications in robotics, visual surveillance, human computer interactions, and intelligent transport systems. In the visual surveillance, fixed cameras are generally used to reduce costs. This has led to the development of a number of methods [1]–[3] based on background subtraction for detecting motion from images captured by fixed camera. One of the successful approach to model the background uses a Gaussian mixture model (GMM) [2]. Since methods based on background subtraction use a top-down approach, object classification at the next step becomes impossible if the object's region is not segmented correctly. A window-scanning approach has been proposed for solving this problem. It was made possible by the great improvements in computer speed in recent years. Recently, several appearance features [4]–[6] for detecting people were proposed. In particular, Dalal *et al.* [5] presented a people detection algorithm that has an excellent detection ability. Each detection window is divided into cells of size $8 \times 8$ pixels, and each group of $2 \times 2$ cells is integrated into a block in a sliding fashion, so the blocks overlap. Each cell consists of a 9-bin Histogram of Oriented Gradients (HOG), and each block contains a concatenated vector of all its cells. This representation has been proven to be powerful enough to classify people using a linear Support Vector Machine (SVM). Because HOG features are invariant to illumination and local geometrical changes, many recent studies have used them to detect people [7], [8].

Several people detection methods using appearance and motion features have been proposed to improve detection accuracy [9], [10]. Viola et al. proposed a method for detecting people using patterns of motion and appearance obtained by Haar-like features [10]. Dalal et al. proposed a method with improved detection accuracy that uses HOG and motion features calculated from the optical flow [9]. The availability of motion information makes it possible to improve the detection

Received September 19, 2009; Revised January 12, 2010; Accepted January 13, 2010.
[1)] *yuu@vision.cs.chubu.ac.jp*, [2)] *hf@cs.chubu.ac.jp*, [3)] *iwahori@cs.chubu.ac.jp*, [4)] *tk@cs.cmu.edu*
DOI: 10.2201/NiiPi.2010.7.5

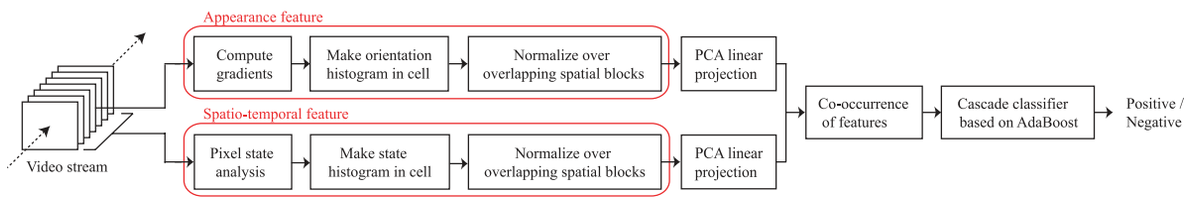



Fig. 1 Flow of proposed method.

performance. One problem with these methods based on appearance and motion is that any features besides appearance cannot be obtained when the object is stationary, such as when people are standing still.

We have investigated the problem of detecting people using images captured by a fixed camera and propose a method for people detection based on the co-occurrence of appearance and spatio-temporal features. Histograms of Oriented Gradients(HOG) are used as appearance features, and the results of pixel state analysis are used as spatio-temporal features. The Pixel State Analysis (PSA) classifies foreground pixels as either stationary or transient. The appearance and spatio-temporal features are projected into subspaces in order to reduce the dimensions of the vectors by Principal Component Analysis (PCA). The cascade AdaBoost classifier is used to represent the co-occurrence of the appearance and spatio-temporal features. The use of spatio-temporal features reduce the number of false object detections, *i.e.*, detection of objects that appear similar to people. Moreover, co-occurrence of appearance and spatio-temporal features makes the detection more effective.

## 2 Overview of proposed method

Our method for using the co-occurrence of appearance and spatio-temporal features is diagramed in Fig. 1. The HOG feature vectors for the detecting window are computed as feature based on appearance. They are projected into subspaces to reduce their dimensions by PCA. PSA is used to compute the spatio-temporal features, and these feature vectors are also projected into subspaces to reduce their dimensions by PCA in the same way as for the HOG features. Both the appearance and spatio-temporal features are combined into one feature by representing their co-occurrence using the joint probability acquired from the combination of each of the features. Finally, the cascade AdaBoost classifier, which is trained in advance, decides whether or not the objects are people. Feature co-occurrence, which captures the similarity of appearance, motion, and spatial information within the people class, makes it possible to construct an effective detector.

## 3 Feature extraction

This section describes the feature extraction and how the co-occurrence between appearance and spatio-temporal features is represented.

### 3.1 Histograms of oriented gradients

HOG representation [5] has several advantages. It captures the gradient structure that is characteristic of the human shape. First, magnitude $m$ and orientation $\theta$ of the gradients are computed using

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2}, \tag{1}$$

$$\theta(x,y) = \tan^{-1}\frac{f_y(x,y)}{f_x(x,y)}, \tag{2}$$

where $f_x(x,y) = L(x+1,y) - L(x-1,y)$, $f_y(x,y) = L(x,y+1) - L(x,y-1)$, and $L(x,y)$ is the brightness of pixel. Each detection window is divided into cells of size $5 \times 5$ pixels and each group of $3 \times 3$ cells is integrated into a block in a sliding fashion, as shown in Fig. 2, so that the blocks overlap with each other. Each cell consists of a 9-bin histogram of HOG features represented by $\boldsymbol{F}_{ij} = [f_1, f_2, \cdots, f_9]$. Each block contains a concatenated vector of all its cells. Each block is thus represented by $\boldsymbol{V}_k = [\boldsymbol{F}_{ij}, \boldsymbol{F}_{i+1j}, \boldsymbol{F}_{i+2j}, \boldsymbol{F}_{ij+1}, \boldsymbol{F}_{i+1j+1}, \boldsymbol{F}_{i+2j+1}, \boldsymbol{F}_{ij+2}, \boldsymbol{F}_{i+1j+2}, \boldsymbol{F}_{i+2j+2}]$. The feature of one block of the $k$ th block represents 81 feature vectors that are normalized to an L2-norm using the following equation, which is the sum of the Euclidean distance in the block.

$$v = \frac{f}{\sqrt{\|\boldsymbol{V}\|_2^2 + \epsilon^2}} \qquad (\epsilon = 1). \tag{3}$$

Each detection window ($30 \times 60$ pixels) is represented by $4 \times 10$ blocks, giving a total of $3,240$ features per detection window.

### 3.2 Pixel state analysis

Objects similar to human are done false detection when only appearance feature is used. Therefore, we use feature vectors obtained from the result of pixel state analysis (PSA) [11] that represent object motion

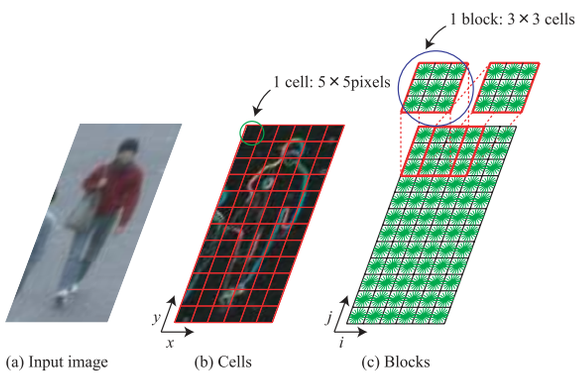



Fig. 2 Cells and blocks.

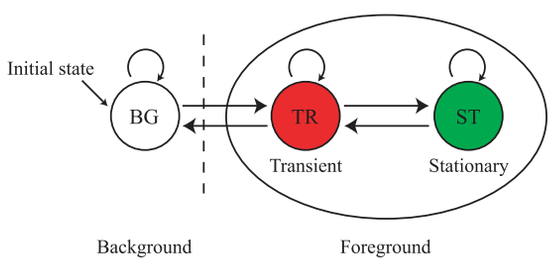



Fig. 3 Diagram of state transition for a pixel.

```
if((M = st OR M = bg) AND (T > th_t)){
    M = tr
}
if((M = tr) AND (S < th_s)){
    if(I = background intensity){
        M = bg
    }else{
        M = st
    }
}
```

Fig. 4 Algorithm for pixel state analysis.

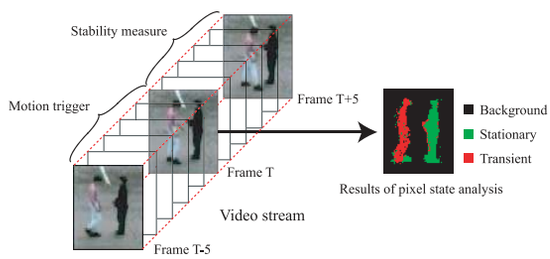



Fig. 5 Examples of pixel state analysis.

and spatial information. This analysis is used to determine whether a pixel is stationary or transient by observing its intensity value over time, as shown in Fig. 3. To capture the nature of changes in pixel intensity profiles, two factors are important: the existence of a significant step change in intensity, and the intensity value to which the profile stabilizes after passing through a period of instability.

Let $I_t$ be some pixel's intensity at a time $t$ occurring $k$ frames in the past. Two functions are computed: a motion trigger $T$ just prior to the frame of interest $t$, and a stability measure $S$ computed over $k$ frames from time $t$ to the present. The motion trigger $T$ is simply the maximum absolute difference between the pixel's intensity $I_t$ and its value in the previous five frames:

$$T = \max\{|I_t - I_{(t-j)}|, \forall_j \in [1,5]\}. \quad (4)$$

The stability measure $S$ [1)] is the variance of the intensity profile from time $t$ to the present:

$$S = \frac{K\sum_{j=1}^{K} I_{(t+j)}^2 - \left(\sum_{j=1}^{K} I_{(t+j)}\right)^2}{K(K-1)}. \quad (5)$$

[1)] Development of variance equation was described in appendix.

Transient map $M$ is defined by the algorithm below (Fig. 4) for each pixel, using three possible values : background = (bg); transient = (tr) and stationary = (st). The background intensity is prepared in advance as a background image. The background is updated by an Infinite Impulse Response(running average) filter to accommodate slow lighting changes and noise in the imagery [1].

Fig. 5 shows an example of the pixel state analysis. We see that most pixels of people walking on the left are transient pixels and that most pixels on the right are stationary ones because the people have stopped. Thus, spatio-temporal feature is extracted from the result of PSA.

For each detection window, a 3-bin histogram (background/stationary/transient) is computed by counting the number of each state. This histogram is normalized by the same procedure of HOG. If the detection window is $30 \times 60$ pixels, the dimension of PSA feature is 1,080 feature vectors.

### 3.3 Principal component analysis

Appearance features (3,240 dimensions) and spatio-temporal features (1,080 dimensions) have very high dimensionality. Because the histogram for a cell is used many times for the normalization, the correlation between feature vectors is strong. We need to identify a valid subspace in order to obtain a compact and low-dimensional representation of the feature vectors. The

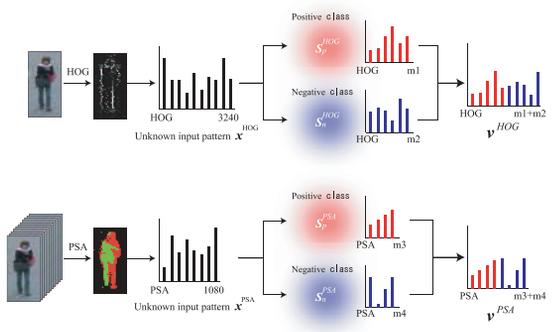



Fig. 6 Projection into subspaces.

PCA is thus performed on a training database. Given $N$-dimensional feature vectors, $\boldsymbol{x}_n = (x_1, x_2, \cdots, x_N)$, mean vector $\boldsymbol{M}$ and covariance matrix $\boldsymbol{CV}$ are given by

$$\boldsymbol{CV} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{M})(\boldsymbol{x}_n - \boldsymbol{M}), \tag{6}$$

and

$$\boldsymbol{M} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n. \tag{7}$$

The principal components assume the first $\boldsymbol{P}$ significant eigenvectors of $\boldsymbol{C}$V; that is, $\boldsymbol{v} = (v_1, v_2, \cdots, v_P)$. Construction of eigen matrix $\boldsymbol{U}_n = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_P)$ with a $d \times P$ dimension enables an arbitrary $N$-dimensional original feature vector $\boldsymbol{x}$ to be represented as a new $P$-dimensional compact vector. The projection matrix for the positive class and negative class of each feature is computed using training images. Number of the dimensions to reduce is decided by cumulative contribution rate. In this paper, we use the dimension where the cumulative contribution rate will be 99%.

Next, the vectors are projected into subspaces, and the features are extracted. Here we consider four projection matrixes, $\boldsymbol{CV}_p^{HOG}, \boldsymbol{CV}_n^{HOG}, \boldsymbol{CV}_p^{PSA}$ and $\boldsymbol{CV}_n^{PSA}$, as shown in Fig. 6.

For each detection window, HOG features $\boldsymbol{x}^{HOG}$ and PSA features $\boldsymbol{x}^{PSA}$ are computed. Next, $\boldsymbol{x}^{HOG}$ and $\boldsymbol{x}^{PSA}$ are projected using the projection matrices, $\boldsymbol{CV}_p^{HOG}, \boldsymbol{CV}_n^{HOG}, \boldsymbol{CV}_p^{PSA}, and\ \boldsymbol{CV}_n^{PSA}$, for the people class (positive class) and the non-people class (negative class). Finally, the features for the classifier are computed using

$$\begin{aligned}
[v_1^{HOG}, \cdots, v_{m1}^{HOG}] &= \boldsymbol{x}^{HOG^T}\boldsymbol{CV}_p^{HOG}, \\
[v_{m1+1}^{HOG}, \cdots, v_{m1+m2}^{HOG}] &= \boldsymbol{x}^{HOG^T}\boldsymbol{CV}_n^{HOG}, \\
[v_1^{PSA}, \cdots, v_{m3}^{PSA}] &= \boldsymbol{x}^{PSA^T}\boldsymbol{CV}_p^{PSA}, \\
[v_{m3+1}^{PSA}, \cdots, v_{m3+m4}^{PSA}] &= \boldsymbol{x}^{PSA^T}\boldsymbol{CV}_n^{PSA}.
\end{aligned} \tag{8}$$

### 3.4 Co-occurrence of features

The features that are newly obtained from PCA are expressed as the co-occurrence of appearance and spatio-temporal features. Our method uses the representation method proposed by Mita et al. [12] to express the co-occurrence between different kinds of features. To improve the generalization performance, we use the weak classifiers that observe multiple features. Feature co-occurrence makes it possible to classify difficult examples that are misclassified by the weak classifiers using a single feature. We represent the statistics of feature co-occurrence using their joint probability. To calculate the joint probability, we binarize the feature value $v_i$. As a result, each feature value is represented by a binary variable, $s$, which takes 1 or 0, specifying people or non-people respectively. The variable $s$ for an example $v_i$ is calculated using

$$s = \begin{cases} 1 & P(C_p|v_i) > P(C_n|v_i) \\ 0 & otherwise \end{cases}, \tag{9}$$

where $s$ is classified by Bayes theorem:

$$P(C_k|v_i) = \frac{P(v_i|C_k)P(C_k)}{P(v_i)} \quad (k = p, n), \tag{10}$$

where $P(v_i|C_k)$ is the probability obtained from the probability density function, $P(C_k)$ is the prior probability, $P(v_i)$ is the existence probability, and $C$ is the class. It is assumed that $P(C_k)$ is equal to 0.5. The probability density function, $P(v_i|C_k)$, is approximated using a smoothed 1D histogram of the $i^{th}$ subspace coefficients that were obtained from the training images.

Feature $c$ is represented by combining the binary variables computed from appearance feature $\boldsymbol{v}^{HOG}$ and spatio-temporal feature $\boldsymbol{v}^{PSA}$ described in 3.3. Feature $c$ representing the co-occurrence between the appearance and spatio-temporal features is described by feature $s$ using each feature one by one. As a result, the feature is described by a value of a total of four patterns. For example, if the appearance feature is 1, and if the spatio-temporal feature is 0, feature $c$ is computed by

$$C(x) = (10)_2 = 2. \tag{11}$$

When $C(x) = c$, the cascade AdaBoost classifier is trained, computed feature $c$ is selected automatically as if the error is minimum.

## 4 Construction of classifier

This section describes the construction of the classifier for people detection.

### 4.1 Cascade AdaBoost

The final strong classifier, $H(x)$, is a linear combination of $L$ weak classifiers, $h_l(x)$:

$$H(x) = \mathrm{sign}\left(\sum_{l=1}^{L} \alpha_l h_l(x)\right), \tag{12}$$

where $x$ is the input data, $\alpha_l$ is the weight of the training data, and $l$ is number of round.

We use superior cascade classifier [13] of calculation efficiency. False detection rate is restrained without decreasing detection rate by constructing classifier to cascade. For each level of the cascade, we construct a strong classifier consisting of several weak classifiers. In each level of the cascade, we keep adding weak classifiers until the predefined quality requirements are met. In our case we require the minimum detection rate to be 0.9975 and the maximum false positive to be 0.3 in each stage.

### 4.2 Weak classifier

A weak classifier, $h_l(x)$, is described in the following equation for the discriminate function based on the conditional probability.

$$h_l(x) \begin{cases} +1 & P_l(y=+1|c) > P_l(y=-1|c) \\ -1 & \text{otherwise,} \end{cases} \tag{13}$$

where $P_l(y=+1|c)$ and $P_l(y=-1|c)$ are the joint probabilities of feature co-occurrence represented by feature $c$ and class label $y_i \in \{+1, -1\}$. Joint probabilities are computed by the following equations,

$$P_l(y=+1|c) = \sum_{i:C_l(x_i)=c \wedge y_i=+1} D_l(i), \tag{14}$$

$$P_l(y=-1|c) = \sum_{i:C_l(x_i)=c \wedge y_i=-1} D_l(i), \tag{15}$$

where $C_l(x)$ is the function used to observe feature $c$, which is used to represent the co-occurrence between feature vectors, and $D_l(i)$ is the weight of training data. An example of the probabilities $P_l(y=+1|c)$ and $P_l(y=-1|c)$ obtained from HOG and PSA features are shown in Fig. 7. The two features yield four combinations of binary variables, which are from $(00)_2$ to $(11)_2$. If $c = (01)_2$ or $(11)_2$ is measured from an input data, it is classified as positive class.

## 5 Experimental results

Proposed method is evaluated through two comparative experiments of people classification using test images.

1. Effectiveness of using co-occurrence of appearance and spatio-temporal features comparing to the conventional method [5].
2. Effectiveness of using spatio-temporal feature obtained by PSA comparing to general motion detection such as the background subtraction and the temporal differencing.

The results were evaluated by the detection error tradeoff (DET) curve expressed using a double logarithmic chart. The horizontal axis represents the false positive rate, and the vertical axis represents the false negative rate.

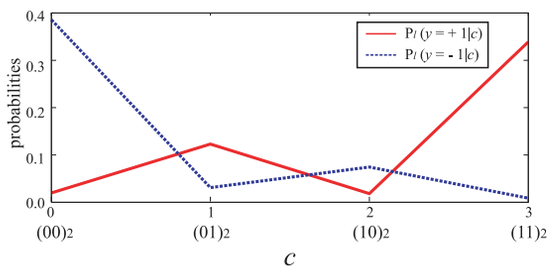



Fig. 7 Joint probability of each class.

Table 1 Reduction of dimensionality of feature vectors by PCA.

| Feature | Before reduction | After reduction |
|---|---|---|
| HOG(Pos.) | 3,240 | 678 |
| HOG(Neg.) | 3,240 | 1,231 |
| PSA(Pos.) | 1,080 | 124 |
| PSA(Neg.) | 1,080 | 109 |

### 5.1 Database

We collected nine video sequences of street scenes. Each sequence consists of 2,000 to 10,000 frames. We used five of the sequences for training. The other four were used for testing. The training data consist of 2,053 positive images and 6,253 negative images, and the test data consist of 1,023 positive images and 1,233 negative images. Fig. 8 shows some examples of each feature in the training data.

### 5.2 Dimensionality reduction of vectors by PCA

The feature vectors was reduced by PCA. Table 1 shows the extent of the reduction for each feature and class.

The reduction for HOG (negative class) was lower than that for the positive one. This is why the negative class was more complex than the positive class. The reduction for PSA (negative class) was larger than that for the positive one due to the background clutter. The variance in the negative data for the training was thus lower because there were mostly static objects in the

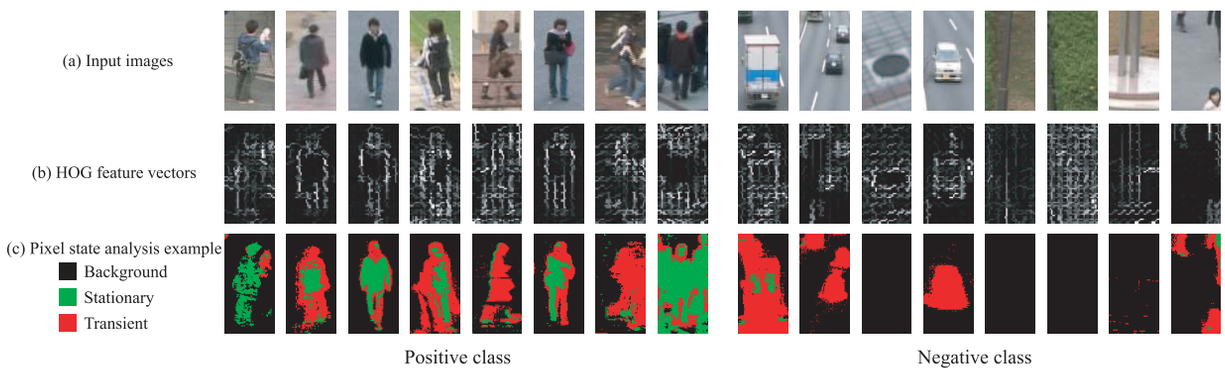



Fig. 8 Some examples of HOG and PSA in the training data.

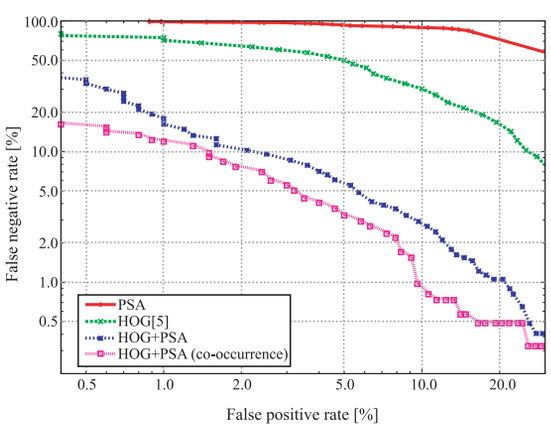



Fig. 9 DET of experiment 1.

training data for the negative class.

### 5.3 Experiment 1: Effectiveness of using co-occurrence

We compare four feature pattern methods, "HOG" (the conventional method) [5], "PSA", "HOG + PSA", and "co-occurrence of HOG + PSA" (our method). As shown in Fig. 9, our method had better accuracy than the "HOG" method. With a false positive rate of 10%, "co-occurrence of HOG + PSA" had a 29.3% lower false negative rate. Compared to the "HOG + PSA" method, our method had 2.8% better detection performance. The conventional "HOG" method, which uses only appearance features, is more likely to falsely detect objects similar in appearance to people and objects with a complex texture. Our method had a lower false detection rate due to the use of co-occurrence of appearance and spatio-temporal features.

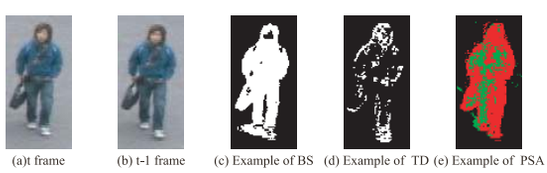



Fig. 10 Examples of background subtraction, temporal differencing and pixel state analysis.

### 5.4 Experiment 2: Effectiveness of using spatio-temporal feature

We compare four pattern methods of features, "HOG" , representing appearance only, "HOG + BS" (BS means background subtraction) representing appearance and spatial information, "HOG + TD" (TD means temporal differencing) representing appearance and motion information, and "HOG + PSA" representing appearance and spatio-temporal information. "Co-occurrence of HOG+PSA" gave the best performance in experiment 1. Therefore, except for "HOG" represents the co-occurrence. Fig. 10 shows examples of BS, TD and PSA.

Background subtraction detects the whole object region (Fig. 10 (c)), while the temporal differencing detects moving regions of the object (Fig. 10 (d)). Both extract the moving object region of cell as with PSA as feature vector.

As shown in Fig. 11, the "HOG+BS" method had a false positive rate of 10.0%, 28.7% lower than that of the "HOG" method. The "HOG + TD" had a 27.0% lower rate than the "HOG" method. These results indicate that using spatial information obtained by BS and using motion information obtained by TD are effective for the people detection.

Next, we verity the difference of quality of BS and TD. To determine the difference in quality between using BS and using TD, we defined a set of positive-class

test images (1,023 images) as universal set $S$. The images correctly classified by the "HOG + PSA" method were defined as subset $P$. Those correctly classified by the "HOG + BS" method were defined as subset $B$. Those correctly classified by the "HOG + TD" method were defined as subset $T$. We then calculated the non-common class ratio:

$$R_{B-T} = 1 - \frac{\mathrm{N}(B \cap T)}{\mathrm{N}(B \cup T)} = 1 - \frac{642}{865} = 0.26, \qquad (16)$$

where $N(X)$ represents the number of elements $x$ constituting set $X$. The ratios for $B$ and $T$ were both about 26% (223 images). This means that classification using either the "HOG + BS" or "HOG + TD" method worked well. Therefore, the feature obtained using BS (spatial information) and TD (motion information) represent the different quality feature.

To determine the effectiveness of the "HOG + PSA" method, we defined the union of sets $B$ and $T$ as $C(B \cup T)$. The non-common class ratios for $P$ and $C$, calculated using

$$R_{P-C} = 1 - \frac{\mathrm{N}((P \cap C) \cup P)}{\mathrm{N}(P \cup C)} = 1 - \frac{961}{974} = 0.02 \qquad (17)$$

were about 2% (13 images). This means that successful images by the "HOG + BS" and "HOG + TD" methods are also approximately possible by using the "HOG + PSA" method. It seems reasonable to think that the spatio-temporal feature obtained by PSA includes both spatial and motion information. Furthermore, as shown in Fig. 11, the performance of the "HOG + PSA" method was better than those of the "HOG + BS" and "HOG + TD" methods. The use of spatio-temporal feature is thus effective for detecting people either walking or standing.

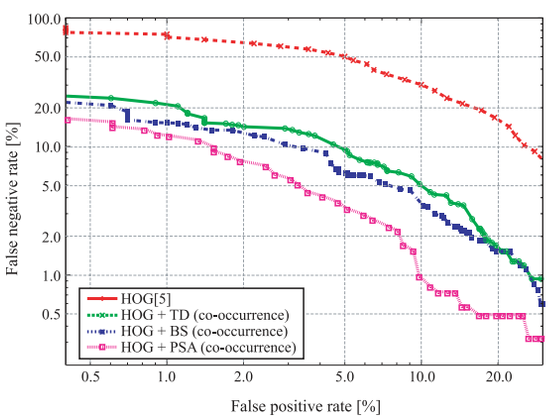
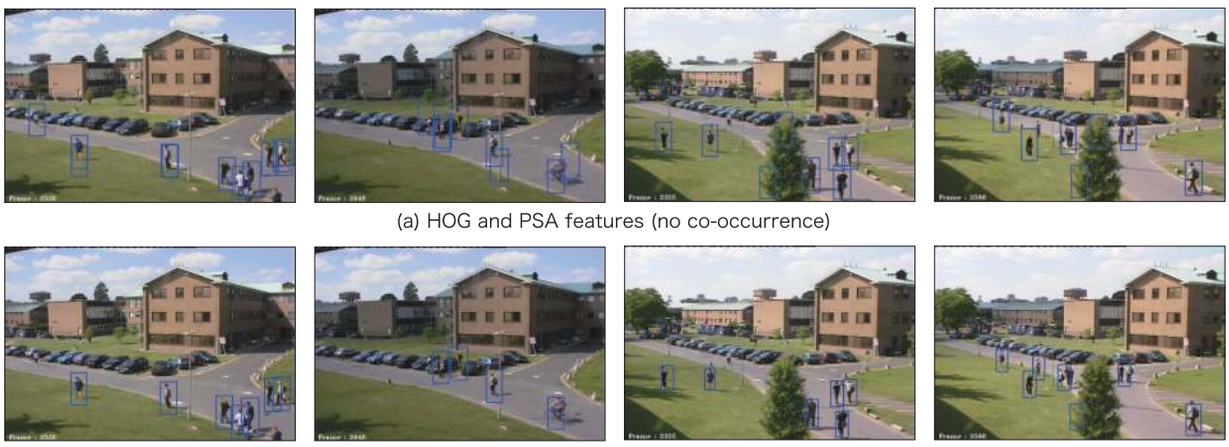



Fig. 11 DET of experiment 2.

### 5.5 People detection examples

When detecting people from an image, a raster scan of the detection window is performed over and over as the scale of detection window is changed. Therefore, our method can accommodate for a change of size of human as in other human detection methods. This resulted for about 10,000 detected windows in each im-


Fig. 12 Examples of people detection.



(a) HOG and PSA features (no co-occurrence)

(b) Co-occurrence of HOG and PSA features

Fig. 13 Detection examples on PETS2001 dataset.

age, and each one possibility contained images of people. Mean shift clustering [14] was used to make the final decision on people detection by placing a box around each detected human.

Some example detections are shown in Fig. 12 and Fig. 13. Fig. 12 shows the comparison examples of people detection method using videos taken in multiple locations. Our method clearly had better detection accuracy, even when people overlapped with the complex backgrounds. The conventional method had several false detections due to the complicated background and objects with a shape resembling that of a human. The HOG method alone is unable to extract shape information with sufficient accuracy for reliable people detection. Because PSA can output more accurate spatial information for people, including motion information, our method works better for the cluttered backgrounds including occlusion situations.

In addition, Fig. 13 shows examples of people detection using PETS2001 dataset[2]. PETS2001 dataset includes images in which people and cars are passing through streets, tree leaves are flickering, and the illumination conditions are varying rapidly. In this hard environment, proposed method using co-occurrence of HOG and PSA features was able to improve the accuracy of detecting people.

## 6 Conclusion

We have developed and tested a method for detecting people that is based on the co-occurrence of appearance and spatio-temporal features. It uses pixel state analysis to obtain spatio-temporal information, which enables it to accurately detect people when there is a complicated background. Future work involves creating a corresponding method for active cameras and camera motion.

[2] http://www.cvg.rdg.ac.uk/PETS2001/

## References

[1] A. Lipton, H. Fujiyoshi and R. Patil, “Moving target classification and tracking from realtime video”, *IEEE Workshop on Application of Computer Vision*, pp. 8–14, 1988.

[2] W. E. L. Grimson, C. Stauffer, R. Romano and L. Lee, “Using adaptive tracking to classify and monitor activities in a site”, *IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 22–31, 1998.

[3] O. Hasegawa and T. Kanade, “Type classification, color estimation, and specific target detection of moving targets on public streets”, *IEEE Machine Vision and Applications*, pp. 116–121, 2004.

[4] K. Levi and Y. Weiss, “Learning object detection from a small number of examples: the importance of good features”, *IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 53–60, 2004.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, *IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[6] B. Wu and R. Nevatia, “Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors”, *IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 90–97, 2005.

[7] F. Suard and A. Broggi: “Pedestrian detection using infrared images and histograms of oriented gradients”, *IEEE Symposium on Intelligent Vehicles*, pp. 206–212, 2006.

[8] Q. Zhu, S. Avidan, M. Yeh and K. Cheng, "Fast human detection using a cascade of histograms of oriented gradients", *IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 1491–1498, 2006.

[9] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance", *IEEE European Conference on Computer Vision*, pp. 428–441, 2006.

[10] P. Viola, M. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance", *IEEE International Conference on Computer Vision*, pp. 734–741, 2003.

[11] H. Fujiyoshi and T. Kanade, "Layered detection for multiple overlapping objects", *IEICE Transactions on Information and systems*, pp. 2821–2827, 2004.

[12] T. Mita, T. Kaneko, B. Stenger and O. Hori, "Discriminative feature co-occurrence selection for object detection", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1257–1269, 2008.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *IEEE CS Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, 2001.

[14] D. Comaniciu and P. Meer, "Mean shift analysis and applications", *IEEE International Conference on Computer Vision*, pp. 1197–1203, 1999.

## Appendix Development of variance equation

Variance $S$ is defined by the following equation.

$$S = \frac{1}{K}\sum_{i=1}^{K}(I_{(t+i)} - \bar{I})^2 \tag{18}$$

Here, the important point to note is that equation (18) need to calculate the average intensity that is a high calculation cost in each frame. Therefore, we develop equation (18) as following

$$\begin{aligned}
&\frac{1}{K}\sum_{i=1}^{K}(I_{(t+i)} - \bar{I})^2 \\
&= \frac{1}{K}\sum_{i=1}^{K}(I_{(t+i)}^2 - 2I_{(t+i)}\bar{I} + \bar{I}^2) \\
&= \frac{1}{K}\sum_{i=1}^{K} I_{(t+i)}^2 - 2\sum_{i=1}^{K} I_{(t-i)}\bar{I} + \bar{I}^2 \\
&= \frac{1}{K}\sum_{i=1}^{K} I_{(t+i)}^2 - 2\bar{I}^2 + \bar{I}^2 \\
&= \frac{1}{K}\sum_{i=1}^{K} I_{(t+i)}^2 - \bar{I}^2 \\
&= \frac{1}{K}\sum_{i=1}^{K} I_{(t+i)}^2 - \Big(\frac{1}{K}\sum_{i=1}^{K} I_{(t+i)}\Big)^2 \\
&= \frac{K\sum_{j=1}^{K} I_{(t+j)}^2 - \big(\sum_{j=1}^{K} I_{(t+j)}\big)^2}{K^2}.
\end{aligned} \tag{19}$$

In the case of unbiased variance, equation (19) represents

$$S = \frac{K\sum_{j=1}^{K} I_{(t+j)}^2 - \big(\sum_{j=1}^{K} I_{(t+j)}\big)^2}{K(K-1)}. \tag{20}$$

In this way, equation (20) takes advantage that there is no need to calculate the average intensity, thus it can enable fast calculation of the variance.

**Yuji YAMAUCHI**

Yuji YAMAUCHI received the BS and MS degrees, both in computer science, from Chubu University, Japan, in 2007 and 2009. He is currently pursuing his Ph. D and has been in the Graduate School of engineering, Chubu University since 2009. His research interests include computer vision and pattern recognition. He is a member of the IEICE.

**Hironobu FUJIYOSHI**

Hironobu FUJIYOSHI received his Ph.D. in Electrical Engineering from Chubu University, Japan, in 1977. From 1997 to 2000 he was a postdoctoral fellow at the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA, USA, working on the DARPA Video Surveillance and Monitoring (VSAM) effort and the humanoid vision project for the HONDA Humanoid Robot. He is now an associate professor of the Department of Computer Science, Chubu University, Japan. From 2005 to 2006, he was a visiting researcher at Robotics Institute, Carnegie Mellon University. His research interests include computer vision, video understanding and pattern recognition. He is a member of the IEEE, the IEICE, the IPSJ, and the IEE.

**Yuji IWAHORI**

Yuji IWAHORI received the B.S. from Dept. of Computer Science, Faculty of Engineering, Nagoya Institute of Technology in 1983, the M.S. degree and the Ph.D. degree from Dept. of Electrical and Electronics Engineering, Tokyo Institute of Technology, in 1985 and 1988. He joined the Educational Center for Information Processing, Nagoya Institute of Technology as a research associate in 1988. He became an associate professor and a professor of the same institute in 1992 and 2002, respectively. Then he has become a professor of Dept. of Computer Science, Chubu University since 2004, a graduate course head of Computer Science during 2007-2008 and a dept. head since 2009. In the meantime, he joined the Laboratory for Computational Intelligence, Dept. of Computer Science, University of British Columbia as a visiting professor. His interests include Computational Vision, Neural Network and Pattern Recognition. He is a member of the IEICE, IEEE Computer Society and Information Processing Society of Japan.

**Takeo KANADE**

Takeo KANADE received his Ph.D. in Electrical Engineering from Kyoto University, Japan, in 1974. After being on the faculty at Department of Information Science, Kyoto University, he joined Computer Science Department and Robotics Institute in 1980. He became Associate Professor in 1982, a Full Professor in 1985, the U. A. and Helen Whitaker Professor in 1993, and a University Professor in 1998. He is the Director of the Robotics Institute since 1992. He served as the founding Chairman (1989-93) of the Robotics Ph. D. Program at CMU, probably the first of its kind in the world. Dr. Kanade has worked in multiple areas of robotics, ranging from manipulator, sensor, computer vision, multi-media applications and autonomous robots, with more than 200 papers on these topics. He has been the founding editor of International Journal of Computer. Dr. Kanade's professional honor includes: election to the National Academy of Engineering, a Fellow of IEEE, a Fellow of ACM, and a Fellow of American Association of Artificial Intelligence; several awards including Joseph Engelberger Award, Yokogawa Prize, JARA Award, Otto Franc Award, and Marr Prize Award.