Incoherent Motion Detection using a Time-series Gram Matrix Feature

Masato Kazui, Masanori Miyoshi, Shoji Muramatsu
Hitachi Research Laboratory
E-mail:masato.kazui.bq@hitachi.com

Hironobu Fujiyoshi
Chubu University
hf@cs.chubu.ac.jp

Abstract

This paper proposes a new method for incoherent motion recognition from video sequences. We use time-series spatio-temporal intensity gradients within a space-time patch. Using a global space-time patch, we found that the gradient feature allows us to distinguish an incoherent motion from a coherent motion without segmentation. Furthermore the algorithm can run in real time even on an embedded device. In this paper, we verify motion recognition performance for actions which we consider coherent (walk/run) and incoherent (turn/squat/inverse walk). To identify the multiple motion classes, we use linear discriminant analysis and the KNN method. As a result, our method can distinguish multiple-class motion patterns with a detection rate of about 80%. Also the detection rate of incoherent motions is 100% with a false positive rate of less than 10 %.

1. Introduction

In recent years, the decreasing cost of CCTV cameras and hard disk recorders has led to the wide-spread use of large scale surveillance systems. However there is a risk that in a network of thousands of local cameras the image server can overflow, causing the server to go down. To avoid such a problem, we need to decrease traffic between the cameras and the server by installing intelligence such as behavior-based correlation [1], event detection [3][4][8], and abnormal action detection [2][5] on the cameras and the server.

E. Shechtman, et.al. have proposed an algorithm for detecting motion patterns using correlation between a spatio-temporal event template and video sequences [1]. This method is based on the continuous rank-increase measure of Gram matrix of a local space-time patch and can detect the query pattern even in a noisy sequence. Yu, et.al. have presented stable contact concept, which comes from extreme points of human contour [5]. The stable contact is trained by Hidden Markov Model, and specific motion patterns, such as fence-climbing and rock-climbing can be detected by this method.

Since our development is aimed at embedded devices such as IP cameras, the computational costs of the space-time volume scanning against an input video sequence [1], or of the spatio-temporal mean shift clustering [3] are critical, e.g., the previous method [1] needs 30 minutes for searching a $60 \times 30 \times 30$ query against a $144 \times 180 \times 200$ video sequence. In order to reduce the computational costs of the Gram matrices and the scanning, we use a large space-time patch in place of the small space-time patch used in [1, 3]. Since the Gram matrix can represent motion coherency or motion discontinuity [1][6] within the patch, use of the large patch allows us to distinguish an incoherent motion from a coherent motion without segmentation. Furthermore, we use time-series Gram matrix components which come from spatio-temporal intensity gradients within the large space-time patch. This is because one space-time patch is not enough for representing multi-motion classes. In our experiments, we verify motion recognition performance for actions which we consider coherent (walk/run) and incoherent (turn/squat/inverse walk). To identify the multiple motion classes, we use linear discriminant analysis and the K Nearest Neighbor (KNN) method.

The rest of this paper is organized as follows: Section 2 describes the incoherent detection algorithm; Section 3 shows the experimental results; and Section 4 summarizes and describes considerations and future work.

2. Algorithm Overview

Our algorithm is based on features from spatio-temporal gradients in a global space-time patch. To detect the incoherent motion, we utilize linear discriminant analysis and the KNN method to a feature vector given by time-series Gram matrices. The algorithm is described in detail below.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

2.1 Time-series Gram matrix feature vector

A feature vector for representing a human motion is derived from a Gram matrix. The matrix is obtained from an optical flow equation as shown below.

$$\nabla P_i(u, v, w)^T = 0 \tag{1}$$

Here (u, v, w) is the homogeneous expression of the 2-D optical flow vector, and $\nabla P_i$is the space-time gradient of the intensity at each pixel within a space-time patch (hereafter, ST-patch) $(i = 0, 1, 2, ...n)$, which is a small video clip, e.g., a $7 \times 7 \times 3$ pixel space-time volume. The ST-patch is used to compute the behavioral similarity between two video segments (e.g., query and reference).

$$\underbrace{\begin{bmatrix} P_{x_1} & P_{y_1} & P_{t_1} \\ P_{x_2} & P_{y_2} & P_{t_2} \\ & \cdots & \\ & \cdots & \\ P_{x_n} & P_{y_n} & P_{t_n} \end{bmatrix}_{n\times 3}}_{\mathrm{G}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}_{n\times 1} \tag{2}$$

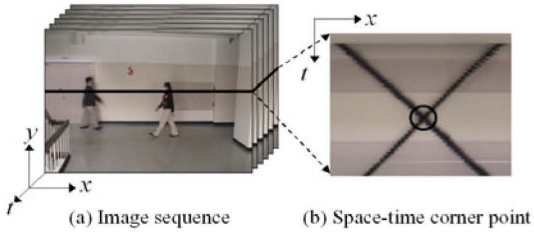
where n is the number of pixels in the patch. We obtain Gram matrix *M* by multiplying both sides of Eq.(2) by the transposition of G, as below.

$$M = G^T G = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y & \Sigma P_x P_t \\ \Sigma P_y P_x & \Sigma P_y^2 & \Sigma P_y P_t \\ \Sigma P_t P_x & \Sigma P_t P_y & \Sigma P_t^2 \end{bmatrix} \tag{3}$$

This matrix is in fact a co-variance matrix of the space-time gradients within the ST-patch. E. Shechtman, et.al. use a small ST-patch and assume the patch includes a locally uniform motion [1]. However, this assumption will be violated at motion discontinuities in a video sequence. They use a set of the patches as a query for detecting a behavior pattern from video clips.

On the other hand, we use a large ST-patch, the size of which is the image size at most, and determine whether or not the large patch includes an incoherent motion. As described in [1], when the ST-patch includes multiple independent motions, the rank of *M* becomes 3.

Here, we assume 1-D motion for simplicity. When multiple motions occur in a scene as shown in Fig.1 (a), the corresponding x-t image, which consists of scan lines, contains two trajectories. As a result, there exist an intersection, i.e., a space-time corner point as denoted by a circle in Fig.1 (b). We can detect such the



(a) Image sequence (b) Space-time corner point

Figure 1. Incoherency of motion

point by use of Harris detector [7]. Harris detector evaluates a co-variance matrix as below.

$$M^\circ = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_t \\ \Sigma P_t P_x & \Sigma P_t^2 \end{bmatrix} \tag{4}$$

If there exists the space-time corner point, then $rank(M^\circ)$ becomes 2. Therefore, $rank(M^\circ) = 2$ implies that there could be multiple motions, i.e., incoherent motion in a summation area. In the case of 2-D motion, the matrix can be expanded into Eq.(3), and the multiple motions make the rank of M 3. We exploit this property for detecting incoherent motion, such as inverse running, falling, or putting an object in a crowd. However, we use the Gram matrix components explicitly, while the rank or continuous rank-increase measure of *M* is used for action detection [1]. This is because we consider the rank of a large ST-patch degenerates the motion properties.

Though the Gram matrix in the ST-patch can describe the motion coherency at each frame, we cannot know a motion type simply through a sequence of frames. In our research, we use a time-series chain of the Gram matrix components as a feature vector x in order to recognize motion patterns, as below:

$$x = [m_t, m_{t-1}, m_{t-2}, \cdots, m_{t-L}]^T \tag{5}$$

$$m = [\Sigma P_x^2, \Sigma P_x P_y, \Sigma P_x P_t, \Sigma P_y^2, \Sigma P_y P_t, \Sigma P_t^2] \tag{6}$$

where L is the number of frames to be used. We don't use duplicate components of *M*. Therefore the dimension of x is $6 \times L$. The length of L depends on the cycle of the motion to be detected. Empirically, it is said that the human walk cycle is about 1 second, i.e., 30 frames. Though some might say we need to use $L = 30$, we set $L = 5$ because we assume that instantaneous motion is a characteristic for detecting an incoherent motion. We need to set the optimum length of L, or eventually, change it dynamically.

2.2 Motion pattern classification

Since m changes frame by frame consecutively, the feature vector x has redundancy. In this research work, we project x onto a low-dimensional subspace by linear discriminant analysis (LDA) of Eq.(7).

$$W^{-1}B, \tag{7}$$

where B is an inter-class covariance matrix and W is an intra-class covariance matrix denoted as:

$$B = \sum_{k=1}^{J} n_k \left(\bar{x_k} - \bar{x}\right)\left(\bar{x_k} - \bar{x}\right)^T, \tag{8}$$

$$W = \sum_{k=1}^{J}\sum_{i=1}^{n_k} \left(x_{k,i} - \bar{x}\right)\left(x_{k,i} - \bar{x}\right)^T, \tag{9}$$

where J is the number of motion classes, n is the number of the total samples, n_k is the number of samples in k-th class, $x_{k,i}$ is i-th input vector of k-th class, $\bar{x}$ is a mean vector of all of $x_{k,i}$, and $\bar{x_k}$ is a mean vector of k-th class. We decide the dimensions of discriminant subspace M using a threshold from Eq.(10):

$$TH \times \sum_{i=1}^{N} \lambda_i \geq \sum_{i=1}^{M} \lambda_i, \tag{10}$$

where λ_i is the i-th eigen value of Eq.(7), and N is the dimension of x:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N. \tag{11}$$

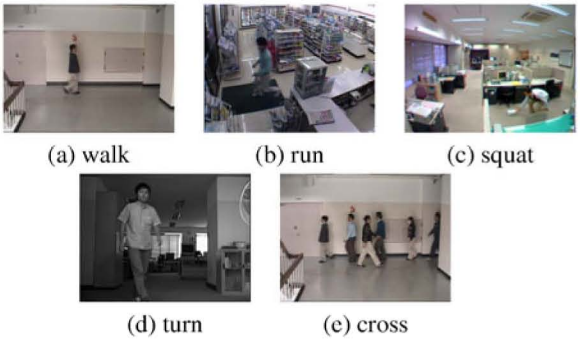
After the projection, the feature vectors make clusters in the discriminant subspace. We use the KNN method for classifying motion patterns in the subspace. In the training phase, the training data set after the projection is quantized by the LBG algorithm [9] to select representative vectors $\hat{y}$ for the each class. Input vectors are classified into motion classes defined using the representative vectors. Then a consecutive frame judgment is applied to the classification result in order to decrease false positives. Fig. 2 presents various motion patterns.

3. Experimental Results

3.1 Conditions and parameter setting

We use frame subtraction edges to obtain the Gram matrix of Eq.(3), while E. Shechtman, et.al. [1] use all the pixels within an ST-patch. This is because the incoherent motion of a small object could be buried under the global noise which occurs in an image. We use a constant value for the frame subtraction, e.g., 15, in this paper. Our method works well for now under the conditions that a pedestrian in an image is not occluded and the size of the pedestrian is over 60×30 pixels against an image of 320×240 pixels. For an occluded or small pedestrian, we need to introduce a-part based approach [3]. To validate the feature vector x of Eq.(5), in this section, we prepared the appropriate sample sequences shown in Fig.1.

As described in Section 2.1, we set the size of the ST-patch to an image size, e.g., 320×240. The dimension of the discriminant subspace M of Eq.(10) is 15 and the number of frames L for the input vector is 6.



(a) walk (b) run (c) squat

(d) turn (e) cross

Figure 2. Motion patterns

Table 1. Confusion matrix

True class →	walk	run	squat	turn	cross
walk	**80.9**	12.8	6.1	4.1	0
run	8.3	**80.1**	3.5	3.4	6.9
squat	10.0	1.5	**83.4**	20.5	0.5
turn	0.8	1.9	7.0	**61.1**	3.2
cross	0	3.7	0	10.8	**91.5**

3.2 Motion pattern classification

First, in order to verify the separation performance of the intra-class of training sequences, we define coherent and incoherent motion classes as follows. Coherent motions are "walk" and "run". Incoherent motions are "squat", "turn", and "cross". Key frames of these motions are shown in Fig. 2. Each of the motion classes has twenty sequences and each sequence includes 60 to 100 frames. Using this dataset, we verify the separation performance by the KNN method in a linear discriminant space. The verification result is shown in Table 1 as a confusion matrix. The table denotes the percentages of detected frames for each motion class. As shown in the table, the separation performance is about 80% for

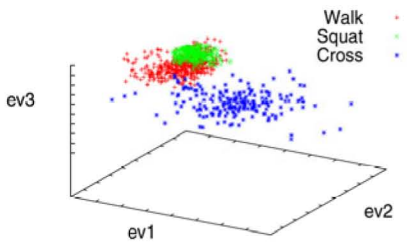


Figure 3. Discriminant space

the each class. The separation performance for "turn" is worse than for the others. This is because "turn" partially includes "walk". On the other hand, "cross" shows the best performance. This is because "cross" includes two independent motions in the ST-patch, maximizing the rank of M in Eq.(3): in other words, the motion of the ST-patch is incoherent.

The separation is not perfect because clusters of the motion classes overlap each other in a linear discriminant space. Fig. 3 shows an example of the clusters (this figure shows only the upper 3 dimensions of M of Eq.(10)). Though it is difficult to boost the classification rate due to the overlaps, we can reduce false positives by using the consecutiveness of a human action. We discuss this property in the next section.

3.3 Incoherent motion detection

We verified a consecutive frame judgment to reduce the false positives described in the previous section. Input sequences are different from the training dataset, and we define a true detection as follows: When the same classification result occurs consecutively in N_c frames within a sequence, we count it as a true detection, where N_c is the number of consecutive frames. This definition arises from the difficulty of motion segmentation. In other words, no matter how we segment motions, there will be frames which belong to more than one motion class, due to the consecutiveness of human motions.

Fig.4 shows comparison results between CHLAC (Cubic Higher-order Local Auto-Correlation) [2] and our method (denoted as "ST-patch"). CHLAC is based on higher-order correlation within a local region, e.g., $3\times3\times3$ pixels for example. CHLAC has several preferable properties, which are shift invariance to data, additivity for data, and robustness to noise in data. Due to the properties, CHLAC has been used for Gait recognition recently. Here, we show the results from "walk", "run", "squat", "turn", and "cross". Each graph shows true positive and false positive rate for both of CHLAC and our method. Altogether, the larger N_c is, the lower the false detection rate becomes. As shown in Fig.4(a), (b), and (c), the performance of the both methods is almost same. On the other hand, the true positive rate of our method is superior to CHLAC in case of "turn" and "cross". Also the detection rate of incoherent motions is 100% with a false positive rate of less than 10 %. From these results, it can be said that our method is suitable for detecting incoherent motions. However, it is difficult for our method to deal with geometric transformations of the sequences and different speed in the human actions. We are going to apply a pedestrian detector and scale normalization to deal with the problems.

Our algorithm can run with a 320×240 pixel image within about 30 milli seconds on a SH4 200MHz processor and Image Processing Accelerator VCHIP-II [10]. This speed is fast enough for an embedded device, such as an IP camera.

4. Conclusion and future work

We have presented an approach for detecting incoherent motion in video clips. In this paper, we have described the method, which uses time-series Gram matrix components, and have shown its performance. The results of our experiment demonstrate that our method can distinguish multiple-class motion patterns with a detection rate of about 80%. Also the detection rate of incoherent motions is 100% with a false positive rate of less than 10 %. Though direct comparison of the run-time is difficult, our algorithm can run with a 320×240 pixel image within about 30 milliseconds on a SH4 200MHz processor and Image Processing Accelerator VCHIP-II, while the previous method [1] needs 30 minutes for seraching a $60 \times 30 \times 30$ query against a $144\times180\times200$ video sequence. In future work, we will enhance our method by using a segmentation based approach [3], in order to deal with the occlusion problem and detect more precise motion patterns.

References

[1] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. CVPR*, 2005.

[2] T. Nanri and N. Otsu. Unsupervised Abnormality Detection in Video Surveillance. In *Proc. MVA*, 2005.

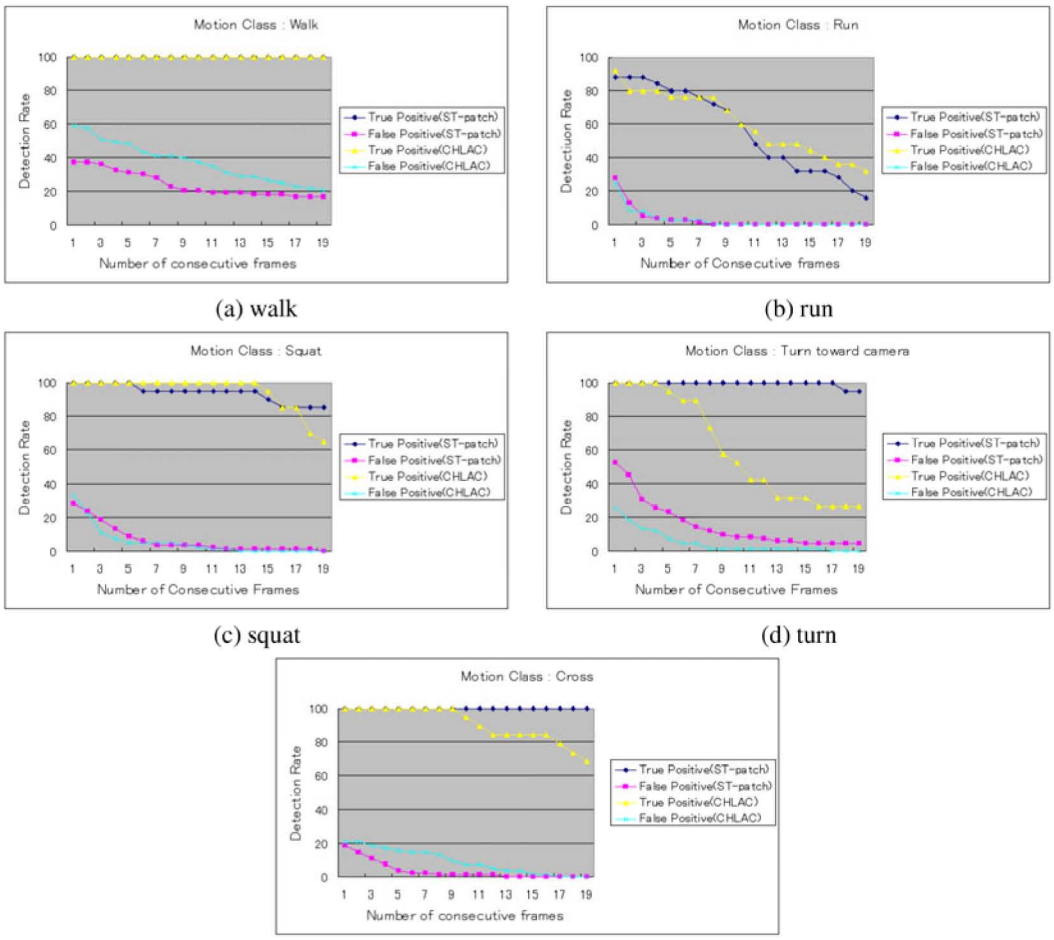


Figure 4. Comparison results of incoherent motion detection

[3] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Cluttered Videos. In *Proc. ICCV*, 2007.

[4] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. ICCV*, 2005.

[5] E. Yu and J. K. Aggarwal. Detection of stable contacts for human motion analysis. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pp.87–94, 2006.

[6] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, 2003.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pp.147–151, 1988.

[8] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. ICCV*, 2005.

[9] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantization design. *IEEE Transactions on Communications*, Vol.28, No.1, pp.84–94, 1980.

[10] S. Muramatsu, Y. Kobayashi, Y. Otsuka, H. Shojima, T. Tsutsumi, T. Imai, and S. Yamada. Development of image processing LSI "SuperVchip" for real-time vision systems. In *Proc. SPIE 4666*, 2002.