Pattern Recognition Letters 33 (2012) 1188–1195

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Motion recognition using local auto-correlation of space–time gradients

Takumi Kobayashi *, Nobuyuki Otsu

National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba 305-8568, Japan

ARTICLE INFO

Article history:
Received 20 November 2010
Available online 20 January 2012
Communicated by F. Tortorella

Keywords:
Motion recognition
Motion feature extraction
Space–time gradient
Auto-correlation
Bag-of-features

ABSTRACT

In this paper, we propose a motion recognition scheme based on a novel method of motion feature extraction. The feature extraction method utilizes auto-correlations of space–time gradients of three-dimensional motion shape in a video sequence. The method effectively exploits the local relationships of the gradients corresponding to the space–time geometric characteristics of the motion. For recognizing motions, we apply the framework of bag-of-*frame-features*, which, in contrast to the standard bag-of-features framework, enables the motion characteristics to be captured sufficiently and the motions to be quickly recognized. In experiments on various datasets for motion recognition, the proposed method exhibits favorable performances as compared to the other methods, and faster computational time even than real time.

1. Introduction

Motion recognition has attracted a great deal of attention in recent decades and is important for numerous applications, such as video surveillance, man–machine interface, and analysis of sports motion. Significant research efforts in computer vision community have been made to categorize human actions and gestures in video sequences. With the development of the image recognition techniques, methods for recognizing motions have also progressed and produced promising results in recent years.

While conventional methods have used *ad hoc* knowledge based on human body parts (for a survey, refer Gavrilla, 1999), recent studies have employed statistical approaches without such knowledge. By regarding a motion image sequence as three-way data in the space–time (*XYT*) domain, the methods that are applied to (two-way) image recognition have been naturally generalized to motion recognition (Dollar et al., 2005; Jhuang et al., 2007; Laptev et al., 2008; Kobayashi and Otsu, 2009; Blank et al., 2005; Kim et al., 2007). The motion is explicitly dealt with as space–time shape by Blank et al. (2005) who extracted human silhouettes from motion images.

In particular, the bag-of-features framework (Csurka et al., 2004) has been successfully applied to motion recognition (Dollar et al., 2005; Laptev et al., 2008; Wong and Cipolla, 2007) as well as image recognition (Bosch et al., 2007). In that framework, the recognition of motion relies on local features which are based on simple histograms of spatial gradient orientations (HOG) (Dalal and Triggs, 2005) and space–time derivatives (Dollar et al., 2005; Zelnik-Manor and Irani, 2006). These local features cannot fully capture the space–time shape of the motions and do not have much discriminative power. Therefore, in the bag-of-features framework, the motion is represented as ensembles of numerous local features extracted around the space–time interest points which are sparsely detected by, e.g., a Harris-Laplace detector (Laptev, 2005) or a nonnegative matrix factorization (NMF) like detector (Wong and Cipolla, 2007). The sparse interest points, however, are not sufficient to characterize the motion (Dollar et al., 2005; Willems et al., 2008; Ballan et al., 2009), since densely detected interests points (like grid points) improve the performance of image classification (Tuytelaars and Schmid, 2007; Bosch et al., 2007). In motion images, the higher dimensionality due to the three-way data increases the number of interest points even for the sparse detection, which requires a larger computational cost for quantizing the local features into *words*, and the denser detection becomes less feasible.

We propose a novel motion feature extraction method and an effective and high-speed motion recognition scheme based on these features. The feature extraction method exploits the local relationships (co-occurrence) among space–time gradients in the *XYT* domain, by developing the gradient local auto-correlation for image recognition (Kobayashi and Otsu, 2008) to extract space–time motion features. The local relationships correspond to geometric characteristics, i.e., gradients and curvatures, which are fundamental properties of space–time motion shape. For motion recognition, we utilize the *frame*-based features which are extracted from sub-sequences sampled at dense (grid) time points along the time axis. In this approach, referred to as the *bag-of-frame-features* approach, the frame-based features sufficiently characterize the motion in the spatial domain in contrast to the local features, and the motion in the entire sequence is described by

* Corresponding author. Tel.: +81 29 861 5491; fax: +81 29 861 3313.
E-mail addresses: takumi.kobayashi@aist.go.jp (T. Kobayashi), otsu.n@aist.go.jp (N. Otsu).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.01.007

the densely sampled features along the time axis. The bag-of-frame-features approach is effective and fast due to the reduced computation of the frame-based features achieved by applying integral histograms (Porikli, 2005) and the small number of the sampling points placed only along the time axis without a requirement for time consuming interest point detection.

This paper has the following three main contributions: (1) to propose a novel motion feature extraction method, (2) to demonstrate the favorable performance of the proposed method for motion recognition on various datasets as compared to the other methods, and (3) to exhibit much faster computational time even than *real time*. In particular, the proposed motion features are based on *co-occurrence* histograms of the space–time 3D gradient orientations and they are employed for frame-based features to *densely* characterize the motion in contrast to recent works which sparsely describe the motions by using simple occurrence histogram of gradient orientations. To facilitate the implementation, we explicitly describe the practical details of the proposed method, such as parameter settings.

The rest of the paper is organized as follows: the next section describes details of the proposed motion feature extraction method. Then, we describe the scheme to recognize motion using the features in Section 3. In these sections, we also describe implementation details, such as parameter values, of the proposed method as practical issues. In Section 4, the experimental results for motion recognition are shown. Finally, Section 5 contains our concluding remarks.

2. Feature extraction

First, we describe the method for extracting features of motion in the space–time domain. The image feature extraction method in (Kobayashi and Otsu, 2008) is developed to deal with space–time volume in an image sequence, and we call the proposed method *space–time auto-correlation of gradients* (*STACOG*). STACOG extracts the local relationships, such as co-occurrence, among the space–time (three-dimensional) gradients by means of the auto-correlation functions regarding the space–time orientations and the magnitudes of the gradients. The local relationships are closely related to the local geometric characteristics of space–time motion shape. In addition, STACOG has the property of *shift-invariance* which is desirable for recognition.

2.1. Space–time gradient

The space–time (three-dimensional) gradient vector is calculated by derivatives (I_x, I_y, I_t) of motion image volumes $I(x,y,t)$ at each space–time point in an image sequence. As shown in Fig. 1(a), the gradient vectors can be geometrically represented by the magnitudes $m = \sqrt{I_x^2 + I_y^2 + I_t^2}$ and two types of angle: spatial orientation $\theta = \arctan(I_x, I_y)$ in an image frame and temporal elevation $\phi = \arcsin(I_t/m)$ along the time axis, where the functions arctan and arcsin output the angles within $[0, 2\pi)$ and $[-\pi/2, \pi/2]$, respectively. The space–time orientation of the gradient defined by these two angles is coded into B orientation bins on a unit sphere by voting weights to the nearest bins (Fig. 1(b)). Then, the orientation is finally described by a B-dimensional vector $\boldsymbol{h}$, called space–time orientation coding (STOC) vector. The STOC vector $\boldsymbol{h}$ consists of the weights voted to B bins and is sparse: The number of non-zero elements is at most four (see Fig. 1(a)).

Practical issue. For coding the gradients, we consider a hemisphere ignoring the opposite directions of the gradients. Thus, bins are located on the hemisphere as follows. Four orientation bins along the longitude are arranged on each of five layers along the latitude, and one bin is located at pole. Thus, there are a total of $B = 21$ bins, as illustrated in Fig. 1(b).

2.2. Definition of STACOG

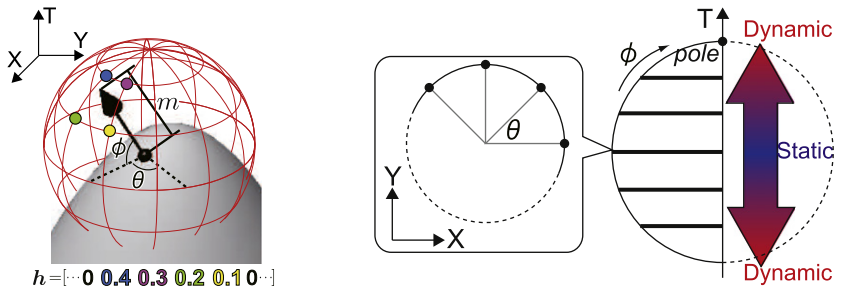
The Nth order auto-correlation function for the space–time gradients is defined by using the magnitude m and the STOC vector $\boldsymbol{h}$ of the gradients as follows:

$$\boldsymbol{R}_N(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N) = \int w[m(\boldsymbol{r}), \ldots, m(\boldsymbol{r}+\boldsymbol{a}_N)]\boldsymbol{h}(\boldsymbol{r}) \otimes \cdots \otimes \boldsymbol{h}(\boldsymbol{r}+\boldsymbol{a}_N)d\boldsymbol{r}, \tag{1}$$

where $\boldsymbol{a}_i$ are displacement vectors from the reference point $\boldsymbol{r} = (x,y,t)$, w is a weighting function, and $\otimes$ denotes the tensor product of the vector. In the tensor products, there are a few non-zero components associated to the gradient orientations of the neighbors indicated by $\boldsymbol{a}_i$. Thus, Eq. (1) extracts the local relationships such as co-occurrence of space–time gradients (Fig. 2(a)).

We restrict the parameters such that $N \in \{0,1\}$, $a_{1x,y} \in \{\pm\Delta r, 0\}$, $a_{1t} \in \{\Delta t, 0\}$, $w(\cdot) \equiv \min(\cdot)$, as in (Kobayashi and Otsu, 2008). For the displacement interval, we use different parameters, Δr and Δt, in the spatial and temporal axes, respectively. For the spatial axes, the interval along the x-axis is made equal to that along the y-axis because of isotropy in the XY plane. On the other hand, the temporal interval Δt may be different from the spatial interval Δr because the resolutions of space and time may differ. With respect to the weight function w, we adopt min in order to suppress the effect of isolated noise on surrounding auto-correlations.

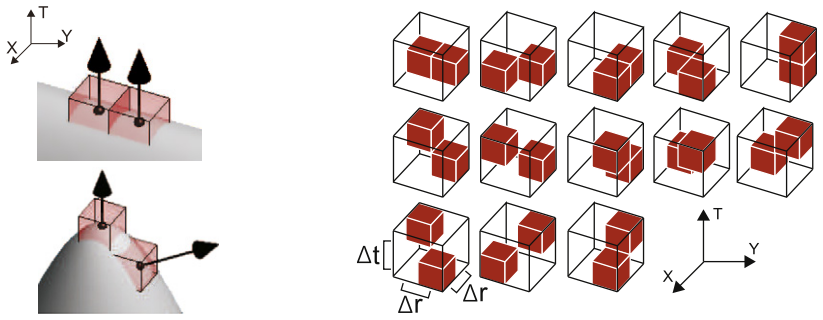
Consequently, we obtain the following practical formulation of STACOG:



(a) Space-time gradient of equi-gray level surface

(b) Orientation bins along latitude and longitude

Fig. 1. (a) The space–time (three-dimensional) gradients are described by the gradient magnitude m and STOC vector $\boldsymbol{h}$ which codes the gradient orientations (ϕ, θ). (b) The orientation coding is based on bins (denoted by black dots) on a hemisphere, ignoring opposite directions along the longitude. The orientation bins are categorized into two types along the latitude: static bins (blue) and dynamic bins (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Auto-correlation (b) Configuration patterns of local pairs

Fig. 2. STACOG exploits the local relationships of the gradient pairs (a) defined by the configuration patterns (b).

$$\textbf{0th order}: \boldsymbol{F}_0 = \sum_{\boldsymbol{r}} m(\boldsymbol{r})\boldsymbol{h}(\boldsymbol{r}), \tag{2}$$

$$\textbf{1st order}: \boldsymbol{F}_1(\boldsymbol{a}_1) = \sum_{\boldsymbol{r}} \min[m(\boldsymbol{r}), m(\boldsymbol{r}+\boldsymbol{a}_1)]\boldsymbol{h}(\boldsymbol{r})\boldsymbol{h}(\boldsymbol{r}+\boldsymbol{a}_1)^T. \tag{3}$$

Note that the tensor product of two vectors simply equals to outer product ($\cdot^T$ denotes transpose) and the features $\boldsymbol{F}_0$, $\boldsymbol{F}_1$ are actually vectorized. There are 13 configuration patterns of $(\boldsymbol{r}, \boldsymbol{r}+\boldsymbol{a}_1)$ after eliminating duplicated ones, as shown in Fig. 2(b). For the first-order features $\boldsymbol{F}_1$, we calculate the outer product of the STOC vector pairs indicated by the configuration patterns and sum them over the space–time region with the magnitude-based weights. Despite the high dimensionality of STACOG features ($d = B + 13B^2$), the required computational cost is low due to the sparseness of STOC vector $\boldsymbol{h}$. Namely, the operations in Eqs. (2) and (3) are applied only to a few (at most, four) non-zero elements of the vector $\boldsymbol{h}$.

Practical issue. As to displacements, $\Delta r/\Delta t$ is closely connected to velocity of target motion. We set $\Delta t = 1$, i.e., temporally adjacent auto-correlation to cope with faster motions, and $\Delta r \in \{1,\ldots,8\}$ in which the three most discriminative ones are selected in the later process, as described in Section 3.3. Note that the displacements are limited to local neighbors since local gradients are assumed to be highly correlated.

2.3. Discussion of STACOG

2.3.1. Statistical property

The zeroth-order features $\boldsymbol{F}_0$ in Eq. (2) are equivalent to a histogram of the space–time gradient orientations which is similar to SIFT (Lowe, 2004), HOG (Dalal and Triggs, 2005) widely used in image recognition and 3D-SIFT (Scovanner et al., 2007). Such histogram indicating simple occurrence is an approximation (quantization) of the orientation probability distribution over space–time region. In the proposed method, the point is that we extract the first-order features $\boldsymbol{F}_1$ in Eq. (3) corresponding to a *joint* (*co-occurrence*) *histogram* of the space–time orientations of local gradient pairs, which is a natural extension of the standard orientation histogram. These first-order features are an approximation of the joint probability distribution of the local orientation pairs like co-occurrence matrix (Haralick et al., 1973) and color correlograms (Huang et al., 1997).

2.3.2. Geometrical property

From the geometrical perspective, in the first-order features $\boldsymbol{F}_1$, the combinations (co-occurrences) of the gradient pairs quantize and patternize the three-dimensional curvatures of the space–time shape, as shown in Fig. 2(a), while the zeroth-order features $\boldsymbol{F}_0$ characterize only the gradients similarly to 3D-SIFT (Scovanner et al., 2007). The gradients and the curvatures extracted by the zeroth-order and the first-order features are fundamental geometric characteristics of the space–time motion shape. Along the time axis, these geometric characteristics also correspond to velocity and accelerations of motion. Note that STACOG can simply extract such geometric features via space–time gradients without explicitly extracting silhouettes (Blank et al., 2005).

2.3.3. Dynamic and static components

The space–time orientations of the gradients seamlessly reflect *dynamic* and *static* situations along the latitude (Fig. 1(b)),[1] and correspondingly the STOC vectors $\boldsymbol{h}$ contain dynamic and static components which are caused by the motion and the static figures, such as background, respectively. Consequently, in STACOG features, the zeroth-order features $\boldsymbol{F}_0$ can be categorized into two types, *dynamic* and *static*, while the first-order features $\boldsymbol{F}_1$ are also categorized into the following three types: The first type is derived from dynamic motion only (correlation of *dynamic* × *dynamic* components in $\boldsymbol{h}$). The second type is derived from the relationship between dynamic motion and static figure (*dynamic* × *static* correlation). Finally, the third type is derived from static figures only (*static* × *static* correlation). Thus, both the dynamic and static information is extracted by STACOG in a unified manner. For motion recognition, however, the *static* features in $\boldsymbol{F}_0$ and the *static* × *static* ones in $\boldsymbol{F}_1$ might be unnecessary and can be eliminated since these features do not contain any information about the motion itself. On the other hand, all types of feature would be useful for tasks that simultaneously require information about motion and static objects, such as scene understanding. We demonstrate how these types of feature component contribute to performance in the experiments of motion recognition (Section 4.1).

2.3.4. Comparison to related work

The proposed method is closely related to the method of (Kobayashi and Otsu, 2009) which also extracts auto-correlations in space–time domain, although the information to be correlated is different. For extracting motion information, 0/1 (static/dynamic) *scalar* values extracted by frame-differencing and binarization are only used in (Kobayashi and Otsu, 2009), while space–time *gradient vectors*, especially their orientations ($\boldsymbol{h}$), are used in the proposed method. Thus, the auto-correlation function itself is differently defined based on tensor products in Eq. (1). The gradient vectors provide richer information for motion recognition, efficiently describing the geometrical characteristics of motion shape as follows.

In (Kobayashi and Otsu, 2009), the geometrical characteristics are captured only by the point configuration defined by mask patterns as shown in Fig. 3(b), and the variations of the characteristics solely rely on those of the mask patterns. However, it is dif-

[1] Pixel values belonging to moving objects fluctuate along the time-axis, resulting in the corresponding space–time gradients being oriented along the time-axis to a certain degree, whereas the orientations in static regions remain on *XY* plane.

(a) STACOG (b) The method of (Kobayashi and Otsu, 2009)

Fig. 3. Comparison of the proposed method and (Kobayashi and Otsu, 2009).

ficult to capture all variations of local geometry by using such discrete and finite mask patterns. On the other hand, in STACOG, even by using quantization of *continuous* gradient orientations, detailed local geometric characteristics can be efficiently captured. It is quite advantageous for extracting motion features which contain the various geometrical characteristics in three dimensions. In addition, the relationship of even two gradients in Eq. (3) can extract the characteristics of curvatures unlike (Kobayashi and Otsu, 2009) which requires the relationship among three points for capturing them (Fig. 3).

The proposed method is naturally extended from (Kobayashi and Otsu, 2008) which is based on spatial gradients for image recognition. In this paper, we define auto-correlation in space–time domain in a unified manner, directly utilizing three-dimensional (quantized) orientations and the magnitudes of space–time gradients. The other related works also employ gradient-based features for characterizing motions; Wong and Cipolla (2005) extract the orientations of the 2D spatial gradients on the motion history images (Bobick and Davis, 2001), and Zelnik-Manor and Irani (2006) construct histograms of the respective 3D space–time gradient component values (I_x, I_y, I_t) for motion features. While these methods extract the motion features densely in the whole image frames or the space–time volume, in recent years, it becomes popular to extract motion features sparsely at interest points (local cuboids) in the framework of bag-of-features (Csurka et al., 2004). Dollar et al. (2005), Wong and Cipolla (2007) directly use, as feature vectors, 3D gradient component values in the cuboids at interest points, Laptev et al. (2008) employ histograms of spatial gradient orientations (HOG) together with those of optical flow orientations (HOF), and Ballan et al. (2009) construct histograms of space–time gradient orientations θ in the spatial domain and ϕ along the time axis, respectively. Scovanner et al. (2007) consider histograms of the two orientations θ, ϕ of the space–time gradients, called 3D-SIFT, which we also employ as primitive STOC vector $\boldsymbol{h}$ in this study (Section 2.1).

Space–time gradients, especially their orientations, extract geometrical characteristics of the space–time motion shape in an essential manner as described in Section 2.3.2. In this paper, the temporal aspect, *static* or *dynamic*, of the space–time gradient orientations is discussed in Section 2.3.3 and its contribution to the performance is experimentally illustrated in Section 4.1. It should be noted that we consider the *co-occurrence* histogram of the space–time gradient orientation pairs in the local neighborhood, while the above-mentioned methods are based on a simple occurrence histogram of the gradient orientations. Such co-occurrence effectively encodes space–time geometrical characteristics as described in Section 2.3.2, attaining high discriminative power, which is advantageously utilized for the subsequent frame-based motion features described in the next section. These points are our main contribution.

3. Bag-of-frame-features scheme

For recognizing motions, we employ the framework of *bag-of-frame-features* taking advantage of discriminative power of STACOG features. In contrast to sparse representation in standard bag-of-features (Dollar et al., 2005; Scovanner et al., 2007; Ballan et al., 2009), this method captures the characteristics of motion sufficiently (densely), and in addition it enables fast computation for motion recognition. This method is based on *frame-based* STACOG features sampled at dense (grid) time points along the time axis. Such densely sampled frame-based features characterize the motion sufficiently in the space–time domain and there are fewer sampling points only along the time axis than interest points in standard bag-of-features. We then apply feature transformation, which is simply Fisher discriminant analysis (FDA) in this paper, to enhance the discrimination between recognition classes while excluding the irrelevant variations, such as background noise, included in these features. The feature vectors are embedded into a lower-dimensional space via FDA and then are quantized into *motion words* at quite low computational cost due to the low dimensionality. An overview of this framework is shown in Fig. 4.

3.1. Frame-based features

The frame-based STACOG features in Eqs. (2) and (3) are extracted by summing up over the full space–time region within a sub-sequence of D time duration (frames). Such frame-based STACOG features are more *global* and sufficiently contain motion characteristics in contrast to the sparse representation using *local* features (Laptev et al., 2008; Dollar et al., 2005). We densely sample the frame-based features at every R frames along the time axis with various time durations D. By exploiting the multiple features of various D, a degree of time-scale variation is allowed. The densely sampled features sufficiently describe the characteristics of the motion in temporal domain without the time-consuming interest point detector (Laptev, 2005), and the number of the features is smaller than that of the local features at space–time interest points. In addition, *shift-invariance* in both the spatial and the temporal domains are rendered by STACOG and the grid time sampling, respectively. The frame-based STACOG features are finally normalized by separately applying SIFT-like normalization (Lowe, 2004; Dalal and Triggs, 2005) to the zeroth-order features $\boldsymbol{F}_0$ and to the first-order ones $\boldsymbol{F}_1$ with the respective threshold values τ_0 and τ_1.

Practical issue. We use parameter values $D \in \{10, 20, 30\}$ and $R = 2$. One-dimensional integral histograms (Porikli, 2005) along the time-axis are applied to calculate the frame-based features in a computationally efficient manner. In particular, for on-line recognition, frames are successively inputted and frame-based STACOG features are calculated incrementally. The threshold values of the normalization are $\tau_0 = 0.03$ for the zeroth-order and $\tau_1 = 0.005$ for the first-order features.

3.2. Motion words

We construct *motion word* clusters by clustering the frame-based feature vectors and then assign these words (cluster IDs) to the features. For motion classification, it is desirable that the

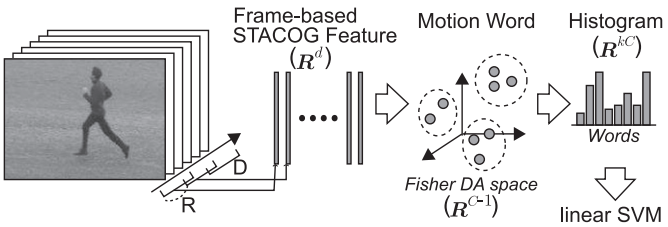


Fig. 4. An overview of bag-of-frame-features for motion recognition.

clusters of *motion words* discriminatively represent motions, and thus we simply apply the Fisher discriminant analysis (FDA) (Duda et al., 2001) to all frame-based features in a learning phase. FDA analytically constructs the lower-dimensional space having a dimensionality of $C-1$, where C is the number of recognition classes. In the FDA space, discrimination between classes is enhanced, which makes *motion word* clusters more favorable, i.e., discriminative, with regard to recognition, while suppressing the irrelevant variations, such as background noise.

Then, k-means clustering is applied to the feature vectors embedded into the FDA space, and *motion word* clusters are obtained. In this paper, in order to avoid bias with respect to the number of samples in each class, the clustering is performed for respective classes; k clusters are constructed in each class and all kC clusters are employed for the motion words. Then, we assign weighted motion words to each frame-based feature as follows (like soft quantization (Philbin et al., 2008)). By considering n nearest neighbor cluster centers to describe the feature, the weights ω_i $(i=1,\ldots,n)$ are defined as

$$\hat{\omega}_i = \begin{cases} 1 & (i=1) \\ \frac{d_1}{d_i} & (i>1) \end{cases}, \quad \omega_i = \frac{\hat{\omega}_i}{\sum_{j=1}^{n}\hat{\omega}_j} = \frac{\prod_{l\neq i} d_l}{\sum_{j=1}^{n}\prod_{l\neq j} d_l}, \tag{4}$$

where d_i is the distance to the ith nearest center, and $\hat{\omega}_i$ is the distance ratio based on the nearest $(i=1)$ distance, which is then normalized to ω_i. Unlike (Philbin et al., 2008) using the weight $\exp(-d_i/\sigma)$, Eq. (4) is based on simple calculation without exp and has only one parameter n. Note that the above soft assignment of the motion words to the frame-based features requires little computational cost since the dimensionality of the FDA space $(C-1)$ is far lower than that of the original feature space $(B+13B^2)$ and the number of extracted features is small. Finally, a motion in a video sequence is described as a histogram of the motion words by gathering the assigned weights ω, and the histogram vector is normalized in L2-norm.

Practical issue. The number of clusters for each class is $k=10$, and so the dimensionality of the final motion word histogram is $10C$ for recognizing C classes. The number of nearest neighbor cluster centers in Eq. (4) is set to $n=3$. These parameter values have been determined empirically to yield favorable performances.

3.3. Classification

For fast computation, we apply linear classification learnt by support vector machine (SVM) (Vapnik, 1998) and a one-against-all approach to handle multi-class recognition.

We have no prior knowledge about the spatial displacement interval Δr, and the number of Δr used must be as small as possible in view of the computational load. Therefore, in the learning phase, the optimal intervals are automatically selected as follows. For each $\Delta r \in \{1,\ldots,8\}$, the motion word histograms $(\in \boldsymbol{R}^{kC})$ are separately constructed by repeating the processes in Sections 3.1 and 3.2. We refer to the histogram associated with each Δr as a *channel*. These channels (histograms) are concatenated into a vector whose dimensionality is $8kC$. For automatically selecting optimal channels, we apply the zero-norm SVM (Weston et al., 2002) to the concatenated vectors in the learning phase. The zero-norm SVM yields sparse weights for each channel and the channels to which non-zero weights are assigned are considered as the optimal (most discriminative) ones for recognition. In this paper, we select *three* optimal channels, and apply the linear classifier described above to the vector into which the optimal channels are concatenated.

4. Experiments

We conducted motion recognition experiments to evaluate the performance of the proposed method by using various datasets (Fig. 5): three human action datasets (KTH, Weizmann and Hollywood2), one human gesture dataset (RWC), and one hand gesture dataset (Cambridge). All of the sequences in the datasets are downsampled into a half spatial resolution. For fair comparison, we employed the standard experimental protocols as in the other works. Details about the datasets and the protocols are given below.

KTH human action: The KTH human action dataset (Schuldt et al., 2004) contains six types of human action; walking, jogging, running, boxing, hand waving, and hand clapping, which are performed several times by 25 subjects under four different conditions: outdoors, outdoors with scale variation, outdoors with different clothing, and indoors. In all conditions, the background is homogeneous (Fig. 5(a)). We follow the original experimental setup of Schuldt et al. (2004) (*1-shot*): the sequences of 16 subjects are used for training, and the remaining sequences of 9 subjects are used for testing. The performance is evaluated based on the average accuracy over action categories.

Cambridge hand gesture: The Cambridge hand gesture dataset (Kim et al., 2007) contains nine hand gestures defined by three primitive hand shapes and three primitive motions, which are performed 10 times by two subjects under five different illumination conditions with a homogeneous background (Fig. 5(b)). We use the sequences acquired under the plain illumination condition for training and those under the remaining four conditions for the test.

Weizmann human action: The Weizmann human action dataset (Blank et al., 2005) contains nine types of human action; running, walking, jumping jacks, jumping forward, jumping in place, galloping sideways, waving two hands, waving one hand, and bending, which are performed once by each of nine subjects. The background has non-homogeneous texture (Fig. 5(c)). We evaluate the performance by employing the scheme of leave-one-subject-out in (Wang and Mori, 2009); the sequences of eight subjects are used for training and those of the remaining subject are for testing, which is repeated for all nine subjects.

RWC human gesture: The RWC human gesture dataset (Hayamizu et al., 1996) contains 17 types of human gesture; up, down, right, left, front, beyond, square, pointing at "this", this size, me, pointing left, pointing right, turn clockwise, turn counterclockwise, stop, expand, and reduce. Each of these gestures is performed four times by 48 subjects. This is a large dataset captured indoors with a homogeneous background (Fig. 5(d)). The performance is evaluated by using the leave-one-subject-out scheme.

Hollywood2 human action: The Hollywood2 human action dataset (Marszalek et al., 2009) contains 12 types of human action collected from 69 movie films (Fig. 5(e)); driving car, eating, fighting, running, answering the phone, getting out of the car, shaking hand, hugging, kissing, sitting down, sitting up, and standing up. In the motion images, multiple persons are shown against cluttered backgrounds with the large intra-class variability of action classes. This is a challenging dataset for action recognition. We follow the experimental protocol in (Wang et al., 2009): "clean" (manually annotated) training and test sets are used, and the performance is evaluated by mean average precision rate (mAP) over action categories.

For the proposed method, we used the parameter settings described as practical issues in this paper. Note that such parameter settings were empirically determined so as to produce favorable performances, especially on KTH dataset.

4.1. Performance analysis on the KTH dataset

First, using the KTH dataset, we analyze the performance of the proposed method in various settings.

Gradient: Gradient computation is the first processing step that may affect the final performance. We applied three types of 3-D derivative filter: Sobel, one-dimensional derivatives (1d-dev;

(a) KTH (Schuldt et al., 2004) (b) Cambridge (Kim et al., 2007)

(c) Weizmann (Blank et al., 2005) (d) RWC (Hayamizu et al., 1996)

(e) Hollywood2 (Marszalek et al., 2009)

Fig. 5. Example images of the datasets.

$[-1,0,1]$), and Roberts filters. As shown in Table 1(a), the 1d-dev filter is the most effective, as in HOG (Dalal and Triggs, 2005), whereas Roberts filter is least effective. The 1d-dev and Sobel filter, which are somewhat larger filters, can capture rather faster motions in contrast to the compact Roberts filter.

STACOG components: We evaluate the several types of STACOG components (Section 2.3.3): *dynamic* and *static* in the zeroth-order features, and *dynamic* × *dynamic*, *dynamic* × *static*, and *static* × *static* in the first-order features. Note that the number of static (orientation) bins in the STOC vector $\boldsymbol{h}$ is four, and the other bins are regarded as dynamic ones, as shown in Fig. 1(b). For the evaluation, we sequentially exclude static-related feature components, and the results are shown in Table 1(b). The right column shows the performance by using all types of feature components, while the left one shows the performance by only *dynamic* components. As expected, the features of *static* in the zeroth order and *static* × *static* in the first order, capturing only static figures, are redundant, thereby slightly worsening the performance, whereas those of *dynamic* × *static* contribute to the performance improvement by extracting the relationship of dynamic and static parts of the human body. The combination of *dynamic* components including *dynamic* × *static* is found to be optimal for motion recognition.

Spatial interval Δr: The spatial displacement interval Δr is dependant on the target scale. The performances for various Δr are shown in Table 1(c) as compared to the combination of three Δr optimally selected by zero-norm SVM (Weston et al., 2002) as described in Section 3.3. The result of the optimal combination is slightly superior to that of $\Delta r = 7$ which is the best result among $\Delta r = 1,\ldots,8$. More importantly, there is no need to manually tune Δr when we use the automatic selection.

Motion word: The effectiveness of the motion words in the FDA space is shown in Table 1(d), compared to standard motion words in original feature space. By applying FDA, the performance is greatly improved. Note that the computational cost for assigning motion words is also decreased due to the dimensionality reduction by FDA as described in Section 3.2.

Optical flow: In addition, we compare STACOG to the features of auto-correlations of optical flow that we newly propose here for the comparison. The optical flow, a popular method for motion feature extraction, is expressed as a two-dimensional vector at each pixel, and we code the vector into eight orientation bins as in Section 2.1. The local relationships of the optical flow pairs can be exploited by Eqs. (2) and (3) in which the coded flow orientations and the flow magnitude are substituted for the STOC vector $\boldsymbol{h}$ and the gradient magnitude m, respectively. We refer to this method as space–time auto-correlation of flows (STACOF). As shown in Table 1(e), STACOG using the space–time gradient is superior to STACOF. STACOF focuses only on movements (corresponding to *dynamic* × *dynamic* in STACOG) of subjects, whereas STACOG additionally consider *dynamic* × *static*, as described above. In addition, the magnitude of the flow is coded in the temporal elevation ϕ of the space–time gradients, while STACOF reduces the magnitude to just a weight for the auto-correlation. Moreover, the computed optical flow is not so reliable due to image noise, whereas STACOG is stable by using the gradient magnitudes as weights, which are roughly related to the confidence weights in the optical flow. The results in Table 1(e) indicate that the fundamental space–time gradients are more suitable for extracting motion features.

Running Time: The computational time for each process of the proposed method is shown in Table 1(f). The method requires only 0.3 s per video sequence, each of which contains about 95 frames (about 4 s at 25fps) on average, by using Xeon 3 GHz PC and Matlab (except for feature extraction implemented by mex-C). This is 10 times faster even than *real time*. From the qualitative viewpoint, it can be said that the method is faster than the other state-of-the-art methods based on the standard bag-of-features scheme because, as described in Section 3, the proposed method

Table 1
Performance analysis of the proposed method in various settings by using the KTH dataset (1-shot).

(a) Gradient computation

Filter	Sobel	1d-dev	Roberts
Accuracy (%)	94.1	**95.6**	91.8

(b) STACOG components

0th	*dynamic*	✓	✓	✓
	static			✓
1st	*dynamic* × *dynamic*	✓	✓	✓
	dynamic × *static*		✓	✓
	static × *static*			✓
Accuracy (%)		92.8	**95.6**	94.6

(c) Spatial interval Δr

Δr	Accuracy (%)
1	94.0
2	93.9
3	93.3
4	93.3
5	93.2
6	93.7
7	94.2
8	93.7
opt.	**95.6**

(d) Motion word

Motion word	in FDA	in original space
Accuracy (%)	**95.6**	82.2

(e) Optical flow

Method	Space–time gradient	Optical flow
Accuracy (%)	**95.6**	90.7

(f) Average running time per video sequence

Process	Feature	Motion word	Classify	Total
Time (s)	0.273	0.0406	0.000174	0.314

unnecessitates time-consuming detection of numerous interest points which consequently yields exhaustive processes for a large amount of local feature extractions and visual word assignments. The method could be made to perform several times faster still by using parallel or multi-thread programming since the processes using the three selected Δr are completely parallel for each Δr and the STACOG feature extraction process can also be parallelized for image data.

4.2. Comparison with the other methods

Next, we compare the performance of the proposed method to those of the other methods on the five datasets. Note that the same parameter setting is used in the proposed method over all these datasets, except that we use larger time grid interval of $R = 5$ for large datasets (KTH, RWC and Hollywood2) due to memory storage and use full eight channels of $\Delta r = 1$–8 for Hollywood2 to slightly improve the performance (∼1%). In addition, unlike in (Jhuang et al., 2007; Wang and Mori, 2009), preprocessing, such as background subtraction, was not applied to motion sequences. On KTH, we additionally employed the leave-one-subject-out (Table 2(a-2)) as well as 1-shot evaluation.

Table 2 shows the performance results; for the other methods, we show the performances reported in the reference papers. The proposed method yields favorable performances on most datasets, as compared to the other methods. Note that the proposed method exhibits the superior performance on the KTH dataset which is widely used to evaluate the performance of motion recognition.

Table 2
Comparison with the other methods on various datasets.

(a-1) KTH dataset (1-shot)

Method	acc. (%)
Schuldt et al. (2004)	71.7
Kobayashi and Otsu (2009)	88.0
Laptev et al. (2008)	91.8
Wang et al. (2009)	92.1
Ballan et al. (2009)	92.1
Ours	**95.6**

(a-2) KTH dataset (Leave-one-out)

Dollar et al. (2005)	81.2
Niebles et al. (2006)	81.5
Wong and Cipolla (2007)	86.7
Kobayashi and Otsu (2009)	90.7
Wang and Mori (2009)	91.2
Bregonzio et al. (2009)	93.2
Liu and Shah (2009)	94.2
Ours	**96.2**

(b) Cambridge dataset (%)

Method	Set1	Set2	Set3	Set4	Avg.
Niebles et al. (2006)	70	57	68	71	66
Kobayashi and Otsu (2009)	82	76	70	82	78
Kim et al. (2007)	81	81	78	86	82
Ours	**94**	**84**	**82**	**96**	**89**
Ours (sobel)	**94**	**87**	**91**	**95**	**92**

(c) Weizmann dataset

Method	acc. (%)
Scovanner et al. (2007)	82.6
Dollar et al. (2005)	86.7
Ballan et al. (2009)	92.4
Riemenschneider et al. (2009)	96.7
Kobayashi and Otsu (2009)	98.8
Jhuang et al. (2007)	98.8
Wang and Mori (2009)	**100**
Ours	**100**

(d) RWC dataset

Method	acc. (%)
Ishihara and Otsu (2004)	95.7
Kobayashi and Otsu (2009)	95.9
Ours	**98.1**

(e) Hollywood2 dataset

Method	mAP (%)
Kobayashi and Otsu (2009)	26.3
Marszalek et al. (2009)	35.5
Wang et al. (2009)	**47.7**
Ours	45.7

The proposed method also produces impressive results on the other datasets captured in various situations, such as textured backgrounds (Weizmann), large recognition classes (RWC), and illumination changes (Cambridge). In addition, these results show that the proposed method is applicable to low-resolution video sequences since all sequences are actually down-sampled to half size. These results demonstrate that the proposed method comprising *co-occurrence* histograms of space–time gradient orientation pairs and its (*dense*) frame-based features is quite effective for motion recognition, compared to the other gradient-based methods such as 3D-SIFT by Scovanner et al. (2007) and HOG+HOF by Laptev et al. (2008) which characterize motions by using sparsely extracted gradient-based local features, i.e., simple occurrence histogram of gradient orientations.

On the Hollywood2 datasets, the irrelevant information derived from the cluttered background is inevitably included in the frame-based features. The performance of the proposed method is degraded, but it is close to that of (Wang et al., 2009) (Table 2(e)) since FDA effectively suppresses the effects of such irrelevant information. The performance might be improved, for example, by employing the other discrimination method (e.g., mutual subspace method (Fukui and Yamaguchi, 2003)) or applying the STA-COG features to the local descriptors in (Wang et al., 2009).

5. Conclusion

We have proposed a method for extracting motion features and thereby have provided an effective and high-speed method of motion recognition. The proposed feature extraction method is based on local auto-correlations of the space–time gradients and effectively captures the geometric characteristics, such as curvatures, of space–time motion shape. The motion is recognized in the framework of bag-of-frame-features, which can sufficiently (densely) extract the motion characteristics in a computationally efficient manner, unlike standard bag-of-features which describes the motion sparsely. In experiments on motion recognition using various datasets, the proposed method exhibited favorable performances, as compared to the other methods. In addition, the results were obtained with less computational load and much faster than real time.

References

Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G., 2009. Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In: International Conference on Image Processing, pp. 3569–3572.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space-time shapes. In: International Conference on Computer Vision, pp. 1395–1402.

Bobick, A.F., Davis, J., 2001. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23 (3), 257–267.

Bosch, A., Zisserman, A., Munoz, X., 2007. Image classification using random forests and ferns. In: International Conference on Computer Vision.

Bregonzio, M., Gong, S., Xiang, T., 2009. Recognising action as clouds of space-time interest points. In: IEEE Conference on Computer Vision and Pattern Recognition.

Csurka, G., Bray, C., Dance, C., Fan, L., 2004. Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–22.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley-Interscience.

Fukui, K., Yamaguchi, O., 2003. Face recognition using multi-viewpoint patterns for robot vision. In: International Symposium of Robotics Research, pp. 192–201.

Gavrilla, D., 1999. The visual analysis of human movement: A survey. Computer Vision and Image Understanding 73 (1), 82–98.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Trans. Systems Man Cybernet. SMC 3 (6), 610–621.

Hayamizu, S., Hasegawa, O., Itou, K., Sakaue, K., Tanaka, K., Nagaya, S., Nakazawa, M., Endoh, T., Togawa, F., Sakamoto, K., Yamamoto, K., 1996. Rwc multimodal database for interactions by integration of spoken language and visual information. In: International Conference on Spoken Language Processing, pp. 2171–2174.

Huang, J., Kumar, S.R., Mitra, M., Zhu, W.-J., Zabih, R., 1997. Image indexing using color correlograms. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 762–768.

Ishihara, T., Otsu, N., 2004. Gesture recognition using auto-regressive coefficients of higher-order local auto-correlation features. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 583–588.

Jhuang, H., Serre, T., Wolf, L., Poggio, T., 2007. A biologically inspired system for action recognition. In: International Conference on Computer Vision.

Kim, T.-K., Wong, S.-F., Cipolla, R., 2007. Tensor canonical correlation analysis for action classification. In: IEEE Conference on Computer Vision and Pattern Recognition.

Kobayashi, T., Otsu, N., 2008. Image feature extraction using gradient local auto-correlations. In: European Conference Computer Vision, pp. 346–358.

Kobayashi, T., Otsu, N., 2009. A three-way auto-correlation based approach to motion recognition. Pattern Recognition Lett. 30 (3), 185–192.

Laptev, I., 2005. On space-time interest points. Internat. J. Comput. Vision 64, 107–123.

Laptev, I., Marszaek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition.

Liu, J., Shah, M., 2009. Learning human actions via information maximization. In: International Conference on Computer Vision.

Lowe, D., 2004. Distinctive image features from scale invariant features. Internat. J. Comput. Vision 60, 91–110.

Marszalek, M., Laptev, I., Schmid, C., 2009. Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936.

Niebles, J., Wang, H., Fei-Fei, L., 2006. Unsupervised learning of human action categories using spatial-temporal words. In: British Machine Vision Conference, pp. 1249–1258.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition.

Porikli, F., 2005. Integral histogram: A fast way to extract histograms in cartesian spaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 829–836.

Riemenschneider, H., Donoser, M., Bischof, H., 2009. Bag of optical flow volumes for image sequence recognition. In: British Machine Vision Conference.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local svm approach. In: International Conference on Pattern Recognition, pp. 32–36.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition. In: International Conference on Multimedia, pp. 357–360.

Tuytelaars, T., Schmid, C., 2007. Vector quantizing feature space with a regular lattice. In: International Conference on Computer Vision.

Vapnik, V. (Ed.), 1998. Statistical Learning Theory. Wiley.

Wang, Y., Mori, G., 2009. Human action recognition by semi-latent topic models. IEEE Trans. Pattern Anal. Mach. Intell. 31 (10), 1762–1774.

Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference.

Weston, J., Elisseeff, A., Scholkopf, B., Tipping, M., 2002. Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. 3, 1439–1461.

Willems, G., Tuytelaars, T., Gool, L.V., 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. In: European Conference on Computer Vision, pp. 650–663.

Wong, S.-F., Cipolla, R., 2005. Real-time interpretation of hand motions using a sparse bayesian classifier on motion gradient orientation images. In: British Machine Vision Conference, pp. 379–388.

Wong, S.-F., Cipolla, R., 2007. Extracting spatio-temporal interest points using global information. In: International Conference on Computer Vision.

Zelnik-Manor, L., Irani, M., 2006. Statistical analysis of dynamic actions. IEEE Trans. Pattern Anal. Mach. Intell. 28 (9), 1530–1535.