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

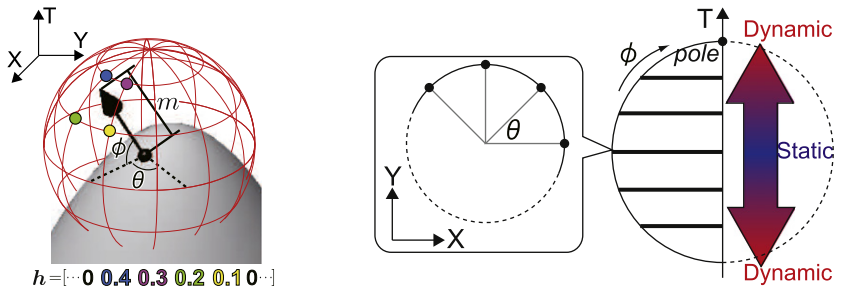
The Nth order auto-correlation function for the space–time gradients is defined by using the magnitude m and the STOC vector $\boldsymbol{h}$ of the gradients as follows:

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

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.)