Signal Processing: *Image Communication* 27 (2012) 96–111

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Human action recognition using Pose-based discriminant embedding

Behrouz Saghafi, Deepu Rajan *

Centre for Multimedia and Network Technology, School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 19 November 2010
Accepted 9 May 2011
Available online 27 May 2011

Keywords:
Human action recognition
Dimensionality reduction
Discriminant embedding
Silhouette
Posture
Action period

ABSTRACT

Manifold learning is an efficient approach for recognizing human actions. Most of the previous embedding methods are learned based on the distances between frames as data points. Thus they may be efficient in the frame recognition framework, but they will not guarantee to give optimum results when sequences are to be classified as in the case of action recognition in which temporal constraints convey important information. In the sequence recognition framework, sequences are compared based on the distances defined between sets of points. Among them Spatio-temporal Correlation Distance (SCD) is an efficient measure for comparing ordered sequences. In this paper we propose a novel embedding which is optimum in the sequence recognition framework based on SCD as the distance measure. Specifically, the proposed embedding minimizes the sum of the distances between intra-class sequences while seeking to maximize the sum of distances between inter-class points. Action sequences are represented by key poses chosen equidistantly from one action period. The action period is computed by a modified correlation-based method. Action recognition is achieved by comparing the projected sequences in the low-dimensional subspace using SCD or Hausdorff distance in a nearest neighbor framework. Several experiments are carried out on three popular datasets. The method is shown not only to classify the actions efficiently obtaining results comparable to the state of the art on all datasets, but also to be robust to additive noise and tolerant to occlusion, deformation and change in view point. Moreover, the method outperforms other classical dimension reduction techniques and performs faster by choosing less number of postures.

1. Introduction

Recognition of human actions from video is an important problem in computer vision that can take advantage of signal processing techniques. It has applications in a wide variety of topics including content-based video retrieval, human–computer interaction and surveillance. The problem is challenging especially in cases of intra-class variations in appearance and of different actions with similar postures.

Some of the common features used in action recognition are optical flow [1], space-time gradients [2], point trajectories [3] and sparse interest points [4–6]. Optical flow vectors are often inaccurate when the videos are of low quality and especially when motion is not smooth. Moreover, it is not robust to illumination changes. Likewise, point trajectories need accurate tracking methods which could fail in cases of fast moving subjects, occlusions and cluttered backgrounds. Also by using sparse representation of interest points, we lose the global structural information, which could otherwise help in the recognition process. On the other hand, silhouettes are informative features for describing actions [7–13].

* Corresponding author.
E-mail address: asdrajan@ntu.edu.sg (D. Rajan).

0923-5965/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2011.05.002

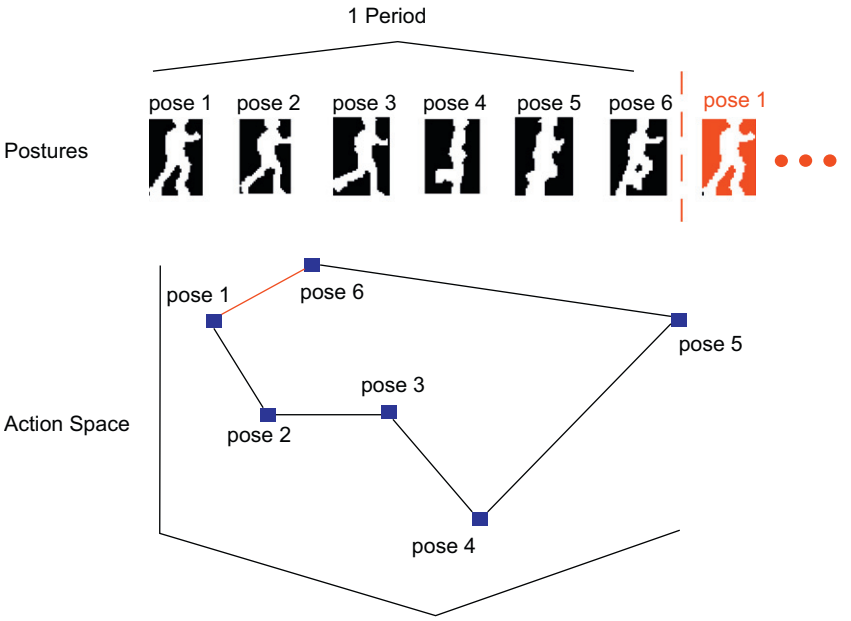

Fig. 1. Examples of postures for *run* and its trajectory in a possible action space.

They are able to capture the spatio-temporal characteristics of motion with possibly lower computational costs [7]. There is no need for an explicit model of the human body. Furthermore, recent advances in foreground extraction from complex backgrounds and in the presence of camera motion have benefited from improved models for segmentation and global motion extraction. In particular, segmentation algorithms have benefited from the area of alpha matting in which a pixel is composed of both background and foreground components and the problem is to estimate these components [14–16].

There have been two general frameworks for using silhouettes in action recognition. Some approaches classify action sequences on a frame-by-frame basis [8]. Thus each frame is independently classified as belonging to one of the actions. The label for the query sequence is obtained based on a voting scheme. All these approaches belong to the *frame recognition framework*. These methods ignore the temporal information and kinematics which is useful in the classification. Then, there are methods that classify the sequence as a whole [9]. These approaches belong to the *sequence recognition framework*. These methods compare sequences based on distances like Spatio-temporal Correlation Distance (SCD) or Hausdorff distance, which are defined between sequences of points. Human action, when represented as a sequence of silhouettes, can be considered as a function of time in which the silhouette of the body changes gradually. Thus, motion information is also included in this kind of representation without using expensive motion features that are difficult to extract. In this paper we utilize the latter framework.

Silhouettes can be considered as points in high-dimensional image space. Consequently, action sequences are described as data trajectories inside image space. Recognition methods which operate in this high-dimensional space suffer from the curse of dimensionality. In addition, the information provided in the high-dimensional image space is way more than required to describe an action. Moreover, the structure of the human body imposes a constraint on possible postures. Hence, it is more efficient to analyze action trajectories in a lower dimensional space. We call this subspace *the action space*. Examples of postures from the action *run* as well as the trajectory in a possible action space is shown in Fig. 1.

1.1. Motivation and overview of approach

There have been previous efforts at learning an efficient action space in order to classify actions. Accordingly, general dimension reduction techniques such as Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) [9], Locally Linear Embedding (LLE) [17], Laplacian Eigenmaps (LE) [18] and Kernel PCA [10], as well as action-specific embeddings like Local Spatio-Temporal Discriminant Embedding (LSTDE) [8] have been used. These methods are explained in more details in the next section. In all these methods, the embedding is defined based on the distance between data points rather than the distance between sequences; thus, they may be efficient in the *frame recognition framework*, but they are not guaranteed to give optimum results when *sequences* are classified. As stated earlier, in the sequence recognition framework, the query sequence is compared to the learned sequences based on distances generally defined between sets of data points, like SCD. In this paper we develop a novel embedding which is optimum in the sequence recognition framework. Specifically, the proposed embedding minimizes sum of the distances between intra-class sequences while maximizing the total distances between inter-class points.

SCD is an effective distance between ordered sequences. In order to compute SCD between two sequences, they need to have the same lengths. First, two sequences are shifted