J. Vis. Commun. Image R. 30 (2015) 22–34

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Human movement analysis around a view circle using time-order similarity distributions ☆

Chi-Hung Chuang [a], Jun-Wei Hsieh [b,*], Hui-Fen Chiang [b], Yi-Da Chiou [c]

[a] Dept. of Learning and Digital Technology, Fo Guang University, No.160, Linwei Rd., Jiaosi, Yilan 26247, Taiwan
[b] Dept. of Computer Science and Engineering, National Taiwan Ocean University, No.2, Beining Rd., Keelung 202, Taiwan
[c] Department of Electrical Engineering, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li 320, Taiwan

ARTICLE INFO

Article history:
Received 26 November 2013
Accepted 9 February 2015
Available online 19 March 2015

Keywords:
Video surveillance
Centroid context
Behavior analysis
View invariance
View circle
Time-order similarity distribution
Action classification
HMM
Symmetrical projection
View alignment

ABSTRACT

This paper presents a new behavior classification system to analyze human movements around a view circle using time-order similarity distributions. To maintain the view in-variance, an action is represented not only from its spatial domain but also its temporal domain. After that, a novel alignment scheme is proposed for aligning each action to a fixed view. With the best view, the task of behavior analysis becomes a string matching problem. One novel idea proposed in this paper is to code a posture using not only its best matched key posture but also other unmatched key postures to form various similarity distributions. Then, recognition of two actions becomes a problem of matching two time-order distributions which can be very effectively solved by comparing their KL distance via a dynamic programming scheme.

1. Introduction

The analysis of human body movements can be applied in a variety of application domains, such as video surveillance, video context retrieval, human–computer interaction systems, and medical diagnoses. In some cases, the results of such analysis can be used to identify people acting suspiciously and other unusual events directly from videos. In the past, many approaches [1]-[26] have been proposed for video-based human movement analysis. For example, Weinland and Boyer [3] proposed a clustering scheme to select a set of key postures for converting action sequences to different feature vectors. Then, a Bayesian classifier is designed to classify them to different event categories. Fathi and Mori [6] detected corner features as well as their motion vectors to represent an event, and then classified it to different event types. In [7], Ju et al. used the SIFT detector to extract different feature points and their trajectories as event descriptors, and then trained a SVM (Support Vector Machine) -based classifier for event classification. In [8], based on dynamic Bayesian networks, Wu et al. integrated RFID and video streams to model various kitchen events. In [9], the MHI (Motion History Image) feature was used to represent human motions and further classified to different event types by a SVM-based classifier. Furthermore, Kratz and K. Nishino [10] used a set of local spatio-temporal motion volumes to represent actions and then proposed a HMM-based framework to analyze the overall behavior of an extremely crowded scene. Some approach decomposes actions into sequences of key atomic action units which are referred to as atoms. In [11], Gaidon, Harchaoui, and Schmid proposed an actom sequence model (ASM) to represent the temporal structure of actions and then recognized actions in videos using a sequence of "atoms". Here, atoms which are specific to each action class should be identified by manual annotation.

In addition to videos, some approaches recognized human activities based on only still images. For example, in [12], Yao and Fei-Fei proposed a data-mining scheme to discover human-object interactions by encoding a set of highly related patches to grouplets with their appearances, locations, and spatial relations. Maji, Bourdev, and Malik [13] trained thousands of poselets to form a poselet activation vector within the bounding box of a person so that the related actions can be recognized from still images. However, the prerequisite that body parts or poses must be well estimated makes this model-based scheme inappropriate for real-time analysis of human behaviors.

☆ This paper has been recommended for acceptance by Prof. M.T. Sun.

* Corresponding author.
E-mail address: shieh@ntou.edu.tw (J.-W. Hsieh).

http://dx.doi.org/10.1016/j.jvcir.2015.02.003
1047-3203/

Another key problem in these approaches is "view invariance". Due to perspective effects, an action will look different from different viewpoints. Large appearance changes will lead to the failure in recognizing an action from different views. To maintain the view invariance, some researchers analyzed human actions using 3-D human models. For example, Shakhnarovichet al. [14] stored a dataset of human models with known 3-D parameters and then estimated best 3-D posture models for action analysis. Another approach for arbitrary view action analysis is to directly estimate 3D shapes and poses from multi-view 2-D image data. For example, Balan et al. [15] uses a triangulated mesh model called SCAPE [31] which employs a low-dimensional, but detailed, parametric shape model for recovering 3-D models directly from a multi-camera system. Weinland et al. [4,5] segmented the human actions into various primitives and clustered them to different action atoms using the feature "motion history volumes". The view invariance can also be achieved by building the connections between different views via a view alignment or matching technique. For example, in [16,17], Farhadi et al. used a codebook technique to construct split-based descriptors and then counted their frequencies to build view-invariant latent models for object recognition and activity analysis. Lv and Nevatia [18] applied the PMK (Pyramid Match Kernel) algorithm [19] to recognize postures and then matched each action from arbitrary views using the Viterbi [20] algorithm. In [21], Souvenir and Babbs used a manifold learning technique to learn a set of action primitives from a single camera to form a viewpoint-invariant representation. In [22], Junejo et al. built trajectory-based self-similarity matrices as pair-wise distances between 2D hand positions of persons for action recognition under different views. Furthermore, Cuzzolin [23] built different bilinear HMMs to analyze human gaits from multiple views. Karthikeyan, Gaur, and Manjunath [24] computed R transforms on action silhouettes and then proposed a probabilistic subspace similarity technique to recognize actions from multiple views by learning their inter-action and intra-action models. However, this scheme requires all action sequences from different views must be collected together before recognize an action.

In addition to 2-D and 3-D data, another trend for behavior analysis is to project the matching problem into a higher dimensional feature space. For example, Lee and Elgammal [25] projected features on a higher dimensional space and then used the techniques of singular value decomposition and manifold learning to model and analyze an articulated object around a view circle. Yan et al. [26] mapped an action to a 4-D model space and then derived its optimal model parameters to recognizing behaviors. Although 4-D features are more informative for human action recognition, their high computation cost makes this approach inappropriate for real-time applications.

This paper proposes a novel matrix-based scheme for recognizing human actions around a view circle using only 2D postures. The flowchart of our method is shown in Fig. 1. Two stages are included in our proposed system, *i.e.*, the training and recognition stages. At the training stage, we sample the viewing angle around the actor and then divide the action space to several subspaces. The basis of each action subspace is a set of key postures which are extracted using a cluster technique. To achieve this clustering task, the centroid context descriptor is constructed to describe a posture up to its syntactic levels. To maintain the view invariance, a larger set of action subspaces must be constructed and thus leads to the inefficiency in action analysis. To reduce the action space, this paper uses the mirror symmetry for reducing the whole view space to only its quarter. At the recognition stage, a novel view alignment scheme is then proposed to search the best view via the Viterbi algorithm. Thus, even though an action is captured from an unknown view angle, it still can be well analyzed and recognized. To avoid false matches and reduce the searching space, a table for recording the transition probability between any two key postures is built in advance at the training stage. Thus, two actions sequences can be compared by using not only their spatial similarities but also their temporal similarities. Once the best view is selected, each action can be converted to a string. However, many errors will happen during the converting process. The novelty of this paper is to code a posture (or frame) using not only its best matched key posture but also its similarities to other key postures. Then, recognition of an action taken from a view circle can be represented with a similarity distribution which changes along the time. After that, two actions can be matched by calculating the Kullback–Leibler distance between their time-order similarity distributions. The performance of the proposed method has been rigorously tested on a variety of behavior videos to prove its superiority in human behavior analysis from both efficiency and robustness perspectives.

The rest of this paper is organized as follows. In Section 2, the descriptor for posture representation will be described. The schemes for action space sampling and model construction are described in Section 3. Section 4 discusses the details of view alignment. The matrix-based representation for behavior analysis is discussed in Section 5. Section 6 reports experimental results and finally a conclusion will be presented in Section 7.

2. Posture representation using centroid context

To describe an action event, this paper uses the descriptor "centroid context" [27] to describe a posture up to its syntactic levels. The descriptor can tolerate large posture distortions than shape context [28]. In addition, its time complexity is much lower than

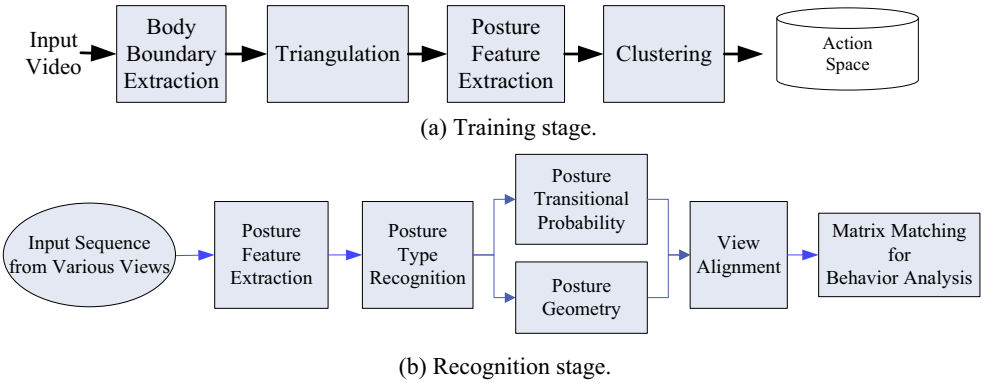

Fig. 1. Flowchart of the proposed system for recognizing behaviors from arbitrary views. (a) Training stage. (b) Recognition stage.