J. Vis. Commun. Image R. 30 (2015) 252–265

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Modeling and recognizing action contexts in persons using sparse representation ☆

Hui-Fen Chiang [a,b,1], Jun-Wei Hsieh [a,*], Chi-Hung Chuang [c], Kai-Ting Chuang [a,1], Yilin Yan [a,1]

[a] Department of Computer Science and Engineering, National Taiwan Ocean University, No. 2, Beining Rd., Keelung 202, Taiwan
[b] Dep. of D. M. Des, Taipei C. S. U. of S. T, Beitou, 112 Taipei, Taiwan
[c] Dept. of Learning and Digital Technology, Fo Guang University, No. 160, Linwei Rd., Jiaosi, Yilan 26247, Taiwan

ARTICLE INFO

Article history:
Received 9 November 2014
Accepted 20 April 2015
Available online 29 April 2015

Keywords:
Sparse coding
Sparse reconstruction error
Occlusions
R transform
Interaction action analysis
Behavior analysis
Person-to-person action recognition
Person-to-object action recognition

ABSTRACT

This paper proposes a novel dynamic sparsity-based classification scheme to analyze various interaction actions between persons. To address the occlusion problem, this paper represents an action in an over-complete dictionary to makes errors (caused by lighting changes or occlusions) sparsely appear in the training library if the error cases are well collected. Because of this sparsity, it is robust to occlusions and lighting changes. In addition, a novel Hamming distance classification (HDC) scheme is proposed to classify action events to various types. Because the nature of Hamming code is highly tolerant to noise, the HDC scheme is also robust to environmental changes. The difficulty of complicated action modeling can be easily tackled by adding more examples to the over-complete dictionary. More importantly, the HDC scheme is very efficient and suitable for real-time applications because no minimization process is involved to calculate the reconstruction error.

1. Introduction

Human action analysis [1,2] is an important task in various application domains like video surveillance [1–38], video retrieval [3], human–computer interaction systems, and so on. Characterization of human action is equivalent to dealing with a sequence of video frames that contain both spatial and temporal information. The challenge in human action analysis is how to properly characterize spatial–temporal information and then facilitate subsequent comparison/recognition tasks. To treat this challenge, some approaches build various action syntactic primitives to represent and recognize events. For example, in [4], Park and Aggarwal used the "blob" concept to model and segment a human body into different body parts from which human events were analyzed using the dynamic Bayesian networks. Instead of body parts, Wang et al. [9] used the *R* transform to extract contour features from different frames and then proposed a HMM-based recognition scheme to analyze human behaviors. The advantage of R-transform is its high tolerance to noise and incomplete contours of foreground objects. Furthermore, Kratz and Nishino [13] used a set of local spatio-temporal motion volumes to represent actions and then proposed a HMM-based framework to analyze the overall behaviors in an extremely crowded scene. Some approaches decompose actions into sequences of key atomic action units which are referred to as atoms. For example, in [36], Gaidon et al. proposed an atom sequence model (ASM) to represent the temporal structure of actions and then recognized actions in videos using a sequence of "atoms" which are obtained by manual annotations. In addition to videos, humans can recognize activities based on only still images. For example, in [42], Yao and Fei-Fei proposed a data-mining scheme to discover human-object interactions by encoding a set of highly related patches to grouplets according to their appearances, locations, and spatial relations. Maji et al. [35] trained thousands of poselets to form a poselet activation vector so that various human actions in still images can be recognized. However, the prerequisite that body parts or poses must be well estimated makes this scheme inappropriate for real-time analysis.

To avoid the difficulty of action primitive or body part extraction, some approaches extract feature points of interest and their motion flows to represent and recognize actions. For example, Fathi and Mori [37] detected and tracked corner features to obtain their motion trajectories for event representation, and then classified events to different layers using the Adaboost algorithm. In [38], Ju et al. used the SIFT detector and a tracking technique to extract different point trajectories with which different event

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author. Fax: +886 2 2462 3249.
[1] Fax: +886 2 2462 3249.

http://dx.doi.org/10.1016/j.jvcir.2015.04.012
1047-3203/

classifiers can be then trained by using the SVM classifier. In [16], Laptev et al. generated the concept of interest points (STIP) from images to flow volumes and then extracted various key frames to represent action events. The success of the above feature-flow methods strongly depends on a large set of well tracked points to analyze action events from videos.

In addition to event features, another key problem in event analysis is how to model the temporal and spatial dynamics of events. To treat this problem, Mahajan et al. [14] proposed a layer concept to divide the recognition task to three layers, *i.e.*, physical, logical, and event layers which correspond to feature extraction, action representation, and event analysis, respectively. In [15], Laxton et al. used dynamic Bayesian networks to model the interaction events between objects. Then, the Viterbi-like algorithm is adopted to seek the best path for event interpretation. In [10], Cong et al. used the sparse representation scheme to model and identify abnormal events from normal ones. Hidden Markov model (HMM) is another commonly used scheme to model event dynamics. For example, Oliver et al. [12] used HMMs to classify human interactions into different event types like meeting, approaching, walking, and so on. Furthermore, Messing et al. [17] tracked a set of corners to obtain their velocity histories and then used HMM to learn activity event models. Kim and Grauman [44] proposed a space–time MRF model to model the distributions of local optical flows within different regions and then estimated the degrees of event normality at each region to detect abnormal events. In addition, Nguyen et al. [39] developed a HHM-based surveillance system to recognize human behaviors in various multiple-camera environments such as a corridor, a staff room, or a vision lab. The challenges related to HMMs involve how to specify and learn the HMM model structure.

A particular action between two objects can vary significantly under different conditions such as camera views, person's clothing, object appearances, and so on. Thus, it is more challenging to analyze human events happening between two objects because of their complicated interaction relations. In addition, occlusions between objects often lead to the failure of action recognition. In [47], Filipovych and Ribeiro used a probabilistic graphical model to recognize several primitive actor-object interaction events like "grasping" or "touching" a fork (or a spoon, a cup). In [4], Ryoo and Aggarwal used a context-free-grammar framework that integrates object recognition, motion estimation, and semantic-level recognition to hierarchically analyze various interactions between a human and an object. In [46], Ping et al. extracted SIFT points and their trajectories to describe events and then used a kernel learning scheme to analyze interaction events caused by two objects. In addition to videos, some previous works address joint modeling of human poses, objects and their relations from still images. For example, in [40,41], Yao and Fei-Fei proposed a random field model to encode the mutual connections of components in the analyzed object, the human pose, and the body parts to recognize human-object interaction activities in still images. However, as described in [43], until now, reliable estimation of body configurations for persons in any poses remains a very challenging problem.

This paper addresses the problem of action analysis between persons (or human-object interactions) by using the sparse representation. As described before, the complicated interaction changes and the occlusion problem between two objects increase many challenges in action recognition. To treat the above problems, this paper proposes a novel dynamic sparse representation scheme to analyze interaction actions between persons. As we know, actors often perform the same action type at different speeds. From the nature of sparse representation, this temporal variation problem can be easily tackled by creating a new code in the training library to capture the speed change. From our knowledge, this is the first work using the sparse representation to analyze interaction actions between persons. The sparse property can make errors (caused by different environmental changes like lighting or occlusions) sparsely appear in the training library and thus increase the robustness of our scheme to environmental changes. In the literature [57], the sparse reconstruction cost (SRC) is usually adopted to classify data to different categories. The calculation of SRC is very time-consuming and unsuitable for real-time applications. Therefore, the second contribution of this paper is to propose a Hamming distance classification (HDC) scheme to classify data without calculating the SRC. Thus, our method is more efficient and suitable for real-time applications. In addition to the efficiency improvement, the HDC scheme is also robust to occlusions because the nature of Hamming code is highly tolerant to noise. The main contributions of this work include:

1. A novel sparsity-based framework is proposed to analyze person-to-person interaction behaviors. Even though the interaction relations are complicated, our method still works successfully to recognize them.
2. A new Hamming distance classification scheme to classify events to multiple classes. It is robust to environmental changes due to the nature of Hamming code.
3. The proposed sparse representation is very efficient and suitable for real-time applications because no minimization process to calculate the SRC is involved.

The remainder of the paper is organized as follows. Section 2 discusses some related work. An overview of our system is introduced in Section 3. Details of feature extraction are described in Section 4. Section 5 discusses the techniques of sparse representation and library learning. The framework of person-to-person interaction action recognition using sparse representation is proposed in Section 6. The experiment results are given in Section 7. We then present our conclusions in Section 8.

2. Related work

There have been many approaches [8,9,13–18,42,44] proposed in the literature for video-based human movement analysis. For example, in [52], Fengjun and Nevatia used a posture-based scheme to convert frames to strings from which actions were then classified into different types using the action nets. Weinland and Boyer [53] selected a set of key postures to convert an action sequence to a vector. Then, the Bayesian classifier is designed to classify the extracted vectors to different categories. In [54], Hsieh et al. proposed a centroid context descriptor to convert postures to different symbols and then detected abnormal events directly from videos. Messing et al. [55] used the velocity history of tracked key-points to represent actions and then combined other features, such as appearance, position, and high level sematic features together to recognize activities of daily living. In [56], Fan et al. modified the Viterbi algorithm to detect repetitive sequential event units and then recognized the predominant retail cashier activities at the checkout station.

To model the interactions between objects, in [45], Wu et al. integrated RFID and video streams to analyze the events happening in kitchen via a Dynamic Bayesian network. In [43], Delaitre et al. proposed a statistical model which integrates order-less person-object interaction responses to recognize actions in still images by calculating spatial co-occurrences of individual body parts and objects. To achieve the property of view invariance, Karthikeyan et al. [8] computed the R transforms on action silhouettes from multiple views and then proposed a probabilistic subspace similarity technique to recognize actions by learning their