Pattern Recognition 48 (2015) 2346–2363

Contents lists available at ScienceDirect

Pattern Recognition

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

Anatomical-plane-based representation for human–human interactions analysis

Rami Alazrai [a,*], Yaser Mowafi [a], C.S. George Lee [b]

[a] School of Computer Engineering and Information Technology, German Jordanian University, Amman 11180, Jordan
[b] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 2 November 2013
Received in revised form
19 February 2015
Accepted 3 March 2015
Available online 11 March 2015

Keywords:
Human–human interaction classification
Human–human interaction prediction
Bag of words
Motion-pose geometric descriptor

ABSTRACT

In this paper, we present a novel view-invariant, motion-pose geometric descriptor (MPGD) as a human–human interaction representation to capture the semantic meaning of body-parts between two interacting humans. The proposed MPGD representation is based on utilizing the concept of anatomical planes to construct a motion profile and a pose profile for each human. Those two profiles are then concatenated to form a descriptor for the two interacting humans. Using the proposed MPGD representation, we study two problems related to human–human interaction analysis, namely human–human interaction classification and prediction. For the human–human interaction classification problem, we propose a hierarchical classification framework consisting of a representation layer and three classification layers. The classification framework aims to realize what is the performed interaction in an input video by understanding how and when each individual performed sub-activities to each other over time. The human–human interaction prediction problem aims to predict the class of ongoing human–human interaction at its early stages. To do so, we propose a prediction framework that utilizes the proposed MPGD to construct an accumulated histograms-based representation for an ongoing interaction. The accumulated histograms of MPGDs are then used to train a set of support-vector-machine classifiers with a probabilistic output to predict the class of an ongoing interaction. In order to evaluate our proposed MPGD representation and both the classification and the prediction frameworks, we utilize a Microsoft Kinect sensor to capture human–human interactions in a video dataset that consists of 12 interactions performed by 12 individuals. We evaluate the performance of our proposed classification framework and compare the results with an appearance-based representation and a representation that combines both the MPGD representation and the appearance-based representation. On the one hand, our proposed MPGD representation performance has shown promising results compared to the appearance-based representation with an average accuracy of 94.86% in classifying human–human interactions. On the other hand, human–human interaction prediction framework has achieved an average prediction accuracy of 82.46% with only 50% of the interaction video being observed.

1. Introduction

Understanding and interpreting the semantic meaning of events that involve human–human interactions are essential in many domains such as robotics, video surveillance, video annotation and indexing. For example, endowing robots with the ability of understanding and interpreting human–human interactions will enable those robots to become versatile and adept in interacting with and assisting humans' daily life activities. However, analyzing human–human interactions is still at its infancy with most of the existing human activity representations (e.g., spatiotemporal interest points [1,2]) suffer from the limitation of capturing the semantic meaning of human body-parts movements. Hence, causing such representations to be more suitable for recognizing simple human actions [2–6] rather than complex human activities (e.g., human–human interactions) that exhibit high similarity in human body-parts movements. Furthermore, several human–human interactions may involve one human occluding other human body-parts, causing considerable ambiguity of human–human interaction representation. In addition, most of the existing approaches do not provide semantic analysis of the spatiotemporal events being displayed in a human activity video (e.g., who did what to whom and when a specific event happened). Instead, the entire human activity video is typically classified to one action class. By and large, the proposed human activity recognition systems are using either 2D video or

* Corresponding author. Tel.: +962 798213151.
E-mail addresses: rami.azrai@gju.edu.jo (R. Alazrai), yaser.mowafi@gju.edu.jo (Y. Mowafi), csglee@purdue.edu (C.S. George Lee).

http://dx.doi.org/10.1016/j.patcog.2015.03.002
0031-3203/

motion-capturing system as an input [7,8] to extract human joint positions. While the use of 2D videos makes the approaches sensitive to the occlusions, clutter background, shadow, variation in illumination, and view-point changes which can lead to low accuracy results. The use of motion-capturing systems via mounting sensing devices on the people can solve the above problems, but the expensive cost of such equipment limits its usability in this field.

That said, the focus of this study is two-fold. The first focus is mainly concerned with human–human interaction representation, in which we propose a novel view-invariant, motion-pose geometric descriptor (MPGD) as a human–human representation that can be used to capture the semantic meaning of body-parts for two interacting humans. The proposed representation is based on utilizing the concept of anatomical planes [9] to construct a motion profile and a pose profile for each human. Those profiles are then concatenated to form a descriptor for the two interacting humans. Such a representation can provide solutions to the first and the third challenging problems mentioned above.

The second focus of this study is concerned with human–human interaction analysis in which we aim to analyze the interactions between two humans in an input video from two different perspectives: (a) human–human interaction classification, and (b) human–human interaction prediction (Fig. 1). The key difference between the two analyses is the progress level of the observed video at the moment we perform the classification or the prediction process. Such that, in human–human classification (Fig. 1a), for each human–human interaction we build a classification model using a set of exemplar videos. Then, for any new exemplar, the entire video is observed prior to be classified using the learned interaction models. In human–human prediction (Fig. 1b) for each human–human interaction, we build a predication model using a set of exemplar videos. Then, for any new exemplar, the class of the ongoing interaction is predicted as early as possible without waiting to observe the entire video.

Interactions between two humans can be viewed as a set of temporal coherent phases in which each temporal phase represents a sub-activity performed by one individual to the other. For example, Fig. 1a shows a hugging interaction in which both the interaction model video and the observed video consist of five temporal phases. In the first phase, both humans are in a stand-still pose. Then, in the second phase, both humans start stretching out and raising their both arms toward each other. In the third phase, both humans have their arms behind the back of each other and their chests and heads as close as possible to each other. In the fourth phase, both humans start withdrawing and moving their arms downward. In the last phase, both humans are back to the stand-still pose. Therefore, recognizing and localizing these phases over the time enable to understand the overall human–human interaction performed during the phases of the entire video. *This is the main focus of our study for human–human interaction classification*. Specifically, given a video that is displaying a human–human interaction, we attempt to answer the following questions: what is the overall interaction displayed in the video and how and when each individual performed different sub-activities over time. Ultimately, the answers to: *what* is the performed interaction, *how* and *when* each individual performed each sub-activity of the observed interaction may provide an understanding and interpretation of human–human interactions in the input video. Thus, the human–human interaction classification reduces to the problem of answering these three questions about human–human interactions, we denote these questions as $\mathbf{Q^3}$. Thus, we propose a classification framework that is capable of answering the $\mathbf{Q^3}$ questions from RGBD input data captured by a Microsoft Kinect sensor.

The proposed human–human-interaction classification framework can be viewed as a hierarchical system consisting of a representation layer and three classification layers. At the representation layer, RGBD images are acquired from a Kinect sensor, and the 3D joint positions of each human are estimated. Since the interaction sequences may be captured from different views, we utilize the proposed MPGD to describe each input frame. With the MPGD representation of human–human interaction, we train a set of support-vector-machine (SVM) classifiers to classify each frame into one of the different states at the first classification layer. The state of each frame describes the spatiotemporal configuration of the two persons in that frame (e.g., both persons are at stand still position, or one person is stretching out his right arm while the other is stand still). At the second classification layer, the input video is segmented based on aggregating consecutive frames with similar predicted states into consecutive temporal phases. Each temporal phase is a collection of consecutive frames that have similar spatial configuration; hence temporal phases describe the sub-activities involved in an interaction. As the interaction video is temporally segmented, we can determine the occurrence of each sub-activity. The resultant temporal phases allow us to answer both the second and the third questions. Finally, at the third classification layer, the constraint dynamic time warping (cDTW) is utilized to provide an answer for the first question. This is done by classifying the entire sequence of states generated from the SVM classifiers into one of the 12 different human–human interactions. Using cDTW for time-series matching allows dealing with the large variation in the duration of interaction video sequences more efficiently.

Unlike human–human interaction classification, several real-world scenarios require to predict the class of the ongoing interaction as early as possible. For example, in surveillance applications, predicting a fighting activity by predicting a kicking or punching interactions can help preventing anticipated violence activities. *This is the second focus of this paper on human–human interaction analysis*. Motivated by the work of [2,4,6,10], in which

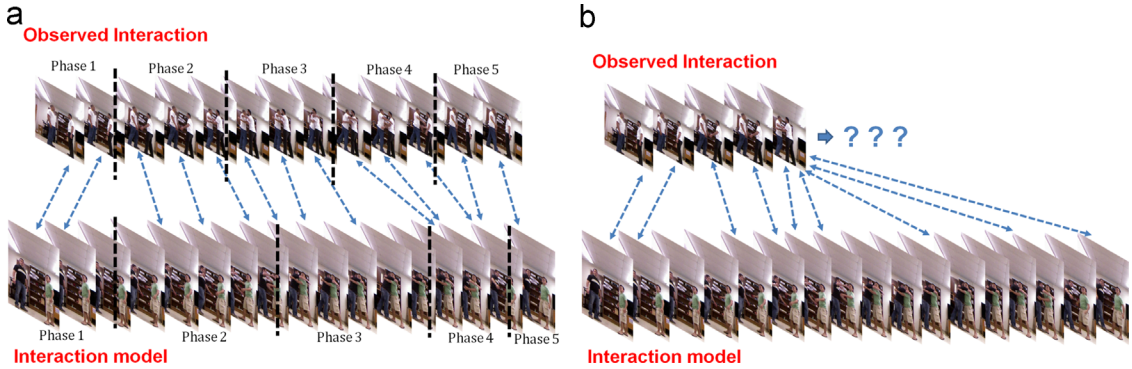

Fig. 1. Human–human interaction analysis. (a) Human–human interaction classification and (b) Human–human interaction prediction.