Computer Vision and Image Understanding 144 (2016) 155–165

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Affective interaction recognition using spatio-temporal features and context

Jinglian Liang[a], Chao Xu[b,*], Zhiyong Feng[a], Xirong Ma[c]

[a] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[b] School of Computer Software, Tianjin University, Tianjin 300072, China
[c] School of Computer Science and Technology, Tianjin Normal University, Tianjin 300387, China

ARTICLE INFO

Article history:
Received 15 December 2014
Accepted 13 October 2015

Keywords:
Interaction recognition
Affective computing
Spatio-temporal feature
Primary visual cortex
Context

ABSTRACT

This paper focuses on recognizing the human interaction relative to human emotion, and addresses the problem of interaction features representation. We propose a two-layer feature description structure that exploits the representation of spatio-temporal motion features and context features hierarchically. On the lower layer, the local features for motion and interactive context are extracted respectively. We first characterize the local spatio-temporal trajectories as the motion features. Instead of hand-crafted features, a new hierarchical spatio-temporal trajectory coding model is presented to learn and represent the local spatio-temporal trajectories. To further exploit the spatial and temporal relationships in the interactive activities, we then propose an interactive context descriptor, which extracts the local interactive contours from frames. These contours implicitly incorporate the contextual spatial and temporal information. On the higher layer, semi-global features are represented based on the local features encoded on the lower layer. And a spatio-temporal segment clustering method is designed for features extraction on this layer. This method takes the spatial relationship and temporal order of local features into account and creates the mid-level motion features and mid-level context features. Experiments on three challenging action datasets in video, including HMDB51, Hollywood2 and UT-Interaction, are conducted. The results demonstrate the efficacy of the proposed structure, and validate the effectiveness of the proposed context descriptor.

1. Introduction

Expression of emotion is part of actions in human interaction [1]. Due to the complexity of the body and lack of formal model for body expressions, the features are difficult to extract [2]. Recently, the majority of affective computing research mainly focuses on facial expressions or speech to make computer understand human emotions. However, it has been demonstrated that body expression plays an important role in emotion understanding as well. In psychology, it is also proved that body expression is as valuable as facial expression to emotion analysis [3]. A psychological research demonstrates the importance of body expressions, like body motion, posture, gestures and actions in human interaction [11]. The importance of body expressions is reflected in two aspects. On the one hand, the changes in body posture reflect changes in affective state [30]. Ekman and Friesen [19,63] propose that postural changes due to affective state endow a person with the ability to cope with the experienced affective state. On the other hand, bodily configuration and orientation are significantly affected by individuals attitude toward his interactive partner [31]. Therefore, some research [4–6] present that using body expression to analyze emotional states will be able to enhance the analysis of emotion. Body postures can help us discriminate not only the emotion, especially the emotion of fear, but also the cause of emotion and the action tendency that will be carried out [7]. The emotional states can be discerned by body expressions much better over long distances, whereas facial features are difficult or unreliable to be used to recognize the emotion at the same distance [8]. In addition, a lot of social emotions are frequently expressed by interactions. For example, loving and congratulating can be easily discriminated by body expressions, such as kissing or clapping, during interaction [7,9,10]. And in the affective interaction, most of the emotional actions are expressed through upper body, such as the head, arms and torso. For example, shaking hands in interaction can express greeting; The emotions to comfort another distressed person can be expressed by hugging, such as a parent giving a hug with a crying child or someone giving a hug with his friend who is troubled with something; Clapping can transmit the information of congratulation or admiration to someone; And striking, which is an open act of

[*] Corresponding author.
E-mail address: xuchao@tju.edu.cn (C. Xu).

http://dx.doi.org/10.1016/j.cviu.2015.10.008
1077-3142/© 2015 Elsevier Inc. All rights reserved.

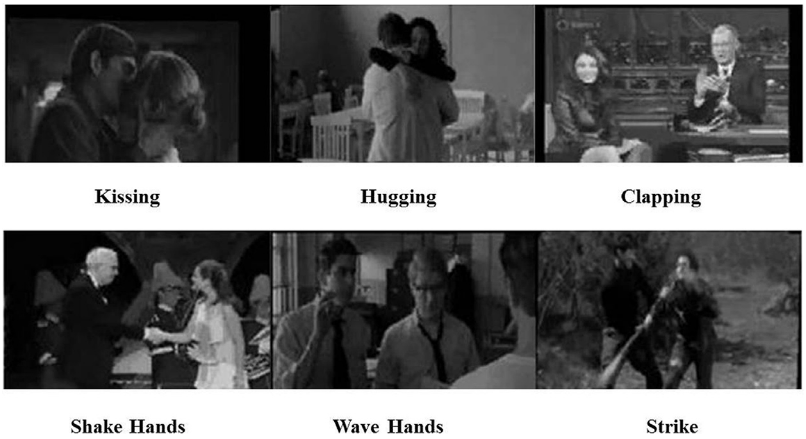

Fig. 1. Interactive body expressions with emotion.

Table 1
Illustration of complex emotions expressed in interaction.

Emotion	Body expression	Scene
Love	Kissing	Two persons kiss with each other, given as a sign of affection
Comfort	Hugging	A person hugs with the other who is troubled with something for comfort
Congratulation	Clapping	A person claps to someone who is excellent in something
Appreciation	Shaking hands	A person shakes hands with the other for his giving
Greeting	Waving hands	A person waves hands for greeting with someone who is coming
Aggression	Striking	A person hits others with anger

aggression by hitting someone, is associated with anger. Motivated by the above analysis, we focus on recognizing the interactive activities associated with emotion from videos in real life, such as hugging, kissing, etc. The illustration for these interactive activities is shown in Fig. 1. Therefore, in this paper, we mainly study on the six types of interactions related to complicated emotions. These interactions are mainly selected in specified scenes. The illustration of emotions, corresponding interactions and scenes is demonstrated in Table 1.

In order to study on the understanding of complicated emotions, we lay emphasis on the action recognition. Action recognition in video is one of the most challenging problems in computer vision [44–47]. Specifically, video representation remains to be an important issue in action recognition. Recently, various approaches have been proposed and achieved great progresses. Inspired by the spatio-temporal nature of actions in video, many works exploit local spatio-temporal feature descriptors combining with Bag-of-Words methods [12–14,22] for action recognition in video. Such approaches are effective and robust to viewpoint and scale changes in unconstrained scenarios for simple actions, and have achieved promising performance. However, they largely ignore the spatial and temporal order of the action among the local spatio-temporal features. In order to solve this problem, in this paper we propose a novel feature representation structure for interaction recognition. In this structure, the spatial and temporal relationships are introduced into action representation, and two layers are used to represent the interaction on the low-level and mid-level respectively. On the first layer, a low-level representation is built using a new hierarchical coding model for the local features. The second layer is built on top of the first layer. On this layer, the mid-level representation in spatial and temporal order are exploited through a Spatio-Temporal Segment Clustering method.

From another aspect, deep architectures for automatic features learning greatly inspire researchers to explore this direction for their promising results in action recognition [49–52]. Similar to deep learning architectures, a lot of biological models are proposed to infer high-level features from the lower-level, so as to learn features automatically. They have been proved to be effective with not so deep structures, and improvement has also been observed in these models for object recognition [36]. Inspired by these works, we propose a biological-motivated Hierarchical Sptio-Temporal Trajectory Coding Model to represent the low-level spatio-temporal motion features in our structure with lower computation costing but higher discrimination of local features. This spatio-temporal motion feature coding method is exploited by extending from a contour coding model for natural images. We advance this model to the spatial and temporal scales, and use it to represent the local motions extracted from actions. And these local motions can be further described as local trajectories. Finally, all these local trajectories make up the whole interaction.

Recently, context information becomes significant for action recognition. For example, identifying the objects involved in the context of an activity can improve the performance of recognition [27]. In the same way, as for the human interaction, the actions of each person related in space and time in interaction rarely occur independently and can be exploited as the context for each other as well. Take “hugging” and “clapping” for instance. The arm movements can be detected from each person respectively in the interactions. However, recognizing from anyone independently cannot help us to understand the interactive activity definitely. Thus, it is not enough to recognize the activities without the context, especially the interactions. On this direction, the role of objects can serve as the context for learning to improve the performance in some approaches [53–55]. Although these approaches have achieved good performance, most of them tend to be complicated and time consuming, because they have to do the preprocessing such as segmentation or distance extraction, etc. This makes us to consider the problem of designing a simple but effective context description method. In this paper we incorporate the context information into the spatio-temporal action features to improve the classification performance. Rather than modeling context information in videos respectively, we extract interactive contours from action sequence as the context, which includes a lot of information. Although the context properties are not extracted respectively as the previous methods, such as the changing relative distance or the concurrent posture states between persons, etc., the interactive contours implicitly incorporate these context properties.