Pattern Recognition Letters 73 (2016) 44–51

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Facial descriptors for human interaction recognition in still images☆

Gokhan Tanisik, Cemil Zalluhoglu, Nazli Ikizler-Cinbis*

Department of Computer Engineering, Hacettepe University, Ankara 06800, Turkey

ARTICLE INFO

Article history:
Received 22 September 2015
Available online 22 January 2016

Keywords:
Human interaction recognition
Facial features
Interaction recognition in still images

ABSTRACT

This paper presents a novel approach in a rarely studied area of computer vision: Human interaction recognition in still images. We explore whether the facial regions and their spatial configurations contribute to the recognition of interactions. In this respect, our method involves extraction of several visual features from the facial regions, as well as incorporation of scene characteristics and deep features to the recognition. Extracted multiple features are utilized within a discriminative learning framework for recognizing interactions between people. Our designed facial descriptors are based on the observation that relative positions, size and locations of the faces are likely to be important for characterizing human interactions. Since there is no available dataset in this relatively new domain, a comprehensive new dataset which includes several images of human interactions is collected. Our experimental results show that faces and scene characteristics contain important information to recognize interactions between people.

1. Introduction

In the last decade, human action recognition has been a very active research area in computer vision due to its various potential applications. A large body of work is dedicated to recognizing singleton activities in videos, whereas some recent work focus on recognizing singleton actions within still images. Human interaction recognition, which constitutes up a significant subset of multi-person activities, is a relatively less studied area. Especially for still images, the prior work is almost non-existent. In this paper, we address this problem of multi-person interaction recognition in images.

When recognizing interactions in still images, the problem gets more complex and harder to solve, due to the explicit need to discriminate foreground from background clutter without the motion information. In videos, motion is shown to be a great clue for identifying the type of the interactions [9,20]. Without motion, the foremost cue becomes the appearance. In this paper, we explore how we can extract and leverage multiple forms of appearance information for interaction recognition in images.

In this context, we propose several novel visual features that captures the intrinsic layout and orientation of face regions. Faces tend to play a great role in characterizing human interactions. People look at each other when they are talking, faces come together when people are kissing, and more. Fig. 1 includes some examples. Fathi et al. [6] use faces in video sequences to describe interactions in first-person (egocentric) videos. Based on inspiration from this work, we explore whether facial features can also be helpful in discriminating multi-person interactions in still images.

Another reason to explore facial features is that, face detection technology is considerably advanced and is able to locate a great deal of faces in images, especially those that are not too small or significantly occluded. Our designed descriptors are based on the observation that relative positions, size and locations of the faces are likely to be important for characterizing human interactions. To extract these descriptors, we first use a face detector. We also estimate the orientations of the faces if possible using the face detector of [25]. In this way, we estimate the size, spatial location and the orientation of the face in a $[-90°, 90°]$ range with 15° resolution. We then make use of these features to propose image-level facial interaction descriptors. For recognition, we combine these multi-person facial descriptors with standard scene descriptors extracted globally from the images. In this context, we also investigate the effect of the state-of-the-art deep learning based features, aka, Convolutional Neural Network (CNN) features to this new problem domain.

Since there is no available dataset in this relatively new domain, a new and comprehensive dataset, which includes a total of 10 human interaction classes, such as boxing, dining, kissing, partying, talking, is collected. This dataset has also been enriched with the manual annotations of the ground truth face locations and orientations to facilitate further comparisons.

Our contributions in this paper are two-fold: (1) we collect a new image dataset for human interaction categorization which

☆ This paper has been recommended for acceptance by Dr. A. Fernandez-Caballero.

* Corresponding author. Tel.: +90 312 297 7500; fax: +90 312 297 7502.
E-mail address: nazli@cs.hacettepe.edu.tr, nazli.ikizler@gmail.com (N. Ikizler-Cinbis).

http://dx.doi.org/10.1016/j.patrec.2016.01.002
0167-8655/

Fig. 1. Faces in interactions. The original images are shown in bottom row and the facial regions are shown on top. In this paper, we explore whether we can predict the type of human interactions in an image based on descriptors extracted from faces and their spatial layout. As it can be seen from this figure, without using any context or scene information, recognizing interactions by only face information can be quite a difficult task even for humans. Our results show that, using facial descriptors together with global scene descriptors yield promising results for human interaction recognition in still images.

includes multi-person interaction instances and (2) we present novel descriptors based on facial regions for human interaction recognition. Our experimental results show that, deep learning based features are effective in recognition of human–human interactions in images, and the proposed facial features that aim to encode the relative configurations of faces also provide useful information, especially when combined with global image features. In the rest of the paper, we first give a brief overview of the related work in Section 2. In Section 3, we introduce our proposed facial descriptors. Section 4 presents the experimental results, and Section 5 concludes the paper with the major findings of our evaluations.

2. Related work

There is a vast literature on human action/activity recognition in videos(for a recent survey, see [1]), whereas human interaction recognition is a relatively less studied topic. For human interaction in videos, a number of studies work on human–object interactions [7], whereas some studies work on human–human interactions [8,12,15].

In image domain, human–object interactions are the focus of a number of studies, which handle the problem by extraction of distinctive feature groups [22], by bag-of-features and part-based representations [3] and by weakly supervised learning [16]. Object–person interactions have also been explored in [4,5,23].

One of the earliest works to recognize the human interactions in still images is the work of [14]. In their paper, four classes are defined: shaking hands, pointing at the opposite person, standing hand-in-hand and intermediate-transitional state between them, and K-nearest neighbor classifier is used to recognize the interactions. Recently, Yang et al. [21] have focused on how people interact by investigating the proxemics between them. They claim that complex interactions can be modeled as a single representation and a joint model of body poses can be learned. Ramanathan et al. [17] look into the problem of detecting social roles in videos in a weakly supervised setting via a CRF model. In our work, we approach the human–human interaction recognition problem by means of several descriptors that encode facial region configurations.

Another study area that could be related to our work is event recognition in still images [2,11,18]. Event recognition research aims to recognize a certain scene or event in images or videos. Datasets in this field are different from ours. Event recognition datasets describe an event like Christmas, wedding, etc. In such images, the main focus is not the people, but visual elements for an event. In multi-person interaction recognition problem, we focus on the presence of people, and try to infer the interaction based on images of people.

In this work, we are inspired from the recent work of Fathi et al. [6], which uses face detection responses to recognize social human interactions in video sequences from a first person perspective camera. They propose to use Markov Random Field for frame based feature representations and a Hidden Conditional Random Field to represent sequence based features. In our work, we propose several simple features based on face regions for recognizing human–human interactions in the images.

3. Our approach

In this section, we describe the facial descriptors and the learning procedure that we have proposed for the purpose of interaction recognition.

3.1. Visual features for human interaction recognition

Our approach begins with the detection of the faces. For this purpose, we first apply the recent algorithm of [25], since it outputs three essential information about the faces: (1) orientation of the face in the range of [−90°, 90°] with a resolution of 15°, (2) location of the face in the image and (3) size of the face in pixels. The orientation of a face is defined as the angle of the face with respect to the imaginary axis that crosses from the midpoint of the chin and the forehead. For reducing the number of false negatives in face detection, we also employ the OpenCV implementation of [19], which only gives the location and size of the images, and whether they are frontal or profile. The face detections from these two approaches are combined in the following way: (1) If both of the detectors find a face in the same region, [25]'s output is used, since it is shown to be more accurate and it outputs face orientation estimates as well as face locations. (2) While using Viola–Jones face detector, if only frontal face is detected, the orientation is assumed to be 0°. If a profile face is detected, then the orientation is assumed to be 90°. If both frontal and profile face detectors fire within the same region, it means that the orientation of the face is between [0°, +/−90°]. To quantize the angle, the intersection ratio is normalized to [0°, −90°] interval.

After detecting the faces, we extract several mid-level descriptors based on the facial regions. Below, we introduce each of these descriptors.

Histogram of Face Orientations (HFO). In order to account of the distribution of the face orientations, we propose to use Histogram of Face Orientations (HFO), which simply is based on the count of face orientations for each angle in an image. In another words, it is the distribution of face orientation frequencies in an image. This descriptor has 13 feature dimensions (13 histogram bins), which corresponds to 15° resolution in [−90°, 90°] interval. Fig. 2 shows some example HFO descriptors.