Computers in Industry 99 (2018) 282–293

Contents lists available at ScienceDirect

Computers in Industry

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

Automated multi-feature human interaction recognition in complex environment

Shafina Bibi[a], Nadeem Anjum[b,*], Muhammad Sher[a]

[a] Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan
[b] Department of Electronic Engineering, University of Engineering and Technology, Taxila, Pakistan

ARTICLE INFO

Article history:
Received 30 August 2017
Received in revised form 7 March 2018
Accepted 15 March 2018
Available online xxx

Keywords:
Median compound local binary pattern
Individual action recognition
Trajectory
Person-to-person interaction recognition

ABSTRACT

The recognition of people's interactions is crucial for making the surveillance applications able to recognize unusual events in complex environments. Generally, multiple cameras are installed to capture videos from different views but these environments suffer from challenging issues: occlusions between persons and light and pose variations etc. We presented a computer vision system to recognize person-to-person interactions in public areas by considering individual actions and trajectory information under multiple camera views. We achieved our goal in two steps, namely, individual action recognition and interaction recognition. Extensive techniques have been used for individual action recognition with very good accuracy. Still, these techniques cannot handle the intricate settings in crowded areas. We have proposed Median Compound Local Binary Pattern (MDCLBP) and combined it with Histogram of Oriented Gradient (HOG). MDCLBP captures the information about the spatial organization of intensities and HOG uses histogram of oriented gradients to describe an image. MDCLBP is a modification of Compound Local Binary Pattern (CLBP). CLBP extracts texture information by using sign and magnitude information. MDCLBP is a variant of CLBP that uses sign information and instead of magnitude, difference from the median value at each 3×3 windows is used to get the descriptor robust to occlusions and light variations. We have combined the individual actions of two persons with trajectory information to recognize person-to-person interactions. Experiments are performed on well-known publically available IXMAS and OIXMAS datasets to demonstrate the effectiveness of our proposed technique for individual human action recognition. Person-to-person interaction recognition method is evaluated on HALLWAY dataset. Experiments carried out on varying views demonstrated that our proposed system achieved better accuracy and can meet the requirements of surveillance applications.

1. Introduction

Visual surveillance is an emerging trend in the area of computer vision for last few decades offering some propitious applications such as smart homes, surveillance aircrafts, automated inspection, facility protection and RADAR systems etc. CCTV cameras are used to capture the videos of targeted areas and monitored at strategic times in control rooms. These areas include shopping malls, airports, railway stations and other public places. Videos captured from such places contain complex information including individual person actions, group actions and interactions among people. Varying lighting conditions, occlusions and different view angles make the scene more complex. Multiple persons interacting with each other in a single frame can be observed in Fig. 1, all the interactions between persons in a scene must be analyzed for the identification of any unusual activity. Due to the significance of action recognition and behavior recognition in surveillance applications, a variety of methods have been proposed to address the problem [1–5]. Several surveys and review papers are also published within the area of human activity analysis and recognition. Poppe [6] and Weinland et al. [7] comprehensively reviewed feature representation and classification methods for vision-based human action recognition. Recently, a good survey covering challenges of the domain, activity representation techniques, classification approaches and datasets is presented by Zhang et al. [8].

Nowadays, multiple cameras are primarily installed to cover larger areas from different views in public environments, hence increasing scene complexity. Each camera is positioned at different location and therefore different views and appearance variations are

* Corresponding author.
E-mail address: nadeem.anjum@stemmainternational.com (N. Anjum).

https://doi.org/10.1016/j.compind.2018.03.015
0166-3615/

Fig. 1. HALLWAY dataset depicting people interactions captured with 3 cameras from three different views. Multiple views, occlusions due to objects/other persons and light variations make this dataset very challenging for interaction recognition.

created. Typically, humans are aided in control rooms to look at TV screens connected to the cameras. This is a tiresome job for humans to view multiple screens all the time. In this context, automated surveillance systems are essentially required to monitor all cameras, since humans are not capable to focus on multiple screens simultaneously. Main challenges in the automated systems are: there is a need to extract pose, scale and occlusion invariant features from each camera view to handle scene variations. Furthermore, similar behaviors can be perceived differently when viewed from different angles. In public places many persons interact with each other performing different activities like talking, queuing, fighting and walking together etc. which need to be monitored and classified as normal or abnormal scenarios. Therefore, machine vision algorithms have become vital in surveillance applications to handle the problems mentioned above.

Most of the existing methods [2,9–11] focused on recognition of individual activities shown in KTH [12] and Weizmann [13] datasets. The surveillance videos contain high-level activities, so interactions between people must be examined to prevent any unusual activity. However, some researchers have focused on joint activities and successfully incorporated contextual information on basis of the assumption that activities in a scene are rarely performed individually [4,13,14], but their work is limited up to a single camera view. In this paper, we proposed machine vision technique to recognize person-to-person interactions by considering individual actions as context. Further, we are considering multiple camera views which make it more challenging task. We employ the concept similar to [15] and [16], which used individual actions and poses for collective activity recognition. Contrary to [15,16], our goal is to recognize multiple interactions in complex sequences captured from multiple views. Unlike [14], persons are represented with their action labels returned by SVM classifier. We used individual action labels and collective poses of both persons for interaction recognition.

Given a person in a video, we have to observe his interaction with nearby persons, subsequently; person-to-person interaction recognition can be further exploited to recognize group activities and prediction of unusual activities. On the basis of this assumption, we consider individual actions as contextual information for person-to-person interaction recognition. Behaviors are divided into two categories; individual actions and person-to-person interactions.

The primary contributions of this research are as follows:

1. A hybrid approach of HOG and MDCLBP is used for identification of individual actions. Both are appearance based features. MDCLBP is novel approach; it is a variation of CLBP [17], where CLBP uses sign difference and magnitude information. MDCLBP uses sign difference, as used in LBP [18] and difference from median value as used in [19] to extract compound local binary patterns.
2. MDCLBP is proposed for texture feature extraction which eliminates the impact of occlusions and light variations, whereas HOG provides view and scale invariant representation. Collectively MDCLBP and HOG create a robust descriptor for individual action recognition under multiple views.
3. Histograms of MDCLBP and HOG are combined to form a feature vector for individual action recognition.
4. Trajectory features (distance, velocities and pose) are extracted from focal person and nearby person.
5. Individual action of each person is combined with trajectory features for person-to-person interaction recognition.

The proposed HOG-MDCLBP is evaluated on IXMAS [20], which is multi-view dataset recorded from different views using 5 different cameras. The robustness of proposed approach against occlusions is evaluated on OIXMAS [21] dataset, that contains many artificial occlusions of different shapes. Fig. 2 shows some example images of IXMAS and OIXMAS datasets depicting multiple views and occluded images. Experimental results show the robustness of HOG-MDCLBP against occlusions and multiple views. HALLWAY dataset [22] is used to evaluate person-to-person interaction recognition module. Results show that HOG-MDCLBP provides 94.06% average accuracy for individual person action recognition and person-to-person interactions are identified with an average accuracy of 98%.

Experimental results are generated in two steps: firstly, an SVM classifier [12] is used for recognizing individual person actions on the basis of HOG-MDCLBP descriptors. Secondly, given a focal person and a nearby person, trajectory features are extracted and combined with individual action labels recognized in the first step. Multiclass SVM classifier is trained by using trajectory features and associated action labels. We have validated the importance and effectiveness of our proposed feature set for person-to-person interaction recognition by performing series of tests; each time the test is performed by excluding one element from feature set.

This paper is organized as follows: Section 2 gives an overview of previous work. Section 3 provides a detailed discussion of proposed framework. We demonstrated the effectiveness of our proposed method with experiments in Section 4. Finally, conclusion and future work are provided in Section 5.

2. Related work

Most common descriptors used for individual action recognition include Histogram of Oriented Gradients (HOGs), Space-Time Interest Points (STIPs), Optical flow, Local Binary Pattern (LBP) and Motion History Image (MHI). Instead of using a single descriptor, researchers focused on combining multiple descriptors to reduce false classifications rate. Dinpakar [23] extracted HOG features from input video frames and identified pattern history by combining HOG features of consecutive frames. Learning and classification are performed using multiclass SVM classifier.

Although the activities are performed by different persons and recorded from different views, this approach can identify only one activity at a time and restricted to only three activities: browsing,