IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 8, 2020, accepted July 19, 2020, date of publication July 28, 2020, date of current version August 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012557

Human Interaction Anticipation by Combining Deep Features and Transformed Optical Flow Components

SHAFINA BIBI[1], NADEEM ANJUM[2], TEHMINA AMJAD[1], GRAEME MCROBBIE[3], AND NAEEM RAMZAN[3], (Senior Member, IEEE)

[1]Department of Computer Science and Software Engineering, International Islamic University, Islamabad 44000, Pakistan
[2]Department of Computer Science, Capital University of Science and Technology, Islamabad 46000, Pakistan
[3]School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, U.K.

Corresponding author: Nadeem Anjum (nadeem.anjum@cust.edu.pk)

ABSTRACT The anticipation of ongoing human interactions is not only highly dynamic and challenging problem but extremely crucial in applications such as remote monitoring, video surveillance, human-robot interaction, anti-terrorists and anti-crime securities. In this work, we address the problem of anticipating the interactions between people monitored by single as well as multiple camera views. To this end, we propose a novel approach that integrates Deep Features with novel hand-crafted features, namely Transformed Optical Flow Components (TOFCs). In order to validate the performance of the proposed approach, we have tested the proposed approach in real outdoor environments, captured using single as well as multiple cameras, having shadow and illumination variations as well as cluttered backgrounds. The results of the proposed approach are also compared with the state-of-the-art approaches. The experimental results show that the proposed approach is promising to anticipate real human interactions.

INDEX TERMS Human interaction anticipation, video surveillance, deep learning, transformed optical flow.

I. INTRODUCTION

The aim of human interaction anticipation is to recognize an interaction before its complete execution [1]. Preliminary studies have been attempted to recognize the actions of a person from single frame and from a few frames [2], [3]. This leads to the concept of anticipating an action from partially observed videos. However, in previous works, the focus was mainly on predicting single person's actions i.e. walk, stand, and sit; rather than complex activities or interactions (e.g. kicking and punching etc.) of more than one person. Although, many studies have been performed for the recognition of fully executed social behaviours and interactions from surveillance videos [4]–[7]; however, in real-world scenarios, often interactions and actions must be anticipated before they are fully executed. Therefore, the anticipation of human interactions and activities is becoming an active area of research. It has grabbed the attentions of research community due to its importance in several applications such as (a) in video surveillance (to generate the alarm before occurrence of any criminal activity (which often involves kicking, grabbing and fighting), (b) in smart homes (to anticipate a fall of elderly before its occurrence) and (c) for human-robot interaction (to make the robot able to aid human by observing human's intentions).

The anticipation of human interaction is a challenging problem since it requires the recognition of interaction from partial observations. Moreover, illumination variations and cluttered background in video surveillance data increase the uncertainty in anticipation task. Furthermore, it is a quite challenging to develop a machine vision algorithm for early recognition of interactions. For this, machines must be provided with extensive knowledge to infer the interactions by looking just a few initial frames. Therefore, interaction representation should be strong enough to recognize unfinished activities [5].

Unlike machines, humans have the ability to predict the future occurrences based on previous experience of such events and thus make themselves able to handle current activity. Therefore, in video surveillance scenarios, humans can perhaps better anticipate an untoward activity if it has to observe single or a couple of cameras' input. However, due

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

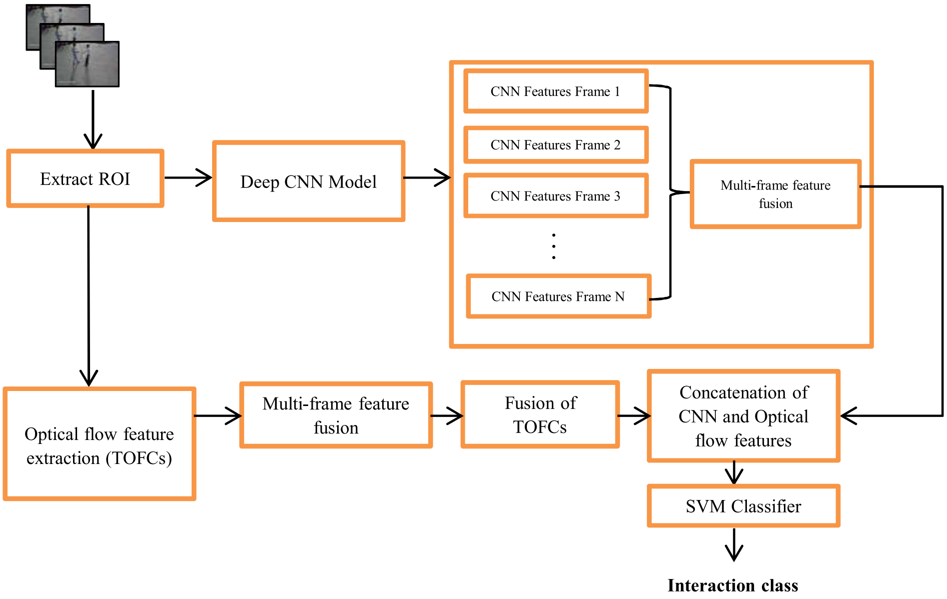


FIGURE 1. Block diagram of the proposed approach.

to increased demand for surveillance, multiple cameras have become an implicit part of our lives. A tremendous number of images and recordings exist within the world. It is a tedious job for a human to monitor multiple TV screens for a long time and thus its natural ability to anticipate an undesired activity degrades substantially. Contrary to human, machines have high-speed processors which can process huge amount of data and they are not tired of the same data being processed [8]. The existing human interaction anticipation/prediction methods have focused on the prediction of interaction from single camera views [8]–[11].

The problem of interaction anticipation becomes trickier when a scene is monitored with multiple cameras in outdoor scenarios. In this paper, our focus is to anticipate complex human activities in real outdoor scenarios in single and multiple camera setups. We propose to combine deep features with handcrafted features to reduce the effects of shadows and illumination variations to provide novel anticipation method. Instead of using more complex deep models (such as 3D CNN) which require very large dataset, we opted to use the features extracted from a single pass of deep network.

Primary contributions of this research include:

1) A novel approach to anticipate real human interactions in single- and multi-camera networks.
2) The introduction of novel Transformed Optical Flow Components (TOFCs) features.
3) Robust interaction representation by combining Deep features with Handcrafted features.

It is noteworthy that with regard to each interaction class, deep features obtained using deep learning models are more descriptive and salient [12]. The temporal information, however, provides useful information about the interaction pattern. The hybrid approach increased the accuracy of interaction anticipation. The proposed method is evaluated on single and multiple camera datasets captured in outdoor scenes having daylight illumination and compared with state-of-the-art approaches. Experimental results show that combining hand crafted and deep features outperform conventional hand crafted features.

The rest of this paper is organized as follows: Section 2 is about the related work; the proposed method is presented in Section 3; Section 4 details the experimental results and analysis; and Section 5 draws conclusions.

II. RELATED WORK

This section reviews the state-of-the art approaches in the context of human action recognition/ anticipation.

A. HANDCRAFTED FEATURES

Spatial and temporal features are mainly used for the representation of interaction and activities in a video frame. Ryoo [8] proposed to formulate the process of interaction prediction as posterior probability. Interactions are represented with integral bag-of-words and dynamic bag-of-words. Lopez *et al.* [13] predicted a portion of trajectories by using a simple trajectory-based representation named as activity description vector (ADV). The ADV is composed of frequency and the four directions of each point i.e. up, down, left and right. Barnachon *et al.* [14] proposed to use a histogram-based representation of poses for the recognition of streamed actions using motion capture data. They proposed to extend classical histogram to integral histogram for the representation of actions. Actions are compared and recognized using dynamic time wrapping paradigm. A former work on human interaction prediction presented by Ryoo [8] represented an activity as an integral histogram of spatio-temporal features. In this

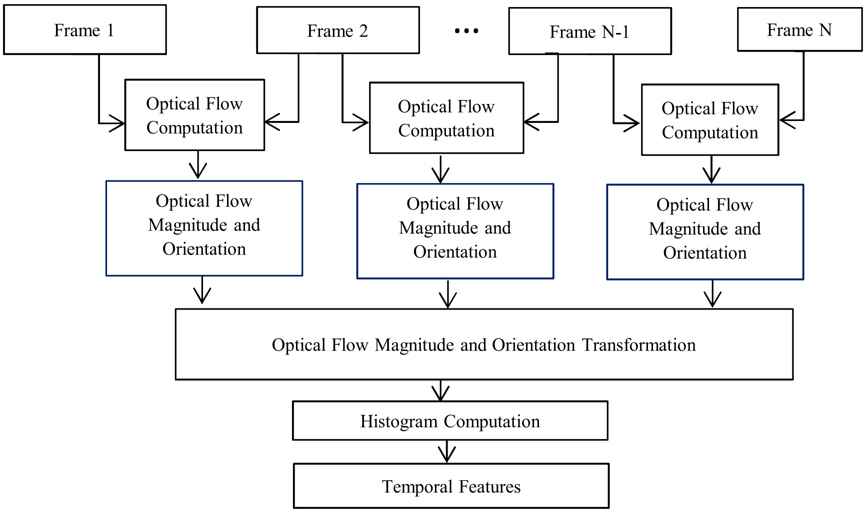

Frame 1
Frame 2
•••
Frame N-1
Frame N
Optical Flow Computation
Optical Flow Computation
Optical Flow Computation
Optical Flow Magnitude and Orientation
Optical Flow Magnitude and Orientation
Optical Flow Magnitude and Orientation
Optical Flow Magnitude and Orientation Transformation
Histogram Computation
Temporal Features

FIGURE 2. Depiction of the process to compute temporal features.

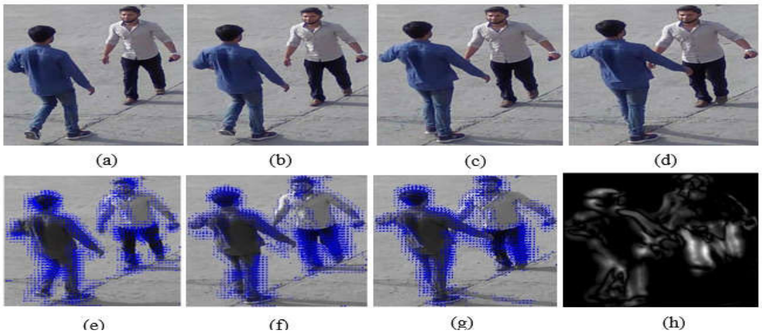


FIGURE 3. (a-d) input images of frame no k-2, k-1, k and k + 1. (e-g) the optical flow vectors. (h) Magnitude image(2nd order difference and thresholding).

work, the activity prediction problem is formulated probabilistically by using integral bag of words approach. A new recognition methodology named dynamic bag of words is developed for the recognition of on-going human activities and interactions. Sun *et al.* [15] detected spatio-temporal interest points and then sparse grouplets are located to represent body parts movement. Wang *et al.* [16] proposed a time series alignment-based activity prediction method. For this, a video sequence is divided into segments and then each segment is represented by local spatio-temporal statistics (Histogram of oriented gradients (HOG) and histogram of optical flow (HOF) using bag of visual words model. They compared the time series of different lengths using temporally-weighted generalized time warping (TGTW) model.

B. DEEP FEATURES AND MODELS

The handcrafted features i.e. HOGs, space time interest points (STIPs), trajectories and optical flow have some impediments in capturing salient motion information for the anticipation of interactions. Recent studies have shown that deep learned representations have boosted the performance of recognition and prediction tasks [22], [23] . Freitas [24] recognized single person actions by using unsupervised feature learning approach and proposed deep belief network (DBN). Choi *et al.* [17] also proposed to extract unsupervised features using multiple restricted boltzmann machines (RBMs) for the prediction of human behaviour in smart homes. Vondrick [23] predicted the visual representation of images in future by training a deep and performed learning from unlabelled videos. Ke *et al.* [9] introduced flow coding images for the representation of temporal information and proposed to extract deep features from temporal images. They proposed to apply temporal convolution on video frames to describe deep temporal information. Ke *et al.* [10] further proposed to use spatial and temporal models learned with longer short-term memory (LSTM) networks. They proposed to combine spatial model, spatial structural model, temporal model and temporal structural model for the prediction of partial observations. Previous work [22], [25] also demonstrated the effectiveness of combining deep features and handcrafted features in classification tasks.

A skeleton based action prediction method is proposed by Bennamoun *et al.* [19]. This approach uses a global regularizer to learn hidden features and a temporal aware cross

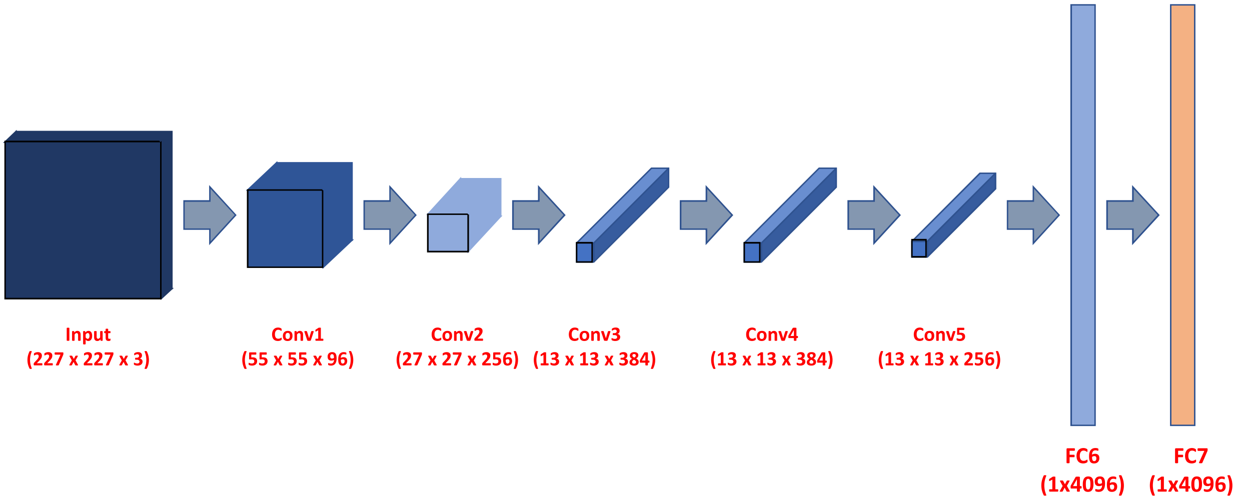


FIGURE 4. Alexnet architecture for extraction of deep features.

TABLE 1. Characteristics of human behaviour anticipation methods.

Ref.	Feature(s)	Classifier(s)	Characteristics
[14]	Histogram of space + time features	Dynamic bag of words	Single view Handcrafted feature Tested in controlled environment
[15]	Sparse group-lets to represent movement of body parts	DTW-E	Single view Single person Handcrafted features No occlusion handling
[11]	Body movements	max-margin learning framework	Single view Less efficient against complex interactions
[10]	Spatial + Temporal models	Long Short Term Memory (LSTM) networks	Single view spatial model+spatial structural model+ temporal model and temporal structural model for the prediction of partial observations
[17]	Space+time features	DBN-R	Single view Handcrafted features Deep learning
[9]	Pre-trained CNN to extract features from Flow coding images	Softmax	Deep temporal architecture
[18]	HOG+action-let/pose-let+CNN features	Probabilistic model	Object affordance based on distance and angular preferences
[19]	Skeleton data	CNN	single camera Skeleton data Requires full sequences
[20]	Skeleton joints	Scale selection network	Single view 3D skeleton data Best results are achieved on 90% observation ratio/ Difficult to predict actions having same motion pattern
[21]	Time-phase feature of the Gaussian model	Multi-feature fusion network algorithm based on parallel Inception and ResNet	Single view Whole-individual detection is required
Proposed	Handcrafted temporal features + deep features	SVM	Integrated handcrafted and deep features Handles illuminations and clutters Single view and multi-view outdoor complex interactions

entropy to address the challenges of diverse motion in an action sequence captured from single view camera. Compelling results have been achieved; however, the proposed network requires information about full video sequences. This is essential to construct the hidden feature layer. Liu *et al.* [20] also predicted actions from 3D skeleton sequences using single view camera. Motion dynamics in input sequences are modelled using dilated convolutional network and a hierarchy of dilated convolutions is used to learn the multi-level representations from input skeleton data. Like [19], Skeleton data is used in [20] for action prediction. This skeleton data is mainly based on precise detection of skeleton and joints; which is often difficult especially in real videos. Both [19] and [20] require extensive training data to effectively train the networks.

In this work, we focused on anticipation of complex interactions by aggregating the motion and CNN features in single as well as multi-view cameras. We performed training on partial sequences, hence full videos are not required for training and testing. Previous work on interaction/activity anticipation focused on single camera scenarios. Use of multiple cameras provides different views of entire scene from distinct positions, which helps to observe the interactions from multiple angles. Multiple camera views are advantageous in outdoor scenarios having occlusions and cluttered background as each camera's input is deemed to make the final decision. Table 1 compiles key characteristic differences between the proposed approach and various state-of-the art approaches.

III. THE PROPOSED APPROACH

Anticipation of ongoing human interactions under multiple camera-views is a challenging problem because:

- unfinished videos only provide the early part of the interaction;
- videos captured in outdoor environments may contain shadows and cluttered background.

Handcrafted features have been extensively used for recognition of simple human activities; however, they underperformed in more complex scenarios. Due to the successful application of convolutional neural network (CNN) features in activity and interaction anticipation tasks; we proposed to combine CNN features with hand-crafted feature for interaction representation. This is due to the fact the hand-crafted features alone are not powerful enough to capture salient motion information in a video [26]. The proposed approach for interaction anticipation is depicted in Fig. 1. It includes five basic steps:

1) Extraction of Optical flow and CNN features.
2) Computation of TOFCs and their representation using histograms.
3) Concatenation of the CNN features and TOFCs features.
4) Construction of final feature vector by employing both types of features computed in every frame.
5) Classification using Support Vector Machines.

It is worth-mentioning that the above-mentioned steps are applied under each camera view and finally the classification results are fused to get the final decision. Details of these steps are presented in subsequent sections.

A. PROBLEM FORMULATION

Let $C = \{C_1, C_2, \ldots, C_M\}$ be a set of M partially overlapping synchronized cameras. Let $V = \{1, 2, 3, \ldots, N\}$ represent a video having N number of frames. Persons are detected and bounding boxes are drawn around detected persons in each frame using aggregate channel features [27]. Person's locations are used as ground truths. We have used the ground truths of one camera and then Homohraphy matrix (H) is applied to estimate the projection from image plane. (x, y, t) to ground truth-plane (x_G, y_G, t), here G symbolizes the ground truth-plane. The Homography matrix is applied as follows:

$$(x_G, y_G, t) = H(x, y, t), \tag{1}$$

where (x_G, y_G, t) is the ground-plane projection of the point (x, y, t). H is the homography matrix that is constructed by selecting the control points of provided ground truths. The proposed method anticipates interactions between two persons. Let R_{pq} be the region of interest around the two persons p and q. Deep features and TOFCs extracted from R_{pq} are represented with f_{deep} and f_{opf} respectively. Our goal is to anticipate ongoing human interactions under multiple camera views by representing partial observations with deep and temporal features.

B. FEATURE EXTRACTION

1) HAND-CRAFTED FEATURES

In this paper, we used hand-crafted features (optical flow) to get the temporal information from successive video frames. Hand-crafted features are extracted by computing optical flow images from the bounding box around two persons in consecutive frames. We used differential method i.e. Horn Schunck optical flow to compute optical flow among four consecutive frames. Optical flow components are transformed to represent the interactions for anticipation [28].

2) TRANSFORMED OPTICAL FLOW COMPONENTS

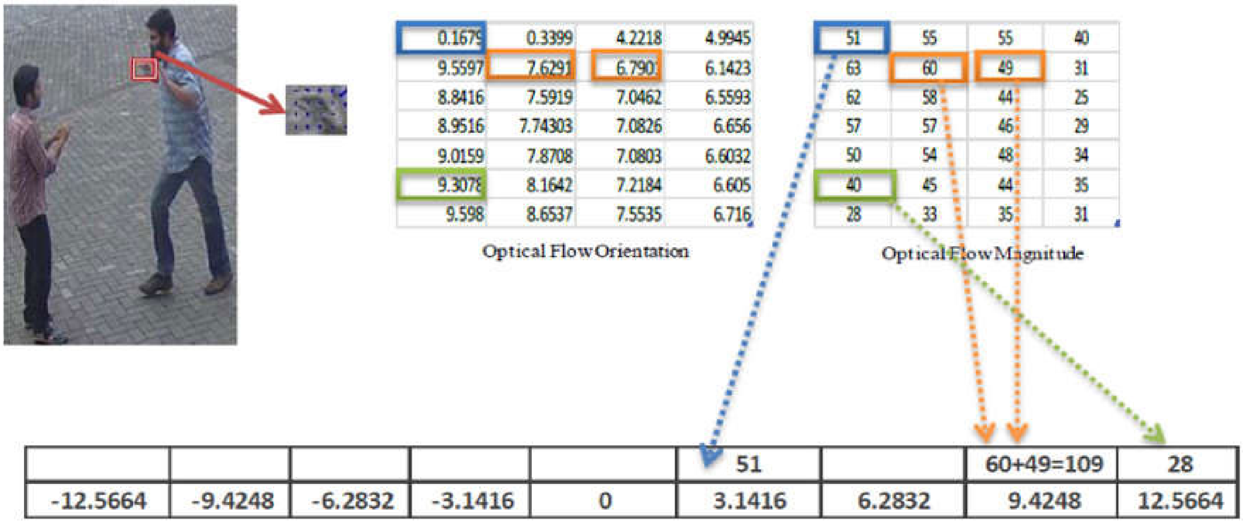
Optical flow magnitude and orientation are computed from optical flow vectors as follows:

$$\mathrm{M}_{x,y} = \sqrt{(r_x^i)^2 + (r_y^i)^2}, \tag{2}$$

$$\ominus_{x,y} = \tan^{-1} \frac{r_y^i}{r_x^i}, \tag{3}$$

where $\mathrm{M}_{x,y}$ represents magnitude and $\ominus_{x,y}$ represents the orientation at location(x, y). Illumination variations and cluttered background in outdoor environment cause the optical flow to include numerous noisy observations in flow field. We propose to transform optical flow components (magnitude and orientation) by applying second order difference on both components and then thresholding them. Contrary to [29], we have not directly threshold optical flow components, rather the second order difference is applied on optical flow components computed from four consecutive frames before thresholding both components. The advantage of second order difference is twofold: (1) thresholding difference components provides fine details and removes small flow variations, (2) temporal information can be specified well by applying second order difference.

The proposed method to represent optical flow magnitude and orientation is named as Transformed Optical Flow Components (TOFCs). TOFCs are then represented with a histogram of transformed optical flow magnitude (HTOM).



0.1679	0.3399	4.2218	4.9945
9.5597	7.6291	6.7901	6.1423
8.8416	7.5919	7.0462	6.5593
8.9516	7.74303	7.0826	6.656
9.0159	7.8708	7.0803	6.6032
9.3078	8.1642	7.2184	6.605
9.598	8.6537	7.5535	6.716

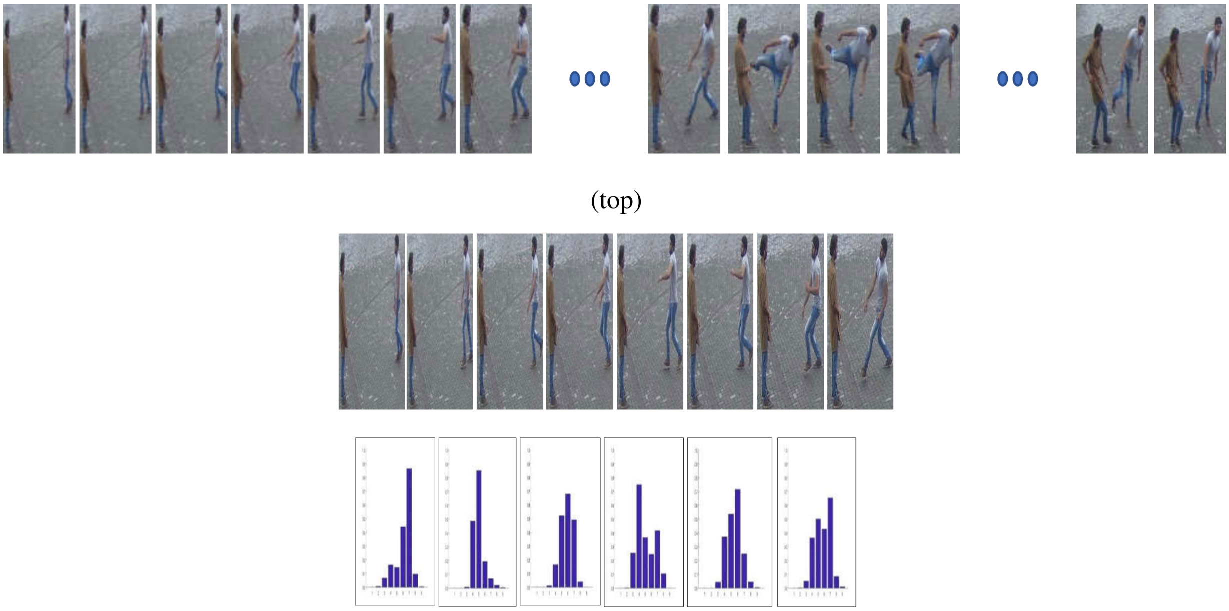
Optical Flow Orientation

51	55	55	40
63	60	49	31
62	58	44	25
57	57	46	29
50	54	48	34
40	45	44	35
28	33	35	31

Optical Flow Magnitude

					51		60+49=109	28
-12.5664	-9.4248	-6.2832	-3.1416	0	3.1416	6.2832	9.4248	12.5664

FIGURE 5. The process to compute HTOM.



(top)

(bottom)

FIGURE 6. (top) Sample frames of entire Kick activity; (bottom) HTOMs of frames used for anticipation of Kick activity.

The process to compute TOFCs is depicted in Fig. 2. To compute TOFCs, second order difference of optical flow magnitude and orientation is computed in four consecutive frames respectively. Magnitude values are scaled between 0-255 by applying linear transformation and then thresholding is applied on magnitude and orientation. Following equations are applied to compute element wise second order difference of magnitude and orientation respectively.

$$\acute{\mathrm{M}}_{x,y} = \mathrm{M}^1_{x,y} - 2\mathrm{M}^2_{x,y} + \mathrm{M}^3_{x,y} \tag{4}$$

$$\acute{\ominus}_{x,y} = \ominus^1_{x,y} - 2\ominus^2_{x,y} + \ominus^3_{x,y} \tag{5}$$

where, M^1 is the magnitude between first and second frame, M^2 between second and third frame and M^3 between third and fourth frame. Similarly, $\ominus^1$, $\ominus^2$ and $\ominus^3$ represent the orientations computed from four consecutive input frames. Eq. 4 and Eq. 5 are applied on both components to enhance temporal information in a frame by considering the flow vectors of previous and next frames. $\acute{\mathrm{M}}_{x,y}$ and $\acute{\ominus}_{x,y}$ are the resultant magnitude and orientation at location x, y. $\acute{\mathrm{M}}_{x,y}$ is

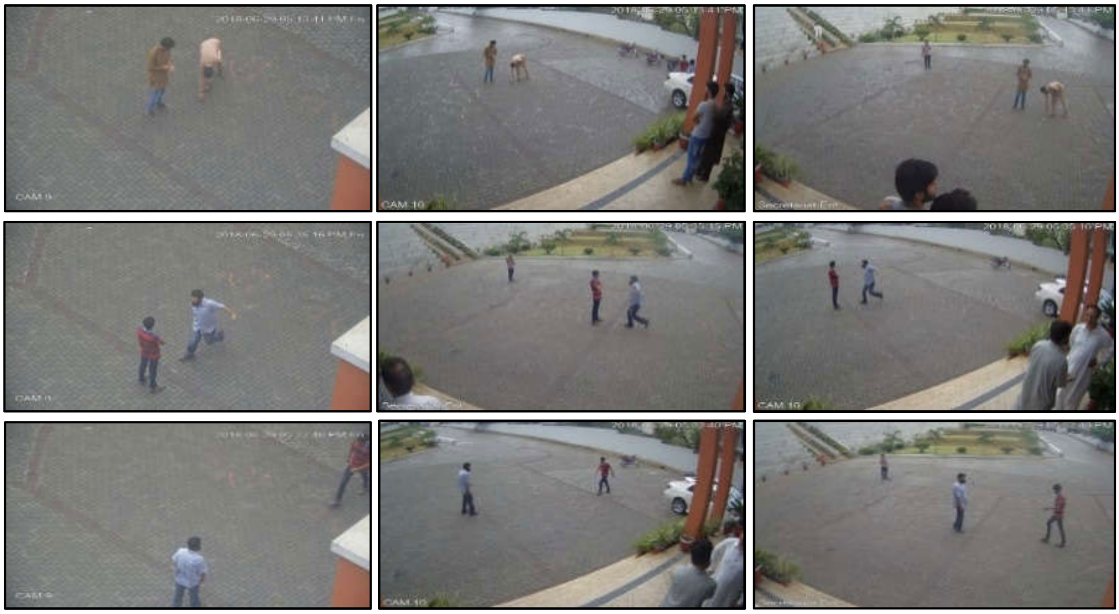


FIGURE 7. Snapshots of MU-Interaction1 from multiple views capturing different activities.

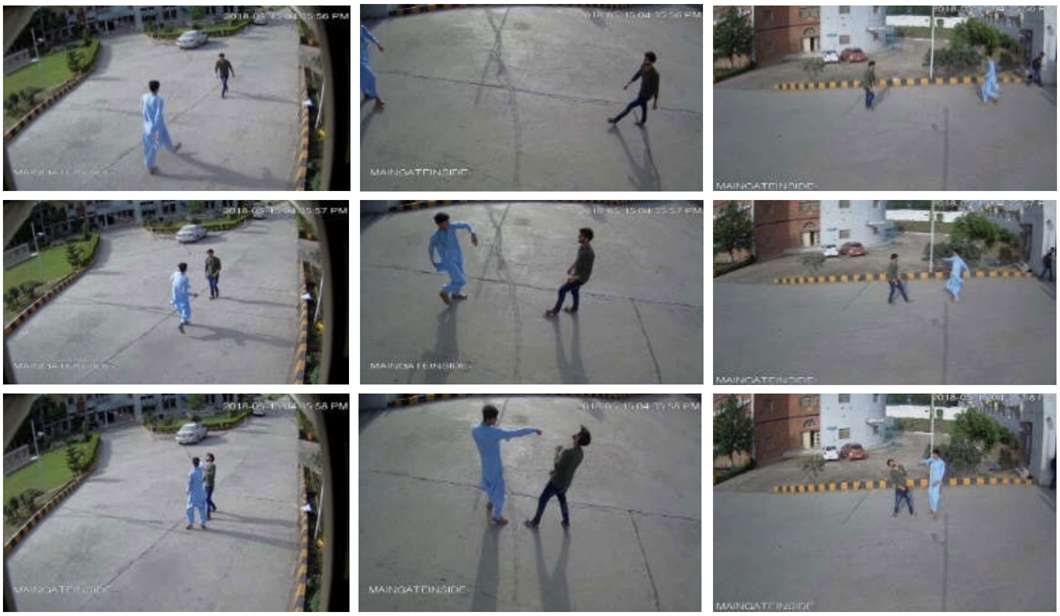


FIGURE 8. Snapshots of MU-Interaction2 from multiple views capturing different activities.

then scaled between 0-255 and thresholding is applied on linearly scaled magnitude as follows:

$$\acute{M}_{x,y} = \begin{cases} 0 & \text{if } \acute{M}_{x,y} < \tau_M, \\ \acute{M}_{x,y} & \text{if } \acute{M}_{x,y} \geq \tau_M, \end{cases} \tag{6}$$

$$\acute{\ominus}_{x,y} = \begin{cases} 0 & \text{if } \acute{\ominus}_{x,y} < \tau_{\ominus}, \\ \acute{\ominus}_{x,y} & \text{if } \acute{\ominus}_{x,y} \geq \tau_{\ominus}, \end{cases} \tag{7}$$

where τ_M and $\tau_{\ominus}$ are empirically selected. Fig. 3 shows the detected flow vectors on four consecutive frames and the resultant magnitude after applying second order difference and thresholding.

C. DEEP FEATURE EXTRACTION

Deep features are extracted using Alexnet [30]. In this paper, we used Alexnet only for the extraction of features, which requires a single pass on input frames. The Alexnet architecture consists of 8 layers in total. Five layers of Alexnet are primarily convolution layers (Conv1, Conv2, Conv3, Conv4, Conv5) and last three layers are the fully connected layers (FC6, FC7, FC8). The architecture of Alexnet network is depicted in Fig. 4. In Alexnet architecture a rectified linear unit (ReLU) is applied after each convolution step and then normalization is applied after ReLU in the first two layers. Max pooling is applied in three layers: first two layers after normalization and in the fifth layer after ReLU. In this paper, we used the output of FC7, which contains 4096 dimension feature vector to represent the interaction between two persons. Deep features are computed in each frame returning 4096 dimensional feature vector. Hence, N4096 dimensional feature is extracted from each input video, where N is the total number of frames in a video. The deep output features of N consecutive frames from a video are concatenated by applying median absolute deviation (MAD).

$$f_{deep}(V) = |\{f_{deep}(k)\}_{k=1}^{N} - median(f_{deep})|. \tag{8}$$

D. INTERACTION REPRESENTATION

This paper proposes to combine deep features and temporal features for the anticipation of ongoing interactions. We simply concatenated both features as follows:

$$feat(V) = [f_{opf}, f_{deep}] \tag{9}$$

where feat(V) denotes the final feature vector after concatenating deep and temporal feature extracted from video V. f_{opf} is temporal feature (either represented by concatenating both components or by using HTOM and f_{deep} is the output deep feature vector. After feature representation is completed, training is performed on features by using SVM classifier.

E. REPRESENTATION OF TEMPORAL FEATURES

Temporal features can be represented in two way: **(a)** by simply concatenating the histograms of magnitude and orientation i.e.

$$f_{opf} = [H(\acute{\mathrm{M}}), H(\acute{\ominus})], \tag{10}$$

where H is the histogram. **(b)** By computing histogram of transformed oriented magnitude (HTOM) to represent transformed components like HOFM in [31]. Unlike HOFM, second order difference and thresholding is applied on optical flow components before computing oriented magnitudes from the overall region of interest. Following are the steps to compute HTOM:

- Orientations are represented with 8-bins in the range -12 to $+12$ which is set as follows: $-8 * \pi/2 : 2 * \pi/2 : 8 * \pi/2$
- The histogram computation is performed by considering the magnitude and orientation values at each pixel location.
- The bins for histogram are chosen from orientations and the votes for bins are chosen on the basis of magnitude.
- For the orientation value greater than 12; magnitude is added to the last bin i.e. 12. Magnitude is added to the first bin if the orientation value is less than -12. Figure 5 depicts the process of computing HTOM.
- The histograms of all frames of a video V are fused by applying Median Absolute Deviation (MAD) as follows [32]:

  $$f_{opf}(V) = |f_{opf}(k)_{k=1}^{N} - median(f_{opf})| \tag{11}$$

  where, $f_{opf}(k)$ is the temporal feature histogram at frame. HTOMs of kick interaction are shown in Fig. 6.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We have performed experiments on multi-view camera datasets for the evaluation of the proposed approach. Experiments are performed on multi-view datasets (MU-Interaction1 and MU-Interaction2) and on publicly available UT-Interaction dataset [33] and results are compared with state-of-the-art approaches.

A. DATASETS

MU-Interaction1 is captured in front of Mirpur University's secretariat building using three Dahua IP cameras and includes 7 interaction classes: Bend, Faint, Handshake, Hug, Kick, Punch and Push. These interactions are performed by 8 people. No restrictions are imposed on the people regarding their positions and actions. Each camera has resolution of 1920 × 1080 pixels and the frame rate is 30 Hz. Figure 7 shows the snapshots of each camera view.

TABLE 2. Interaction anticipation accuracies (percentages) of proposed method on MU-Interaction1 dataset (TOFCs are represented with concatenated histograms) (Average accuracy = 91.5%). Rows: Predicted labels Columns: True labels.

Bend	85.71	14.29	0	0	0	0	0
Faint	0	100	0	20	0	0	0
Handshake	0	0	100	0	0	0	0
Hug	0	0	0	100	0	0	0
Kick	0	0	0	0	91.67	8.33	0
Punch	0	0	0	14.285	0	71.43	14.285
Push	0	0	0	0	0	8.33	91.67

MU-Interaction2 is a challenging outdoor dataset captured at university's entrance using 3 Dahua IP cameras having the resolution of 1920 × 1080 and 10fps, it includes 5 interaction class videos: Hug, Handshake, Kick, Punch and Push which are recorded by 8 persons. 70 samples are collected totally under each camera. In this dataset, shadows and illumination variations are very prominent. One camera captures the scene from the top and the other two cameras are placed on the left and right of the entrance gate. Snapshots of this scenario are provided in Fig. 8.

UT-Interaction dataset is a publicly available dataset containing videos recorded with single camera and having 6 interaction classes: Handshake, Hug, Kick, Punch, Push and Point with 10 instances of each class. Figure 9 shows the snapshots of the dataset.

B. EVALUATION METHOD

For experiments, the region of interest (ROI) includes both interacting people, chosen by combining both people's bounding boxes. Deep features are extracted from ROI using Alexnet model and the output of FC7 is used as deep feature vector. Temporal features (TOFCs) are extracted by extracting optical flow from four consecutive frames and applying second order difference on optical flow components (magnitude and orientation). The resulting components are

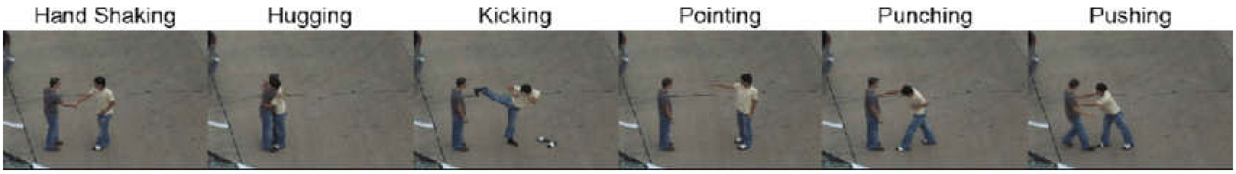


FIGURE 9. Snapshots of UT-Interaction dataset.

TABLE 3. Interaction anticipation accuracies (percentages) of proposed method on MU-Interaction1 dataset (TOFCs are represented with HTOM) (Average accuracy = 92.72%). Rows: Predicted labels Columns: True labels.

	Bend	Faint	Handshake	Hug	Kick	Punch	Push
Bend	100	0	0	0	0	0	0
Faint	0	100	0	0	0	0	0
Handshake	0	0	80	20	0	0	0
Hug	0	0	16.67	83.33	0	0	0
Kick	0	0	0	0	100	0	0
Punch	0	0	0	0	0	85.71	14.29
Push	0	0	0	0	0	0	100

further thresholded on the basis of an empirically chosen threshold value τ ($\tau = 20$ for MU-Interaction1 and $\tau = 35$ for MU-Interaction2). The lower value of τ retains the small variations caused by cluttered background and illumination changes. The higher the value of τ, many useful flow values will also be thresholded. The threshold value is different for both datasets because each dataset is recorded in different environment. The extracted TOFCs are then represented with histograms and Median Absolute Deviation is applied to combine the histograms of one video stream. Deep features and temporal features are combined to represent the interaction between two persons. SVM classifier is used for training and classification. SVMs are binary classifiers and we opted one-against-one method to perform multi class classification. Classification is performed under each camera view and the results are fused to get the final decision. We applied majority voting to decide the final interaction class among all classifiers [34]. If all classifiers select distinct interaction classes, the classifier whose probability of correct classification is high will be considered as reliable classifier [35]. The effectiveness of the proposed complex human interaction method is tested by using cross-validation under each camera view separately. Leave-one-out cross-validation is applied to assess the performance of the proposed method. Instead of providing complete video frames, partial observations are given for anticipation. Experiments are carried out on distinct observation ratios, from 0.2 to 1.0, with step size of 0.1 after the same procedure as in [9]. Here 0.2 indicates that 20% of total frames are utilized for classification. If there are total N frames in a video then $[1, round(0.3 * N)]$ means that 30% of N frames are used to anticipate the interactions.

C. EXPERIMENTAL RESULTS

1) EXPERIMENTS ON MU-INTERACTION1 DATASET

The first set of experiments on this dataset is performed by representing TOFCs with concatenated histograms (eq. 7). Deep features and histogram of TOFCs from one video are concatenated resulting 1×4068 dimensional feature vector. The proposed method is applied by selecting different observation ratios. 30% accuracy is achieved with observation ratio 0.2, the accuracy of the proposed method enhanced by 20% if 30% of the entire interaction is used for anticipation. 15% improvement is observed with 40% observation ratio. The accuracy attained on 60% observation ratio is further noteworthy (92.59%). The observation ratios above 60% are closer towards the completion of interactions so we have chosen 0.6 of the entire video for the anticipation task. 8% of error rate is observed with leave-one-out cross-validation. The improvement of 2.5% (accuracy = 94.5%) is observed when classification is performed on the entire video (the recognition of complete interaction pattern).

TABLE 4. Interaction anticipation accuracies (percentages) of the proposed method on MU-Interaction2 dataset (TOFCs are represented with concatenated histograms) (Average accuracy = 86.34%). Rows: Predicted labels Columns: True labels.

	Handshake	Hug	Kick	Punch	Push
Handshake	71.42	14.29	0	14.29	0
Hug	11.11	88.89	0	0	0
Kick	0	0	100	0	0
Punch	7.14	0	0	92.86	0
Push	0	7.14	0	14.29	78.57

TABLE 5. Interaction anticipation accuracies (percentages) of the proposed method on MU-Interaction2 dataset (TOFCs are represented with HTOM) (Average accuracy = 90.95%). Rows: Predicted labels Columns: True labels.

	Handshake	Hug	Kick	Punch	Push
Handshake	85.71	14.29	0	0	0
Hug	5.56	83.33	0	0	0
Kick	0	0	100	0	0
Punch	0	0	0	100	0
Push	0	0	0	14.29	85.71

The confusion matrix of interaction anticipation accuracies attained using 0.6 of entire observation is depicted in Table 2. Rows of confusion matrices correspond to the predicted labels and columns correspond to true labels. These results are accomplished by fusing the classification results of all camera views. Majority voting based fusion is performed i.e. all instances of faint are recognized in Cam_1 and Cam_3 (accuracy = 100%), hence on the basis of majority voting,

TABLE 6. Average precision, recall and f-measure of proposed method on both datasets.

	MU-Interaction1		MU-Interaction2	
	Deep feat.+concat. TOFCs	Deep feat.+HTOM	Deep feat.+concat. TOFCs	Deep feat.+HTOM
Avg. Accuracy	0.91	0.927	0.86	0.90
Avg. Precision	0.90	0.93	0.87	0.91
Avg. Recall	0.92	0.93	0.86	0.91
Avg. F-measure	0.91	0.93	0.86	0.91

TABLE 7. Results of applying T-test.

			95% confidence interval of the difference	
	Sig.	Mean	Lower	Upper
Deep features	0.025	62.500	30.734	94.266
TOFCs	0.027	58.500	26.734	90.265
Deep features+TOFCs	0.006	91.835	80.950	100.000

all instances of faint are correctly identified. It is noteworthy that in the context of this paper, ‘‘fainting down’’ and ‘‘falling down’’ are same action patterns. However, we distinguish ‘‘fainting down’’ or ‘‘falling down’’ from normal ‘‘bending down’’ event by considering the movements of nearby person(s). Often, If the person is falling down or fainting down, nearby people shall run to aid that person. In case of normal bending, motion of nearby people will not matter.

The second set of experiment on MU-Interaction1 is performed by combining deep features with HTOM. The computation of HTOM is similar to the process described in [31], except we have transformed optical flow components before extracting HTOM. The TOFCs are represented with 9-bin histogram and concatenated with deep features returning 1 × 4105D feature vector. Experiments are performed on different observation ratios and the performance of the proposed method is improved 1.5% to 2% in each observation ratio as compared to the concatenation method. 94% accuracy is attained by the proposed method when 60% observations are provided as input. Confusion matrix depicted in Table 3 explicates that anticipation accuracy of all classes is improved with this method.

TABLE 8. Accuracy comparison with state-of-the-art approaches.

Method	Accuracy
Proposed	94%
Ke et al. [10]	86.67%
Ke et al. [9]	88.3%
Lan et al. [11]	88.1%
Ryoo et al. [8]	70.0%
Ye et al. [21]	91.7%

2) EXPERIMENTS ON MU-INTERACTION2 DATASET

The first experiment on MU-Interaction2 is performed by representing TOFCs with concatenated histograms and combined with deep features for interaction representation. This dataset is very challenging as it is captured in an outdoor environment having illumination variations and shadows. Learning and testing are performed under each camera view and classification results are fused to get a final decision. Experiments are performed on different observation ratios (as on MU-Interaction1 dataset). The experimental results revealed that accuracy increases in first 8 observation ratios, accuracy decrement of 0.5% is observed when the entire video is used for training and testing. By utilizing 60% observations, the proposed method attained 88% accuracy for the anticipation of complex interactions. Results are given in Table 4. These results are attained after fusing the accuracies of separate classifiers under all camera views.

The second set of experiments on MU-Interaction2 dataset is performed by representing TOFCs with HTOM and combining with deep features. The anticipation accuracy of the proposed method on this dataset is improved when TOFCs are represented with HTOM.

An overall 3.3% improvement in accuracy is observed and the improvement ratio of accuracy is also improved when experimented with distinct observation ratios. An overall 91.30% accuracy is achieved using leave-one-out cross-validation and the results are depicted in Table 5. The average precision, recall and f-measure values are shown in Table 6 which clearly depicts that the proposed method attained acceptable results using deep + HTOM features.

3) SIGNIFICANCE TEST

T-test is applied to measure the significance of classifier on proposed features. The rule to check the significance is that if $p >= 0.05$ then the results are significant. Table 7 shows the results of T-test applied on separate and combined feature elements. Sig. (p value) on combined features is 0.006 which indicates that the results of proposed features are significant as compared to separate elements. Mean accuracy of proposed features is 91.84 which is also significantly different from the mean accuracies on the results of separate features.

4) COMPARISON

Comparison of the proposed method is performed with five state-of-the-art approaches on UT-Interaction dataset [8]–[11], [21] . Videos in this dataset are recorded with single camera and having 6 interaction classes: Handshake, Hug, Kick, Punch, Push and Point with 10 instances of each class. Training and testing split is not provided with this dataset. The Performance is assessed by using leave-one-out cross validation. Table 8 shows the accuracy of the proposed method and previous methods for UT-Interaction dataset. The proposed method attained better performance and outperforms other approaches. Ke *et al.* [10] combined the structure of interaction context with the spatial and temporal information of input videos. Long short-term network (LSTM) is used to learn the spatial and temporal models. The proposed approach when compared with [10] achieved the improvement of about 8% on 60% observation ratio. An improvement of about 6% is achieved as compared to [9] which utilized temporal images for deep feature extractions. The hierarchal movements [11] achieved 83.1% accuracy on 50% observation ratio. The proposed method outperform the integral histogram based activity representation method [8] by 24%. This is an early work on interaction prediction that achieved 70% accuracy on half observations. Finally, Ye *et al.* [21] achieved accuracy of 91.7%, which is lower by 2.3% compare to the proposed approach.

V. CONCLUSION

In this research, we have proposed a novice method for the anticipation of ongoing person-to-person interactions from multiple camera views in outdoor environments. The anticipation of ongoing interactions in an outdoor environment is a challenging problem due to cluttered background, shadow and illumination variations. We proposed to represent the interactions with deep features and temporal features. Deep features are extracted by using Alexnet model and temporal features are extracted by applying optical flow. It is further proposed to transform optical flow components by applying second order difference and thresholding the transformed components. In addition, the proposed approach is tested on real out door scenarios. The proposed method achieved 92% and 89% average accuracy on MU-Interaction1 and MU-Interaction2 datasets, respectively. The proposed method is also tested on single camera-view dataset (UT-Interaction) and compared with state-of-the-art approaches. The proposed approach attained the accuracy of 94%, which is at least 6% better than existing state-of-the art approaches. In future, we are intended to extend this work to perform human interaction anticipation in multiple camera scenarios having both partially overlapping and non-overlapping views to cover more area in public places. Also, we intend to explore deep learning for classification purpose in the context of real world event anticipation problems; depending upon the computational viability of these networks.

REFERENCES

[1] B. Rodriguez, C. Fernando, and H. Li, "Action anticipation by predicting future dynamic images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 89–105.

[2] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[3] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1091–1104, Aug. 2002.

[4] Y. Kong and Y. Fu, "Close human interaction recognition using patch-aware models," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 167–178, Jan. 2016.

[5] C.-W. Chen and H. Aghajan, "Multiview social behavior analysis in work environments," in *Proc. 5th ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Aug. 2011, pp. 1–6.

[6] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1709–1718.

[7] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. CVPR*, Jun. 2011, pp. 3273–3280.

[8] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1036–1043.

[9] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, "Human interaction prediction using deep temporal features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 403–414.

[10] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Leveraging structural context models and ranking score fusion for human interaction prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1712–1723, Jul. 2018.

[11] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 689–704.

[12] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Proc. Sci. Inf. Conf.* Cham, Switzerland: Springer, 2019, pp. 128–144.

[13] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, and J. Garcia-Rodriguez, "A novel prediction method for early recognition of global human behaviour in image sequences," *Neural Process. Lett.*, vol. 43, no. 2, pp. 363–387, Apr. 2016.

[14] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture," *Pattern Recognit.*, vol. 47, no. 1, pp. 238–247, Jan. 2014.

[15] Q. Sun, H. Liu, M. Liu, and T. Zhang, "Human activity prediction by mapping grouplets to recurrent self-organizing map," *Neurocomputing*, vol. 177, pp. 427–440, Feb. 2016.

[16] H. Wang, W. Yang, C. Yuan, H. Ling, and W. Hu, "Human activity prediction using temporally-weighted generalized time warping," *Neurocomputing*, vol. 225, pp. 139–147, Feb. 2017.

[17] S. Choi, E. Kim, and S. Oh, "Human behavior prediction for smart homes using deep learning," in *Proc. IEEE RO-MAN*, Aug. 2013, pp. 173–179.

[18] V. Dutta and T. Zielinska, "Predicting human actions taking into account object affordances," *J. Intell. Robot. Syst.*, vol. 93, nos. 3–4, pp. 745–761, Mar. 2019.

[19] Q. Ke, J. Liu, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Global regularizer and temporal-aware cross-entropy for skeleton-based early action recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 729–745.

[20] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1453–1467, Jun. 2020.

[21] Q. Ye, H. Zhong, C. Qu, and Y. Zhang, "Human interaction recognition based on whole-individual detection," *Sensors*, vol. 20, no. 8, p. 2346, Apr. 2020.

[22] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.

[23] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 98–106.

[24] N. De Freitas, "Deep learning of invariant spatio-temporal features from video," Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., 2010.

[25] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Combining deep learning and hand-crafted features for skin lesion classification," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Dec. 2016, pp. 1–6.

[26] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2014, pp. 303–314.

[27] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[28] B. K. Horn and B. G. Schunck, "Determining optical flow," *Proc. SPIE*, vol. 281, pp. 319–331, Nov. 1981.

[29] C. Caetano, V. H. C. de Melo, J. A. dos Santos, and W. R. Schwartz, "Activity recognition based on a magnitude-orientation stream network," in *Proc. 30th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2017, pp. 47–54.

[30] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[31] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017.

[32] F. Mosteller and J. W. Tukey, *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA, USA: Addison-Wesley, 1977.

[33] M. Ryoo and J. Aggarwal. (2015). *Ut-Interaction Dataset, ICPR Contest on Semantic Description of Human Activities (SDHA) (2010)*. [Online]. Available: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

[34] C. Nadal, R. Legault, and C. Y. Suen, "Complementary algorithms for the recognition of totally unconstrained handwritten numerals," in *Proc. 10th Int. Conf. Pattern Recognit.*, vol. 1, 1990, pp. 443–449.

[35] A. Mi, L. Wang, and J. Qi, "A multiple classifier fusion algorithm using weighted decision templates," *Sci. Program.*, vol. 2016, pp. 1–10, Sep. 2016.

SHAFINA BIBI received the master's degree and Ph.D. degree in computer science from International Islamic University Islamabad (IIUI), Pakistan, in 2009 and 2020, respectively. In her Ph.D. research, she focused on recognition and anticipation of person to person interactions under multiple camera views. She is currently interested in anticipation of abnormal behaviors in crowded environments.

NADEEM ANJUM received the master's degree (Hons.) in computer science from International Islamic University, Islamabad, in 2001, the M.S. degree from Queen Mary University of London, U.K., in 2006, and the Ph.D. degree, in 2010. In 2010, he joined Riphah International University, Islamabad, as an Assistant Professor. In 2013, he joined the University of Engineering and Technology Taxila, as an Assistant Professor, and has served the organization, for three years. In 2016, he joined Stemma International (Pvt.), Limited, as the Founding Director. He has been involved in research and development for more than 18 years. His current research interests include deep learning for human activity recognition, computer vision, and machine learning. He received the Best Paper Award at the IEEE AVSS 2009.

TEHMINA AMJAD received the Ph.D. degree (Hons.) in computer science from International Islamic University, Islamabad, Pakistan, in 2015, under a split Ph.D. program, where she conducted her research work at Indiana University Bloomington, Bloomington, IN, USA. She is currently working as an Assistant Professor with the Department of Computer Science and Software Engineering, International Islamic University. She has been actively involved in academic and research activities of various levels for the last 15 years.

GRAEME MCROBBIE received the B.Sc. degree (Hons.) in physics and the Ph.D. degree in electronic engineering from the University of Strathclyde, in 1989 and 1992, respectively. He was a Research Fellow with Glasgow University, until 1996. He is currently a Senior Lecturer and a Programme Leader of mobile Web development with the University of the West of Scotland. He is also interested in providing computing solutions to SMEs, and areas of specialism include business intelligence, data management, and Web and mobile development. His current research programmes included the Gamification of BI Tools to enhance User Engagement, Realtime BI for the Mobile Environment, Development of Data Management Frameworks for SMEs, and the Development and Implementation of a Cloud-Based Customer Relations Management Tools for SMEs.

NAEEM RAMZAN (Senior Member, IEEE) received the M.Sc. degree in telecommunications from the University of Brest, France, in 2004, and the Ph.D. degree in electronics engineering from the Queen Mary University of London, London, U.K., in 2008.

He is currently a Full Professor of artificial intelligence and the Director of the Affective and Human Computing for Smart Environment (AHCSE) Research Centre, University of the West of Scotland (UWS), U.K. He has authored or coauthored more than 200 research publications, including journals, book chapters, and standardization contributions. He has authored a book and coedited some books as well. His research interests are cross-disciplinary and industry focused and include AI/machine learning, affective computing and multimedia processing, analysis and communication, video quality evaluation, brain-inspired multi-modal cognitive technology, big data analytics, affective computing, the IoT/smart environments, natural multi-modal human–computer interaction, and eHealth/connected Health.

Dr. Ramzan's article was awarded the Best Paper Award 2017 of the IEEE Transactions on Circuits and Systems for Video Technology and number of conference papers were selected for the Best Student Paper Award. He has been a Lead Researcher in various nationally or EU sponsored multimillion-funded international research projects (total funding as PI secured over £20m). He has been awarded the Scottish Knowledge Exchange Champion Award 2020 and numerous other awards, such as Staff Appreciation and Recognition Scheme (STARS) Award for Leadership in 2019 and awarded STARS award 2015 and 2017 for Outstanding Research and Knowledge Exchange (the University of the West of Scotland) and Awarded Contribution Reward Scheme 2011 and 2009 for outstanding research and teaching activities (the Queen Mary University of London). He is, a Senior Member of the IEEE Fellow, a Senior Fellow of the Higher Education Academy (HEA), the Co-Chair of MPEG HEVC verification (AHG5) Group and a Voting Member of the British Standard Institution (BSI). In addition, he holds key roles in the Video Quality Expert Group (VQEG), such as the Co-Chair of the Ultra High Definition (UltraHD) Group, the Co-Chair of the Visually Lossless Quality Analysis (VLQA) Group, and the Co-Chair of the Psycho-Physiological Quality Assessment (PsyPhyQA). He is also the Co-Editor-in-Chief of VQEG eLetter. He has served as a Guest Editor for a number of journals. He is also a Founding Associate Editor of *Journal of Quality and User Experience* (Springer) and an Associate Editor of number of Journals. He has chaired/co-chaired/organized more than 25 workshops, special sessions, and tracks in International conferences. He has developed a highly innovative portfolio of post graduate studies, including the M.Sc. degrees in advanced computing, big data, the IoT, and eHealth/digital health.

• • •