Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5596312, 10 pages
https://doi.org/10.1155/2021/5596312

Research Article

Machine Learning Model for Group Activity Recognition Based on Discriminative Interaction Contextual Relationship

Smita S. Kulkarni [1,2] **and Sangeeta Jadhav** [3]

[1]*Electronics and Telecommunication, D. Y. Patil College of Engineering, Akurdi, Pune, India*
[2]*Electronics and Telecommunication, MIT Academy of Engineering, Alandi, Pune, India*
[3]*Information Technology, Army Institute of Technology, Dighi, Pune, India*

Correspondence should be addressed to Smita S. Kulkarni; smitak0103@gmail.com

Received 21 February 2021; Revised 26 June 2021; Accepted 22 July 2021; Published 4 August 2021

Academic Editor: Essam Houssein

Copyright © 2021 Smita S. Kulkarni and Sangeeta Jadhav. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper represents the recognition of group activity in public areas, considering personal actions and interactions between people from the field of computer vision. Modeling the interaction relationships between multiple people is essential for recognizing group activity in the video scene. In artificial intelligence applications, identifying group activities based on human interaction is often a challenging task. This paper proposed a model that formulates a group action context (GAC) descriptor. The descriptor was developed by integrating the focal person action descriptor and interaction joint context descriptor of nearby people in the video frame. The model used an efficient optimization principle based on machine learning to learn the discriminative interaction context relations between multiple persons. The proposed novel group action context descriptor is classified by support vector machine (SVM) to recognize group activity. The proposed technique effectiveness is evaluated for group activity recognition by performing experiments on a publicly available collective activity dataset. The proposed approach infers a group action class when multiple persons are together in the video sequence, especially when the interaction between people is confusing. The overall group action recognition model is interrelated with a baseline model to estimate the performance of interaction context information. The experimental result of the proposed group activity recognition model is comparable and outperforms the previous methods.

1. Introduction

Multiple person activity recognition algorithms have established significant attention in the field of computer vision as well as artificial intelligence. However, group activity recognition from video sequences is often a challenging task due to the dynamic interaction between multiple people. Group activity recognition is important in many applications such as computer-human interaction [1], video surveillance [1], content-based video recovery [2], video summarization [3], and healthcare [1]. In surveillance, medical, and social care fields, these algorithms are used to detect abnormal activities in healthcare fields and in public spaces such as air terminal and metro station places. In [4], for recognizing human activities from videos, a computationally storage efficient approach is proposed. In [5], *k*-nearest neighbors' techniques are developed for human activity recognition.

Most traditional methods in the computer vision system are focused on the recognition of an individual person's activities [6–9]. Although several recent works [10–16] have been handling group activities in real time, scenes often involve multiple persons in action along with their interrelated actions. Group activity recognition recognizes actions that are performed by multiple people.

It is normally hard to discriminate the activities of multiple people based on the appearance of an individual person alone. The visual appearance of the highlighted person in Figure 1(a) is just a standing action as an individual. However, the person is waiting in the queue or talking with other persons. The highlighted focal person in

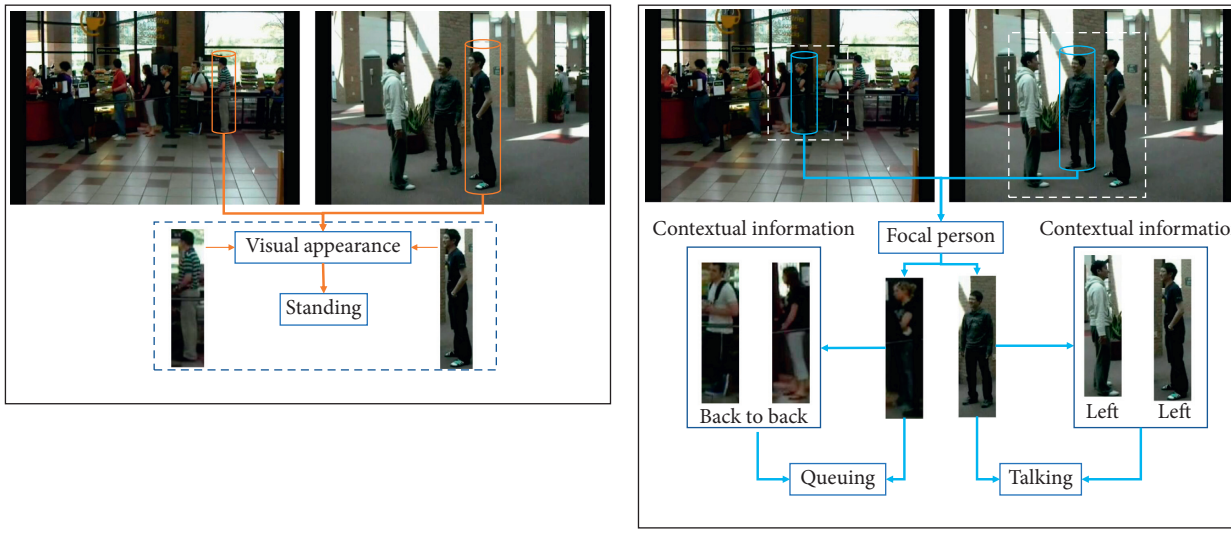

Figure 1: A schematic overview for group activity recognition considering contextual interaction information. (a) Group activity recognition is usually difficult to discriminate based on the visual appearance of only an individual one. (b) This framework considers interaction context information between the focal person and people nearby for group activity recognition.

Figure 1(b) is interrelated with the people nearby. In the interest of group activity recognition, it is essential to deal with the context information between the focal person and surrounding people nearby. Hence, context modeling is necessary for recognizing group activities. The proposed model detects the focal person and interaction context information. In this direction, several researchers are investigating the contextual information to analyze a group activity through interactions between multiple people, denoted as "group activity" or "collective activity" recognition [13–16].

The focus is on modeling the group activity descriptors by integrating focal person's action descriptor and interaction joint context information in the direction of group activity recognition. The performance of the proposed technique is estimated on the collective activity dataset [10]. This approach contributes two main purposes: the first is to improve the misclassification of group activities descriptors by eliminating the confusion of similar actions in the scenes and the second is to recognize the group activities to streamline the interaction inference technique.

Furthermore, the interaction joint context is used to develop an innovative group action context (GAC) descriptor model for efficient group activity recognition process. The proposed approach contributes, first, to the development of an interaction joint context based on Bag-of-Words (BoW) approach representation for individual actions and pose interactions along with the dominant poses and actions in the video within the frame. An algorithm is developed based on the dominant pose and action to determine the interaction within multiple people in the video scenes. Secondly, it proposes a novel group action context descriptor (GAC) that encodes the interaction between joint context and action descriptor of the focal the person. A group action context descriptor is classified through the SVM classifier for group activity recognition.

The rest of the paper is presented as follows. In Section 2, we described review work related to group activity recognition with different approaches. Section 3 explains the detailed discussion of the proposed framework. Section 4 demonstrates the effectiveness of the propped technique through experimental results and evaluations. In Section 5, performance of the proposed approach for a group activity recognition is concluded.

2. Related Work

Recent methods have outperformed in recognition of individual actions [7, 8]. In computer vision, human action recognition has diverse applications in intelligent surveillance, sports analytics systems, etc. In [17], the concept of human activity recognition specifically for video surveillance was explored. In this application activity, understanding is important for improving human-computer interactions. However, recognizing the group actions from multiple people was not restricted to only recognizing the actions of individuals in the group. In recent research work, group activity recognition is based on the actions of persons and the interaction context within multiple people [10, 12–14]. In [18], activity recognition was represented by global representations with local representations approach. The interaction between multiple persons within group most often encodes as context information. Machine learning (ML) techniques encourage an effective framework for modeling the interaction context between multiple people. In [19], various ML approaches for group activity recognition were discussed. Furthermost of the present methods, consider that most of the people present in a scene exhibit

singular activities as group actions. However, this is not true especially in surveillance sports videos. In addition to that, people might show different pose interactions within a group which exhibits a specific activity.

In [10, 11, 16, 20], a multiple person action recognition technique has been discovered in video. In some research work, context information among multiple people has been proposed for group activity recognition. In [10, 21], contextual information is integrated by extracting feature descriptors from multiple persons. This context information is a more significant feature descriptor to analyze the interaction for group activity recognition. However, in this model, the action of everyone is classified independently because this spatial and temporal constancy in the group interaction is not always confirmed. In [12, 14, 15, 22], the proposed graph structure model described the interaction among persons.

There exists contextual pose interaction information which differs the overall group activity as shown in Figure 1. The hierarchical AND-OR graph model is proposed in [20, 21] for group activity, which models temporal and framewise relations in the video. However, this method was expensive to apply.

In [9], spatiotemporal local (STL) descriptor considers spatial variation, and this descriptor generalizes. In [23] RSTV was proposed which captured the context of the person in the crowd but it failed in the noisy pose, hence proposing 3D MRV. In [14], an action context (AC) descriptor was proposed to capture contextual information through HoG feature vector. However, it does not consider the person's posture context interaction information. It considers the action scores of focal persons and all nearby people in the context region. However, this descriptor is sensitive to changes in viewpoint. In [24] the relative action context (RAC) descriptor is proposed, which encodes the relative relations within the activity to represent the viewpoint invariants. In the model of [15] we considered temporal consistency within the group, but the interactions considered limited only in successive frames. Due to this, temporary misclassification results in these models.

In [13, 25], to model interactions between people, a spatiotemporal pattern, hierarchical graphical model was proposed, which involved composite preprocessing and inference processes. As in [14], contextual information is considered only in the adjacent region due to this temporal and spatial uniformity missing due to this misclassification in a group activity classification.

In [26], the interactions between people are integrated through fully connected conditional random fields (CRFs) to avoid misclassification of group actions. These multiscale features are considered which are integrated through CRF to represent the interaction context. The approach in [27] uses a model of human behavior considering semilocal parts and interactions between them, by which the classified multiclass activities developed reasonable functionality. In [28], a graph-based clustering method was proposed for recognition of group action in a crowded scene by considering motion and local interaction information.

However, it is very difficult to handle complex interaction context information based on graphical models. This approach is competent only for modeling human level trajectory info, which is inadequate to recognize confused group activity such walking and crossing. These activities can be recognized by human action and pose appearance.

Fan et al. [29] offer technique for understanding human gaze communication by studying human interaction in social videos. In this design a spatiotemporal graph neural network is used to model dynamic human interaction by passing messages over the graph. To capture the temporal dynamics LSTM based temporal reasoning module is incorporated to predict atomic gaze communication. Paper [30] detected shared attention intervals spatially and predicted shared attention location in video frames by proposing spatial-temporal neural network. The convolutional Long Short-Term Memory network is employed to optimize temporal domain in the shared attention intervals. The Graph Parsing Neural Network (GPNN) is a framework proposed in [31] for detecting and recognizing human-object interactions (HOI) in images and videos. The proposed GPNN signifies HOI structure and automatically analyzes the optimal graph structures and this method is valid for spatial and spatial-temporal domains.

In computer vision deep learning approaches a significant improvement in image classification, human activity recognition, and video classification is shown. The deep neural network learning model is presented in [32] for recognizing the activities performed by multiple people based on contextual relationships. In surveillance scenes for group activity recognition, Deng et al. [33] integrated hierarchical graphical models and deep networks. In [34], deep LSTM-based temporal hierarchical structure model was proposed to learn sport activity data and in [35] Confidence-Energy Recurrent Network (CERN) encompasses two-level hierarchy of LSTMs. In recent work of group activity recognition, Tora et al. [36] proposed pretrained CNN model combined with LSTM recurrent neural network to capture interaction context information. Paper [37] developed multilevel hierarchical recurrent network to model interaction context framework for group activity recognition. The power of deep learning RNN model [38] captured person-level temporal context information. The interaction related to long motion time of individual is aggregated by Bi-LSTM by proposing a novel Participation-Contributed Temporal Dynamic Model (PC-TDM) in [39] which improves performance of group activity recognition. Multistream spatiotemporal architecture by a convolutional fusion is proposed in [40] for collective activity recognition. Tang et al. [41] proposed Coherence Constrained Graph LSTM (CCGLSTM) to model the relevant motions of individuals to effectively recognize group activity, by suppressing the irrelevant motions. The problem of group activity recognition is solved by modeling person-level and group-level actions

in [42] by proposing graph LSTM framework by exploiting temporal features.

The existing research work presented is based on a learned handcrafted feature descriptor. The techniques are evaluated in the direction of context modeling. It has been observed that the context descriptor considers spatiotemporal features which improves the classification of group activity. Most of the previous group activity recognition methods do not handle flexible interactional context information. Owing to this in this research group action context descriptor is formulated using joint interaction context information for recognition of group activity.

The experimental results presented in the paper outperform the graph structure models [14, 23, 26]. The proposed model of group activity recognition has focused on interaction context. Context information considered how the focal person connected with nearby person's actions and pose for group activity. This model produces context modeling which has discriminative interactional features to handle varying number of persons in a group and is flexible to model scalable context information.

3. Approach

Group actions are categorized by pose as well as actions of persons along with interactions within multiple people. The interaction context information exhibits an important role in recognizing the group action. However, it poses a challenging problem owing to the change in people's actions and more precisely the variation in the human pose which exhibits variation in interactions within the group action. In the group activity, the recognition key purpose is to ensure the positional appearance of an individual through interaction context cues in each group.

This section describes the strategy of the group activity recognition method. Thus, the group action context (GAC) descriptor is formulated from the people interaction in a scene and then this descriptor is classified into group activity category by using a multiclass SVM classifier.

The proposed method constructs GAC by combining the focal person action descriptor and joint interaction context descriptor. However, it is assumed that the head poses of people and 3D trajectory space are available in the database [10]. The proposed framework of group action recognition is presented in Figure 2.

As shown in Figure 2, the persons are detected in the video frame and let I denote a person. In the center of the video frame one person I_m is detected as the focal person in the frame, and people nearby in the region of I_m can be considered for the interaction joint context as J_m. In each frame, the focal person is selected and corresponding to its interaction, the joint context feature is computed. The proposed group action context descriptor G_y^m learns through weighted function W_c, between the focal person action descriptor I_m and interaction joint context descriptor J_m. In the proposed group interaction model, the assumption is that the focal person action descriptor would be extremely related to the interaction context as a group action, which is affected by the multiple people pose and actions in the video frame.

In the following, model, the formulation is illustrated in Section 4 that learns an optimal W_c, by optimizing the model for the inference of activity. Before that, Section 3.1 describes the features in detail.

3.1. Feature Details

3.1.1. Focal Person Action Descriptor. It has been assumed that the video frame is preprocessed, and persons are detected along with available locations [10]. The feature is extracted from the detected person using histogram oriented gradient (HoG). HoG is [43] an appearance-based feature vector which is an anticipated technique in complex environments to diminish occlusion and illumination variations for individual action recognition. Owing to this, HoG transform is used to extract the feature from the detected person. However, instead of directly using raw HoG features, here is anticipated the individual action descriptor [13]. In the proposed framework for individual action descriptors, the KNN classifier is trained on HoG features for 5 action classes. In the center of the frame, focal person action descriptor is selected as I_m, where I_m is a vector with a classification score for the 5 action classes.

3.1.2. Pose Context Feature. In support of this spatial information around a person, we explore by proposing BoW as a pose feature *Pi* which includes eight pose categories: right, front-right, front, front-left, left, back-left, back, and back-right. In addition to this most influential pose feature is also considered in the video frame that imitates the interaction context between the focal person I_m and nearby person I_i, $i \neq m$ in the region of I_m. This influential pose feature is discovered as the contextual relationship between the focal person's pose and nearby people.

Let B_{mi} represent pose context feature vector, which is formulated in equation (1) by contacting the most influential pose feature DP_{mi} and BoW of nearby people's pose P_i surrounded to the focal person I_m, where $i \neq m$:

$$B_{mi} = [DP_{mi}; P_i]. \quad (1)$$

3.1.3. Interaction Joint Context Descriptor. In a video frame, for example, consider multiple people with two actions, crossing and walking, in which sometimes they have the same collective pose representing that persons are following each other. In such a case, it is significant to integrate individual actions and interaction context information between these activities to discriminate them. To investigate these complex structures which involve spatial dependencies, the interaction joint context plays a significant role. Thus, to capture the interactions within the group, we propose an interaction joint context descriptor which encodes pose context features as well as individual action descriptor. The proposed interaction joint context model

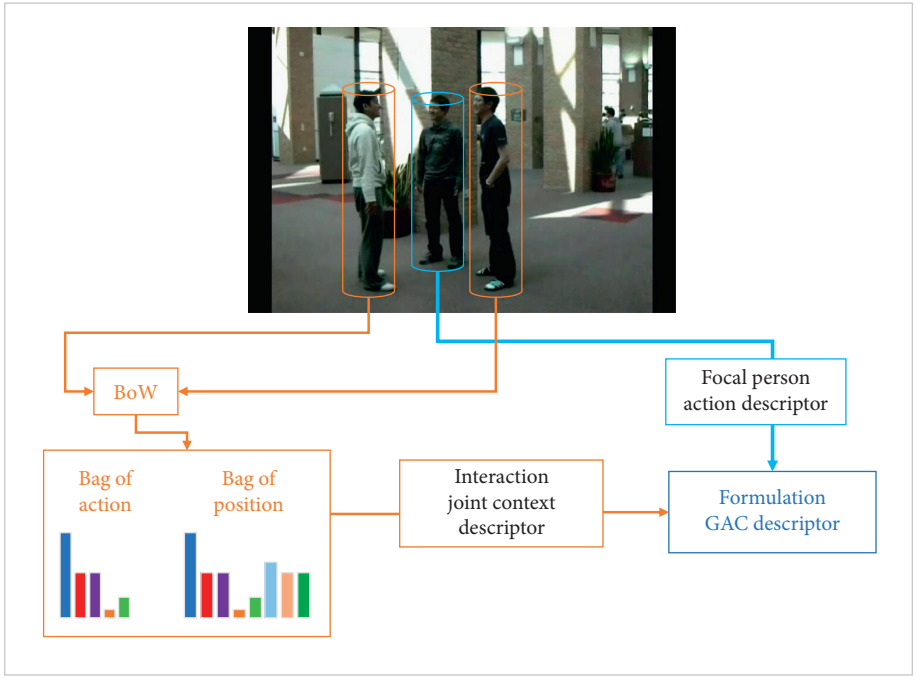

Figure 2: Overview of the proposed approach. (1) Interaction joint context developed by Bag-of-Words (BoW) approach. (2) Group action context descriptor formulation.

formulates BoW illustration for the pose and individual actions of a person in the video frame. In addition, the dominant pose and actions present in the frame are involved in the descriptor. The interaction joint context descriptor is described as

$$J_{mi} = [B_{mi};\ A_i], \tag{2}$$

where B_{mi} is a pose context feature and A_i is BoW of actions as shown in Figure 2. Hence, it helps to consider spatial relations between people which are capable of discriminating against the activities.

3.1.4. Group Action Context Descriptor. The group action context (GAC) descriptor is formulated by integrating the focal person action descriptor and interaction joint context descriptor of nearby people. Group activities involved individuals' actions and positional movements which effect interaction context information. Hence, it is essential to encode contextual features into a novel group action context descriptor.

As shown in Figure 2, proposed GAC descriptor is focused on a focal person and illustrates the relative interaction context between the focal person and other people nearby. In the interest of group activity recognition, it is required to have an effective group action context descriptor which formulates the interaction between persons. In addition, GAC encodes the discriminative information of everyone in the group activity.

The significance of the GAC descriptor is essential which formulates the interaction context future which enhances the performance of the classification algorithm. GAC represents a contextual relationship between the focal person and people nearby in BoW style, which captures actions and pose features. Thus, GAC captures spatiotemporal dynamic information which supports enhancing the learning performance of SVM classifier to recognize group activity.

4. Model Formulation

The person in the frame has been detected with the given location of persons [10]. In addition, to train the descriptor data, in a supervised learning mode, each person in the frame is labeled by action and pose labels.

With multiple people, interaction signifies cooccurrence between the actions of all individuals and the positional head pose to form a group action. The main purpose of the proposed approach is to estimate the group action context descriptors through the interaction context for the group activity recognition.

Let $\{V_m\}$ be the video sequence captured with a camera and having M set of images.

Let $I = \{I_i^K : i = 1, \ldots, m\}$ be the set of m persons in k^{th} frame. Having this information, the main aim is to extract interaction context features between multiple people. The HoG feature descriptor is extracted from the person I as x_i, which is trained using KNN classifier to develop individual action descriptors.

In the frame all the person actions denote $h = (h_1, h_2, \ldots, h_m)$ as action label where $h \in H$, and H is possible action label set. The video frame K is allied through a group activity label as $y \in Y$, where y is possible to group activity labels set. The focal person action descriptor I_m is selected at the center of frame K. The action descriptor for focal person I_m is given as

$I_m = [S1(i); S2(i); \ldots; SA(i)]$, for h possible action where $SA(i)$ is the classification probability score of people I_m to action h.

The group action context (GAC) descriptor is computed as follows:

$$G_y^m = I_m * W_c * J_m, \tag{3}$$

where G_y^m can be observed as the interaction function between I_m & J_m which is optimized by W_c as weight matrix and the weighted relationship between the focal person and interaction joint context Jm is learned. If there is a maximum correlation among focal person descriptor and interaction context for the given y^{th} group activity owing to G_y^m being maximum or else, it is not in equation (4). By considering the interaction modeling function among a set of m persons,

$$\begin{aligned} k &= \arg \max_{C \in 1,\ldots,y} G_y^m, \\ G_y^m &= \sum_{m=1}^{M} G_C^m. \end{aligned} \tag{4}$$

4.1. Learning. In each video sequence, the aim is to recognize group activities. Each video frame is signified as a group action context descriptor that encodes the focal person action descriptor along with the surrounding person interaction joint context. The GAC descriptor implicitly infers group actions during learning and inference. There are N training frames, $\{i \in 1, \ldots, N\}$ and $\{y_i \in 1, \ldots, Y\}$ group activity label which belongs to that video frame. In equation (3), the group action context descriptor is learned through weight matrix W_c by integrating the focal person's action descriptor Im and it is surrounding an interaction joint context descriptor J_m. The matrix W_c should represent the interaction context structure within a group activity. The SVM multiclass [28] classifier is trained on GAC descriptor by optimizing W_c through parameter tuning for correct group activity recognition.

Assume there are m people in the video frame K and the corresponding focal person and interaction joint context are I_m and J_m, correspondingly. Then, equation (5) optimizes the group interaction context response concerning a group activity class Y:

$$G_y^K = \sum_{m=1}^{MK} \left(I_m^k\right) * W_c * J_m^k. \tag{5}$$

Multiclass linear SVM classifier is trained on GAC which is represented as the feature descriptor to learn the weighted matrix W_c to classify group activities for the video sequence once model is trained. Two-category ground-truth response of group activity to other group activities defines a margin δ in

$$\sum_{m=1}^{MK} \left(I_m^k\right) * W_c * J_m^k > \delta. \tag{6}$$

The ultimate formulation of the model becomes

$$\begin{aligned} &\min_{\Delta_{kl}\xi^K} -\frac{1}{2}\left\|\Delta_{kl}\right\|^2 + \sum_{st=k} \xi^K, \\ &\text{s.t.} \sum_{m=1}^{MK} \left(I_m^k\right) * w_c * J_m^k > 1 - \xi^K, \quad \xi^K \geq 0. \end{aligned} \tag{7}$$

Based on the above formulation model, optimize all Δ_{kl}, where $k, l \in 1, 2, \ldots, Y,\ k \neq l$.

In equation (7) Lagrangian Relaxation (also known as Dual Decomposition) constraints are employed on the model which solves by optimization cost function Δ_{kl}.

4.2. Inference. In equation (3) the interaction context response is computed based on the weight matrix Wc by developing the voting scheme for each group activity. $\Delta_{kl} = w_k - w_l = -\Delta_{lk}$ is used to optimize all the weight matrices W_C, where $C \in \{1, 2, \ldots, Y\}$ through learning classifier $k \neq l$.

In frame K, compute Bag-of-words for pose and action labels for an individual and calculate the consistent joint interaction context J_m across the focal person I_m. Then, we compute the value

$$\sum_{m=1}^{MK} \left(I_m^k\right) * \Delta_{kl} * J_m^k. \tag{8}$$

In equation (8), Δ_{kl}, on each k, $l \in \{1, 2, ..., Y\}$, $k \neq l$ *is* computed for recognition of group activity from a video sequence V_m. If equation (6) score is greater than zero, it gets the maximum votes, which means more likely from y^{th} group activity class.

5. Experiment and Results

This section describes the performance results of the proposed GAC descriptor model for group activity recognition.

5.1. Dataset. In this paper, for experimentation, Collective Activity Dataset is selected that is proposed in [10]. The dataset provides automatic person detection and trajectory generation and represents real-time noisy occlusion observations. The proposed dataset is labeled around bounding boxes of the person carrying out an action with their pose and activity class for the recognition purpose of every 10th frame. The proposed dataset contains video frames of five group activity classes, together with waiting, walking, crossing, talking, and queuing, and eight poses with right, front-right, front, front-left, left, back-left, back, and back-right. In the dataset, 44 short video sequences are involved with different multiple people's actions. The performance of the proposed interaction contextual relationship was validated on 5 group action category datasets [10].

The proposed framework's focus is group activity recognition and interaction contextual modeling is used to improve the performance of the GAC classifier model. Although it is observed that optimization improves action recognition through interaction joint context, in the experimentation leave-one-video-out, a cross-validation scheme [22] is performed.

5.2. Experimentation. Group action classification: based on Sections 3.1.1–3.1.4, the obtained feature context descriptor set can be sent to the SVM classifier for training and testing, and the performance of the model is estimated. In the learning and inference phase of the classifier, it is essential to provide a random and independent feature descriptor set. Owing to this, to ensure SVM classifier accuracy, performance in the experimentation feature set of video sequence is divided into 70% training set and 30% testing set randomly and independently. After training SVM classifier based on 10-fold cross-validation group action classification performs on the test data feature set for 5 action classes of crossing, waiting, queuing, walking, and talking. To evaluate the performance of the model, following performance indicators are investigated: accuracy, precision, recall, *F*-1 score, and confusion matrix and compared with baseline model.

5.3. Results. The experimental results are compared with existing techniques in this section. This paper found that the proposed model achieves improved performance compared to [14, 23, 26] as shown in Table 1. In Figure 3, the recognition accuracy for each activity is listed by confusion matrix. It presented that the proposed GAC model achieves a significant improvement compared to the baseline model. Additionally, the proposed GAC descriptor optimization method can capture the diverse forms of interaction context in group actions.

The GAC model is trained with a SVM classifier intended for group action recognition by utilizing the libSVM library [44]. The efficient parameter tuning of SVM had significant improvement on the classifier's accuracy. The group descriptor is robust as in the existence of noisy observations since the descriptor builds on the interaction context. The proposed model average classification accuracy is 88.8% on collectivity activity dataset [10] shown in Table 1 and compared with state-of-art methods.

To evaluate the performance of the model, following performance indicators are investigated: accuracy, precision, recall, and *F*1 score, compared with baseline model in Table 2.

The proposed GAC descriptor exploited context information in terms of focal person, action descriptor, and interaction join context descriptor. The proposed framework automatically infers person interaction context information through an optimal GAC descriptor. In Table 2, it is observed that GAC descriptor improves the performance compared to baseline model. In the interaction, the joint context encodes the information about the focal person actions and multiple people poses and actions in BoW formation. This context feature-level method offers an efficient way to comprise the interaction context temporally and spatially. In the joint context, the descriptor considers the dominant pose and the action of nearby people, which is a benefit for the activity which is not discriminative based on the pose of multiple people (e.g., walking). In this case, bag-off word technique shows promising performance. Thus, formulating an interaction context descriptor gives more intellect for diverse group activities.

In Figure 4 we reported qualitative results on a 5-activity dataset indicating group action using the proposed method. In Figure 4, the first two sequences successfully recognized the action and improved the classification results, whereas the final row represents a failure in classification due to wrong action label recognition which causes misclassification of activity. The group activity classification result is visualized in Figure 4, which is the learned GAC structure of person interaction context descriptor by SVM Machine Learning algorithm. Note that the person in red color made an incorrect prediction for individual action. Owing to this in this frame there is an incorrect classification, which reduces the performance of the model.

5.4. Discussion on Result. The proposed model performance is evaluated along with state-of-art methods and baseline models. The baseline model uses only an interaction joint context descriptor. The GAC framework demonstrated that joint interaction context capturing performs successful group activity classification using SVM. The interaction descriptor considers a BoW representation of a person's actions and pose in the frame. This paper mainly proposed a group action context descriptor that extracts contextual information between the focal person and people nearby. Owing to this, the GAC model achieves improvement in performance in Table 1 over the baseline methods for group activity recognition. The results of AC descriptor [14], RSTV + MRF [23], and AC-RAC + FC-CRF [26] used for group activity classification improved by the proposed GAC descriptor method. In [14] AC descriptor considers the action score in the context region although the interaction pose context does not consider it. In the proposed GAC descriptor, consider the pose context among multiple people, which improves the classification performance.

It is observed that in [14, 26] the classification between walking and crossing was ambiguous. The proposed model inference techniques are improved performance of recognizing group activity that is involving confusion in activities like walking, waiting, and crossing. By integrating interaction contextual information confusion between these activities reduced because the pose context feature fetches important cues for discriminating these activities. As a pose context feature, consider BoW as the pose features along with the most influential pose feature in a group of people. In crossing, persons always cross the street with the same pose and in waiting people stand in the same pose direction, hardly their posture facing each other. However, by considering the most influential pose and action feature, the performance of group activity classification accuracy improved. This implies that the GAC contextual modeling

Table 1: Performance of the proposed method compared with the state-of-art methods for group activity recognition accuracy on collective activity dataset [10].

Method/Activities	Crossing (%)	Waiting (%)	Queuing (%)	Walking (%)	Talking (%)	Accuracy (%)
AC [13]	68	69	76	80	99	78.4
RSTV + MRF [23]	76.4	76.7	78.7	36.8	85.7	70.9
AC + FC-CRF [26]	67	84	86	49	75	72.2
Interaction joint context descriptor	68.3	76.2	83.1	80.6	86	78.84
Proposed GAC descriptor	85.2	87.1	87	84.4	98.2	88.38

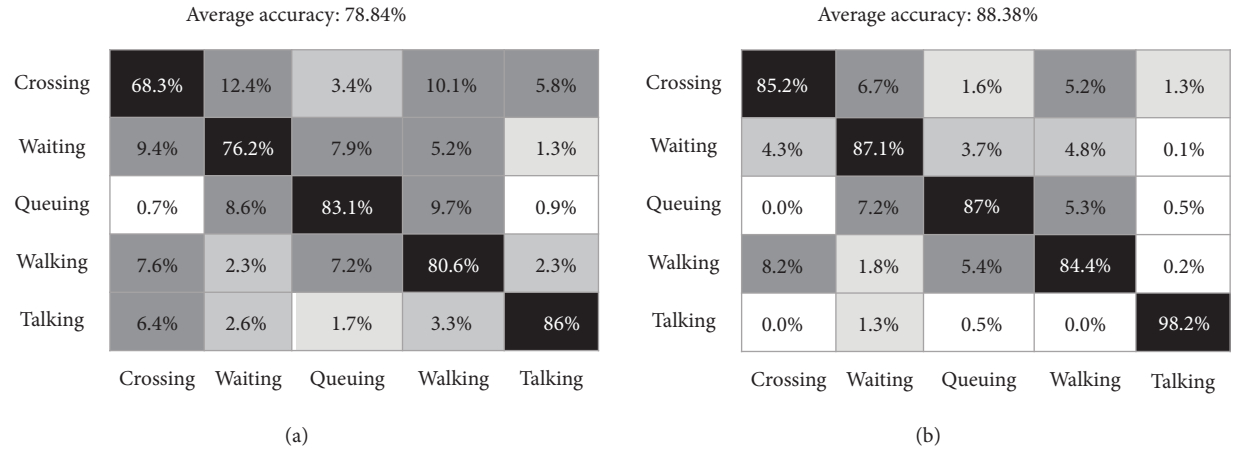

Figure 3: Present accuracy for 5-class [10] group activity through confusion matrix. (a) Baseline Model. (b) GAC descriptor.

Table 2: Performance of GAC descriptors compared with baseline approach.

Method/activities	Accuracy (%)	Precision (%)	Recall (%)	$F1$ score (%)
Interaction joint context descriptor	68.3	80.76	80.89	71.67
Proposed GAC descriptor	85.2	88.22	88.22	80.82

Figure 4: Proposed model visualizations results from the collective activity dataset [10]. Green color bounding box indicates the correct classification of group activity class (crossing, waiting, queuing, walking, and talking) and the red bounding box recognized with the different activities which affect the recognition accuracy performance.

between the focal person and the nearby person is an efficient mechanism for interaction context information detection among multiple people.

6. Conclusion

This paper presents a model for group activity recognition in video. This group activity recognition task was resolved by considering interaction contextual information between multiple people. Based on that, a novel group action context (GAC) descriptor is proposed to model the interaction context between focal person actions and nearby people within a group activity. The group descriptor incorporates focal person action and interaction joint context information to discriminate different group activities. The GAC descriptor model infers group activities efficiently by establishing an effective optimization algorithm SVM. The best average accuracy of 88.8% of the proposed model has shown significant performance as compared to the state-of-art methods on collective activity dataset for group activity recognition. The proposed algorithm utilizes interaction joint context information which can be effective for the development of group action context descriptor. Furthermore, the GAC descriptor is robust for confusing activities such as crossing and walking. In surveillance applications for high-level activity and behavior analysis, the proposed model is rendered easily. Future scope includes investigation of other useful context information such as scene context and research on effective automated context feature learning.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank Professor Sangeeta Jadhav for her valuable supervision.

References

[1] X. Yang and Y.L. Tian, "Super normal vector for activity recognition using depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 804–811, Columbus, OH, USA, June 2014.

[2] L. Yao, Y. Liu, and S. Huang, "Spatio-temporal information for human action recognition," *Journal on Image and Video Processing*, vol. 39, 2016.

[3] S. N. Gowda, "Human activity recognition using combinatorial deep belief networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6, Honolulu, HI, USA, July 2017.

[4] P. Sujitha and P. Simon., "A computationally efficient method for human activity recognition based on spatio temporal cuboid and super normal vector," *Journal of Intelligent & Fuzzy Systems Preprint*, vol. 40, pp. 1–9, 2020.

[5] J. M. Cadenas, M. C. Garrido, R. Martinez-España, and A. Muñoz, "A *k*-nearest neighbors based approach applied to more realistic activity recognition datasets," *Journal of Ambient Intelligence and Smart Environments*, vol. 10, no. 3, pp. 247–259, 2018.

[6] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[7] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[8] Y. Yang Wang and G. Mori, "Human action recognition by semilatent topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762–1774, 2009.

[9] W. Xiao and Y. Lu, "Daily human physical activity recognition based on kernel discriminant analysis and extreme learning machine," *Journal of Mathematical Problems in Engineering*, vol. 2015, Article ID 790412, 2015.

[10] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proceedings of the IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1282–1289, IEEE, Kyoto, Japan, September 2009.

[11] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 1593–1600, IEEE, Kyoto, Japan, September 2009.

[12] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *Advances in Neural Information Processing Systems*, pp. 1216–1224, Springer, Berlin, Germany, 2010.

[13] T. Lan, Y. Wang, G. Mori, and N. R. Stephen, "Retrieving actions in group contexts," in *Proceedings of the European Conference on Computer Vision*, pp. 181–194, Springer, Crete, Greece, September 2010.

[14] T. Lan, Y. Wang, W. Yang, N. R. Stephen, and M. Greg, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–1562, 2011.

[15] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 215–230, Springer, Florence, Italy, October 2012.

[16] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 187–200, Springer, Florence, Italy, October 2012.

[17] R. M. Raval, H. B. Prajapati, and V. K. Dabhi, "Survey and analysis of human activity recognition in surveillance videos," *Intelligent Decision Technologies*, vol. 13, no. 2, pp. 271–294, 2019.

[18] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of Healthcare Engineering*, vol. 2017, Article ID 3090343, 2017.

[19] S. Kulkarni, S. Jadhav, and D. Adhikari, "A survey on human group activity recognition by analysing person action from

video sequences using machine learning techniques," in *Optimization in Machine Learning and Applications*, pp. 141–153, Springer, Berlin, Germany, 2020.

[20] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu, "Monte Carlo tree search for scheduling activity recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1353–1360, Sydney, Australia, December 2013.

[21] M. R. Amer, P. Lei, and S. Todorovic, "Hirf: hierarchical random field for collective activity recognition in videos," in *Proceedings of the European Conference on Computer Vision*, pp. 572–585, Springer, Zurich, Switzerland, September 2014.

[22] S. Khamis, V. I. Morariu, and L. S. Davis, "A flow model for joint action recognition and identity maintenance," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1218–1225, IEEE, Providence, RI, USA, June 2012.

[23] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proceedings of the CVPR 2011*, pp. 3273–3280, IEEE, Colorado Springs, CO, USA, June 2011.

[24] T. Kaneko, M. Shimosaka, S. Odashima, R. Fukui, and T. Sato, "Viewpoint invariant collective activity recognition with relative action context," in *Proceedings of the European Conference on Computer Vision*, pp. 253–262, Springer, Florence, Italy, October 2012.

[25] S. Khamis, V. I. Morariu, and L. S. Davis, "Combining per-frame and per-track cues for multi-person action recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 116–129, Springer, Florence, Italy, October 2012.

[26] T. Kaneko, M. Shimosaka, S. Odashima, R. Fukui, and T. Sato, "A fully connected model for consistent collective activity recognition in videos," *Pattern Recognition Letters*, vol. 43, pp. 109–118, 2014.

[27] B. Antic and B. Ommer, "Learning latent constituents for recognition of group activities in video," in *Proceedings of the European Conference on Computer Vision*, pp. 33–47, Springer, Zurich, Switzerland, September 2014.

[28] K. N. Tran, A. Gala, I. A. Kakadiaris, and S. K. Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," *Pattern Recognition Letters*, vol. 44, pp. 49–57, 2014.

[29] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5724–5733, Seoul, South Korea, October 2019.

[30] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6460–6468, Salt Lake City, UT, USA, June 2018.

[31] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 401–417, Munich, Germany, September 2018.

[32] S. A. Vahora and N. C. Chauhan, "Deep neural network model for group activity recognition using contextual relationship," *Engineering Science and Technology, an International Journal*, vol. 22, no. 1, pp. 47–54, 2019.

[33] Z. Deng, M. Zhai, L. Chen et al., "Deep structured models for group activity recognition," 2015, http://arxiv.org/abs/1506.04191.

[34] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1971–1980, Las Vegas, NV, USA, June 2016.

[35] T. Shu, S. Todorovic, and S.-C. Zhu, "CERN: Confidence-energy recurrent network for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5523–5531, Honolulu, HI, USA, July 2017.

[36] M. R. Tora, J. Chen, J. James, and Little, "Classification of puck possession events in ice hockey," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 147–154, IEEE, Honolulu, HI, USA, July 2017.

[37] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3048–3056, Honolulu, HI, USA, July 2017.

[38] T. Bagautdinov, A. Alexandre, F. Fleuret, F. Pascal, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4315–4324, Honolulu, HI, USA, July 2017.

[39] R. Yan, J. Tang, X. Shu, Z. Li, and T. Qi, "Participation-contributed temporal dynamic model for group activity recognition," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1292–1300, Seoul, Korea, October 2018.

[40] C. Zalluhoglu and N. Ikizler-Cinbis, "Region based multi-stream convolutional neural networks for collective activity recognition," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 170–179, 2019.

[41] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph LSTM for group activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[42] X. Shu, J. Tang, G. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1110–1118, 2021.

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886–893, IEEE, San Diego, CA, USA, June 2005.

[44] R.-En Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.