Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Optimised ARG based Group Activity Recognition for Video Understanding

Pranjal Kumar (pranjal@nith.ac.in)
NIT Hamirpur: National Institute of Technology Hamirpur https://orcid.org/0000-0002-6167-9884
Siddhartha Chauhan
NIT Hamirpur: National Institute of Technology Hamirpur

Research Article

Keywords: group activity recognition, graph convolution network, video understanding, video analytics, activity recognition

Posted Date: March 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-277936/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Noname manuscript No.
(will be inserted by the editor)

Optimised ARG based Group Activity Recognition for Video Understanding

Pranjal Kumar · Siddhartha Chauhan

Received: date / Accepted: date

Abstract Video understanding identifies and classifies various actions and events in the video. Many previous works, such as video annotations, have demonstrated promising success in generating general video understanding. However, a fine summary of human activities and their interactions using state-of-the-art video captioning techniques is still difficult to produce. The comprehensive explanation of human actions and collective behaviors is important information for real time CCTV video tracking, medical treatment, sports video analysis etc. This research suggests a form of video understanding that focuses primarily on identifying group activity by learning the similarities between the pair and the actors appearance. In order to measure the similarity between the pair appearances and construct an actor relations map, the Zero Mean Normalized Cross-Correlation (ZNCC) and the zero-mean sum of absolute differences(ZSAD) are proposed to allow the graph convolution network (GCN) to learn how to distinguish group actions. We recommend that MNASNet be used as the backbone to retrieve features from any video frame. A visualization model is also developed to visualize every input video frame and predict individual behavior or collective activity with projected bounding boxes on a human object.

Keywords group activity recognition · graph convolution network · video understanding · video analytics · activity recognition

1 Introduction

A deeply studied topic in the field of video content analysis is video understanding[1]. Current neural networks, specifically LSTM[2], are utilized for training

Pranjal Kumar
NIT Hamirpur, H.P, India-177005
Tel.: +918637511985
Fax: +91-1972-223834
E-mail: pranjal@nith.ac.in

Siddhartha Chauhan
E-mail: sid@nith.ac.in
NIT Hamirpur, H.P, India-177005

modeling with video-sentence pairs [1,3–5] by traditional video captioning techniques such as LSTM-YT[3] and S2VT[4]. The models learn how to connect the sequence of video frames with the sequence of the sentences in order to generate a video description[4]. Krishna et al. stated that these video subtitles work with only one significant event for a short video[5].They have therefore introduced a new module to describe all events in a video clip using contextual information from the timeline[5]. However, the description of human actions and their interactions in the form of video captions remains very limited[6]. The capability to describe further human actions and group actions[7,8] is demonstrated by recent studies in pose estimation, human action recognition and group activity recognition.

A major problem in video understanding is recognition of human action and recognition of group activities[9]. The techniques of action and activity recognition have been widely used, for example in the fields of social behavior understanding, sport video analysis and video monitoring. It is important to better understand a video scene with several people and to understand the action and collective activity of all individuals. The group activity recognition based on the Actor Relation Graph (ARG) is the state-of-the-art model that aims to capture the appearance and position of the actors in the scene, and to identify the action and the group activity[9]. In this paper, we suggest several ways to improve the functionality and efficiency of the Actor Relation Graph model in order to improve video understanding, which is focused primarily on group activity recognition. To enhance human actions and group activity recognition, MNASNet is used in the CNN layer and to calculate the pair-wise appearance similarity in building an Actor Relationship Graph, Zero Mean Normalized Cross-Correlation (ZNCC) and the zero-mean sum of absolute differences (ZSAD) is used. We also implement a visualization model, which shows each video input frame and predicts bounding boxes on each human object.

2 Related Work

2.1 Video Captioning

Video captioning is an important field of study for video understanding. In 2015, S Venugopalan et al. proposed an end to end model that used the recurrent neural network (LSTM [2]) to use video sentence pairs to create a sequence of images in a video with sequential words to describe the event as captions[2]) and to develop a sequence that would combine videos and sequential words. To learn the temporal structure and sequence model of the generated sentences of the frames a stack of two LSTMs was employed. The whole video sequence must be encrypted in this approach using an initial LSTM network. Long video sequences might lead to gradient vanishing problem and avoid successful training of the model[5].

R. Krishna et al. introduced a model for Dense Captioning Events (DCE) which detects several events and provides a description for each event using past, current and future contextual information in one single video pass[5]. The process is divided in two steps in this paper: event detection and event description. In order to localize temporal proposals of interest to short and long video sequences, the DCE model utilizes a multi-stage variant of the deep actions predictor model. A LSTM captioning model is also introduced to use the historical and future contexts with

a mechanism for attention.
X.P. Li et al. have developed a novel attention-based framework, called LSTM (Res-ATT [10]). This new model takes advantage of the existing mechanism of attention and integrates residual mapping into a two level LSTM network to avoid the loss of word information previously generated. The decoder model based on a residual attention has five different parts: a sentence encoder, time attention, visual and sentence fusion layer, residual layer and MLP[10]. The phrase encoder is an LSTM layer exploring important syntactic information from the phrase. Temporal attention is designed to recognize the significance of each frame. The LSTM Visual and sentence fusion Layer works to mix natural language information with image characteristics and the remaining layer is proposed to lower the transmission loss. The MLP layer predicts the word to create a natural language description[10].

2.2 Pose Estimation Activity Recognition

It is essential to understand the actions of each individual to better understand a video scene which includes several people. OpenPose is a real-time open-source system used to detect pose in 2D multi-persons[8]. It is now widely used in video frames[11, 7, 12] for body and facial detection of markers. It generates a spatial encoding for a variable number of people, followed by a greedy bipartisan graph that matches all people on the image with the 2D keypoints. This approach refines both the prediction of PAFs and the detection of trusted maps in each phase[8]. This improves real-time performance while maintaining each component's accuracy separately. The OpenPose online library collected on one single image, provides 2D-Human-pose-Evaluation for our proposed system to detect the body, hand, and facial keypoints.
The Long Short Term Memory (LSTM) recurrent neural networks (RNNs) are frequently used for the recognition of human actions through the emerging methods of human activity recognition. F. M. Noori et al. proposes an approach in which, after considering movements on the video frames, the anatomical keypoint is first removed from the RGB images using the OpenPose library and then the features of RNN are classified into the corresponding LSTM activities [7]. Improved performance is demonstrated by the organization of various subjects from different angles. However, their efforts to improve accuracy are still underway on the basis of their multi-person action classification.

2.3 Actor Relation Graphs for Group Activity Recognition

A major problem in video understanding is recognition of human action and recognition of group activities[9]. J Wu et al. suggested that Actor Relational Graph (ARG) be used to model actor-to-actor relationships and so that group activity learning be recognized with multiple participants[9]. The relationship between actors from a similarity to the appearance and the relative location is determined and captured using the ARG in a multi-person scene. When compared to the use of CNN to extract person-level features and then add them into a scene-level feature, or the application of RNN to collect temporal information from densely sampled frames, the computationally costly and flexible way of learning with ARG is less

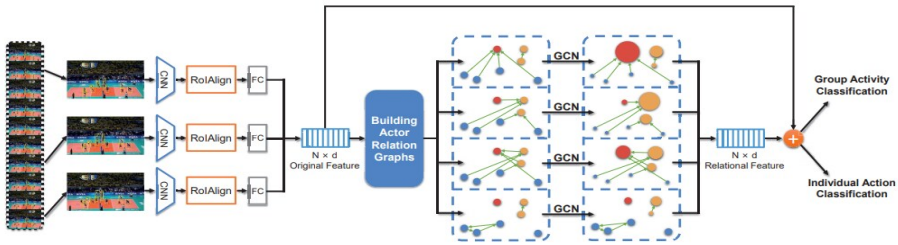

Fig. 1 ARG-based group activity recognition framework.

while addressing variations in group activity. With a video sequence with bounds for actors in the scene, the trained network can recognize individual actions and group activity in a multiperson scene. The network can also recognize action. ARG efficiency is improved for long range video clips by forcing a relational connection in a local neighborhood alone and by dropping several frames alone while maintaining the diversity of training samples and reducing the risk of overfitting. Initially, the features of the actors will be extracted from the provided bounding boxes by model CNN and RoIAlign[13]. After the feature vectors of actors in the scene have been obtained, several graphs are created to represent the diverse information of the same set of actors. Finally, the GCN is designed to perform learning in order to identify the actions and activities of individual groups based on ARG. The pooled ARG applies respectively to two classifiers used for each action and group recognition activity. The representation in scene-level is generated by maxpooling individual actors, that later use them for classification of groups of activities.

3 Proposed Method

We propose to use an improved model based on the Actor relations graph, which focuses on recognition of group activity. The overview is shown in Fig. 1 of the original ARG model.

Firstly, the model extracts the actor from sampled video frames with CNN and RoIAlign bounding boxes[14]. It then builds a N by d feature matrix with a d dimensional vector to represent each bounding box of the actor and N to present the number of video frame bounding boxes. The graphs of the actor relation are then constructed to capture the appearance and position of each actor in the scene. The model will then analyze from the ARG the relations of each actor using Graph Convolutional Networks (GCN). Finally, two distinct classifiers aggregate and utilize the original and relational features to carry out actions and recognition of group activities[9]. As the study focuses mainly on the recognition of group activities, individual actions are not very accurate since the model uses the area of interest and CNN only to recognize actions.
Although high accuracy predictions of group activities can be made using the ARG model. Some areas for improvement still exist. Figure 2 illustrates the improved ARG-based model for human actions and group recognition.

We proposed to use MNASNet for extracting image characteristic maps in CNN and to calculate the pair-wise appearance similarity for the Actor Relation Graph

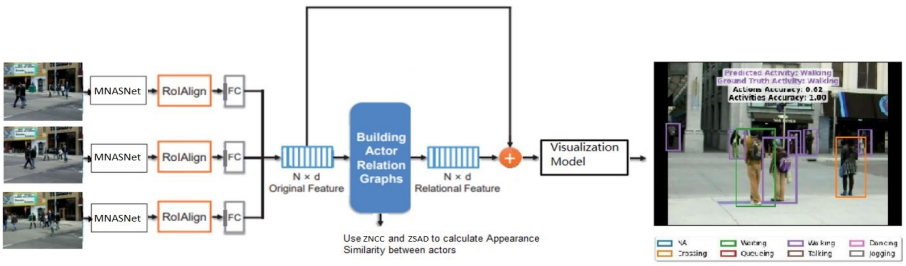

Fig. 2 Optimized ARG-based human actions and group activity recognition

by using Zero mean normalized cross-correlation (ZNCC) and the Zero mean sum of absolute differences (ZSAD). Section V provides further details on our proposed methodology.
In order to give our model a more visualized result we also introduce a visualization model to display every video frame with forecast bounding boxes on each human object. Fig 3 shows the output examples.

4 Methodology

The key to building this Actor relationship graph is in our model. J. Wu et al. has shown that in each frame, the ARG can represent the graphical structure of the information in pairs between the pairs of actors, and use the related information to understand group activity[9]. Both features of appearance and position information are used to construct the ARG to better understand the relation between two actors.The value of the relationship is defined as a composite function f_a below which indicates the relationship of appearance, and f_s indicates position relationship. x_i^a and x_j^a refers to the appearance features of the actor i and actor j while x_i^s and x_j^s refers to the location features (the center of the bounding box of each actor) of the actor i and actor j. Function h combines appearance along with position to scalar weight[9].

$$G_{ij} = h\left(f_a\left(x_i^a, x_j^a\right), f_s\left(x_i^s, x_j^s\right)\right) \tag{1}$$

The normalisation of every actor node with SoftMax function is further adopted in order to ensure that the sum of all values corresponding to each actor node is always one[9]:

$$\mathbf{G}_{ij} = \frac{f_s\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) \exp\left(f_a\left(\mathbf{x}_i^a, \mathbf{x}_j^a\right)\right)}{\sum_{j=1}^{N} f_s\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) \exp\left(f_a\left(\mathbf{x}_i^a, \mathbf{x}_j^a\right)\right)} \tag{2}$$

4.1 Appearance Relation

The Embedded Dot-Product is used in J Wu's paper to calculate the similitude between the appearance features of both actors (the image within each actor's boundary box) in the space where they are embedded[9]: The following is the corresponding function:

$$f_a\left(\mathbf{x}_i^a, \mathbf{x}_j^a\right) = \frac{\theta\left(\mathbf{x}_i^a\right)^{\mathrm{T}} \phi\left(\mathbf{x}_j^a\right)}{\sqrt{d_k}} \tag{3}$$

Both θ and ϕ are functions utilizing $Wx + b$, where W nad b are learnable weights. The learned changes of original features can better comprise the relationship between two actors in a subspatial environment. Two other methods for appearance calculation are: Zero Mean Normalized Cross-Correlation (ZNCC) and Zero mean sum of absolute differences (ZSAD).
A method to evaluation the degree of similarity among two comparable images is the normalized cross correlation(NCC). The lighting and exposure conditions can cause the images to be different in brightness, so that they can first be normalized to have a more accurate similarity score using NCC. The advantage of the normalized cross-relation is that it is less sensitive to linear changes in the illumination amplitude of the two comparative images and can be written in the following way[13]:

$$\varphi'_{x_i^a x_j^a}(t) = \frac{\varphi_{x_i^a x_j^a}(t)}{\sqrt{\varphi_{x_i^a x_i^a}(0)\varphi_{x_j^a x_j^a}(0)}} \tag{4}$$

The quantity $\varphi'_{x_i^a x_j^a}(t)$ vary between -1 to 1. The value of the NCC can help us better understand the relationship of appearance between the two actors. Another method we evaluate in order to calculate the relationship of appearance of ARG is the sum of absolute differences (SAD). By calculating the amount of absolute difference between the matrix components as the formula, the SAD computes the distance between two matrices:

$$SAD\left(x_i^a, x_j^a\right) = \sum_k^n \left|x_{ik}^a - x_{jk}^a\right| \tag{5}$$

SAD is stronger against extreme data values, making it robust when comparing appearance characteristics and better capturing the relationship between appearances.

4.2 Position Relation

Furthermore, spatial structural information is considered to better capture the relationship between actors. To obtain signals from entities not remotely distant, a distancing mask has been applied. As relation is crucial to the understanding of group activities in a local context in comparison to the global relationship, an evaluation of Euclidean distance G_{ij} between two actors is taken as follows:

$$f_s\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) = I\left(d\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) \leq \mu\right) \tag{6}$$

Where $I(\cdot)$is the indicator function and $d\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right)$ evaluates the Euclidean distance between the center coordinates of two actor's bounding boxes, μ is a distance threshold.

5 Results & Discussion

5.1 Datasets

This paper uses a data set known as a collective activity dataset and an augmented dataset[15] to train and test our model. This dataset has a 74 video scene which contains several people in each scene. Every person is also marked with manually defined boundaries and the ground truth of actions and group activity.

5.2 Experminets

The minibatch size is 16, the learning rate is 0.0001 and we train our network through 100 epochs. The single action weight loss $\lambda = 1$ is used. The GCN parameters are determined as $d_k = 256$, $d_s = 32$, and the distance mask threshold μ is adopted as 1/5 of the image width. The default CNN network backbone for function extraction is set to Inception-v3, and an embedded dot-product is the default appearances relating function. The PyTorch framework and individual instance of Tesla K80 GPU are the basis for our implementation.

- *Experiment 1: Evaluation with different backbone networks*
  In this section we conduct extensive studies on the dataset of the collective activities, in order to understand the modeling relationship of the proposed backbone network using group activity prediction precision as the assessment metric. Table I presents the results of the experiments.
  During our two-stage training, the ImageNet pretrained backbone network

Table 1 Group Activity Prediction from Two Backbone Networks

Backbone Network	Stage	Time Cost(seconds)	Best Group Activity Prediction Accuracy
Inception-v3	1	25035	91.03%
Inception-v3	2	25360	92.82%
MNASNet(Proposed)	1	15395	88.63%
MNASNet(Proposed)	2	15395	91.37%

is first finalized with each of the samples randomly selected from the training. Then, in stage 2 the weights of the backbone network feature extraction are fixed. We continue training the GCN network and use the embedded dot-product to calculate the appearance relationship. Inception v3 as a backbone network is the start of our experiments. It takes about 6.95 hours to complete the first training stage, while about 7.04 hours to complete the second stage. Inception v3 achieves the group activity accuracy of 91.03% after stage 1 training. Our model delivers a higher recognition accuracy of 92.82% with additional training of GCN in stage 2.
We also use MNASNet as the backbone network to increase our model speed. MNASNet is a low-weight but efficient deep convolution neural network. The training time for stage 1, which reduces the time used at stage 1 by 38.50%, is

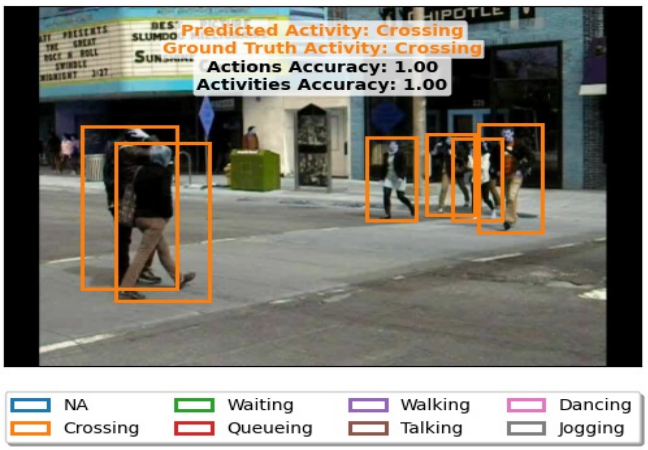

Fig. 3 Optimized ARG-based human actions and group activity recognition

4.27 hours with MNASNet. The training time of stage 2 is 5.37 hours, 23.7% below the training time of Inceptionv3 at the same time. However, the accuracy of activity recognition decreases slightly from 92.82% to 91.37%.

Table 2 Group Activity Prediction on Backbone Network Inception-v3 with Different Appearance Relation Functions

Method	Best Group Activity Prediction Accuracy
Embedded Dot-product	92.06%
ZNCC(our model)	93.58%
ZSAD(our model)	94.37%

– *Experiment 2: Evaluation with different appearance relation functions*
In this experiment, we analyze the performance of the group activity recognition with different functions. The group activity recognition performance is first trained and validated on the default Inceptionv3 backbone and the embedded dot-product for calculating appearance-relation. We have 92.06 % as the best result. The appearance relation function is then updated to the zero mean normalized cross correlation (ZNCC), and the best outcome is 93.58 %. The zero mean sum of the absolute distance (ZSAD) function is further evaluated in order to calculate appearance similarity and 94.37 % is the best result we can achieve. In this Experiment 2, we prove to be more accurate in our proposed model with the ZNCC or ZSAD as an apparence-related calculative function. Table II shows the results of the experiments. Fig 3 shows the output from visualization model.

6 Conclusion & Future Work

This paper uses the model based on the actor relation graph (ARG) for the recognition of group activity. To improve the performance of our model, we learn ARG to use zero mean normalized cross-correlation (ZNCC) and the zero mean sum of absolute difference in the graphs (ZSAD). We also have MNASNet as the backbone network in our proposed model to improve computational speed. In addition, extensive experiments show that the suggested methods for improving precision and speed on the Collective Activity dataset are robust and effective. Because our project focuses primarily on group activity recognition, individual actions are not very accurate because the region of interest (RoI) and CNN are used only to achieve action recognition. We want to further finalize and apply skeleton extraction for the backbone network MNASNet in future to achieve greater prediction precision for individual actions.

Table 3 List of Abbreviations

ARG	Actor Relation Graph
ZNCC	Zero Mean Normalized Cross-Correlation
ZSAD	Zero Mean Sum of Absolute Difference
GCN	Graph Convolution Network

7 Declarations

7.1 Availability of data and materials

For further references see http://vhosts.eecs.umich.edu/vision//activity-dataset.html

7.2 Competing interests

Not Applicable

7.3 Funding

Not Applicable

7.4 Authors' contributions

Category 1
Conception and design of study: Pranjal Kumar
acquisition of data: NILL
analysis and/or interpretation of data: NILL
Category 2
Drafting the manuscript: Pranjal Kumar

revising the manuscript critically for important intellectual content: Siddhartha Chauhan
Category 3
Approval of the version of the manuscript to be published:Siddhartha Chauhan

7.5 Acknowledgements

I would like to express my very great appreciation to Dr Siddhartha Chauhan for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated.

7.6 Authors' information

Pranjal Kumar is pursuing Ph.D., in Department of Computer Science & Engineering, NIT Hamirpur. Phone: +91-8637511985 Email: pranjal@nith.ac.in

Siddhartha Chauhan, joined NIT Hamirpur as Lecturer in CSE Department on 1997. Presently working as Associate Professor in CSE Department. Specializes in Sensor Networks, Ad hoc Network, QoS in Networks, Mobile Computing (Wireless Sensor Networks), Microprocessor Microcontroller Design Phone: 01972-254428 Email: sid@nith.ac.in

References

1. Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.
2. Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.
3. Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.
4. Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.
5. Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.
6. Basura Fernando, Cheston Tan, and Hakan Bilen. Weakly supervised gaussian networks for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 537–546, 2020.

7. Farzan Majeed Noori, Benedikte Wallace, Md Zia Uddin, and Jim Torresen. A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *Scandinavian conference on image analysis*, pages 299–310. Springer, 2019.
8. Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.
9. Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.
10. Xiangpeng Li, Zhilong Zhou, Lijiang Chen, and Lianli Gao. Residual attention-based lstm for video captioning. *World Wide Web*, 22(2):621–636, 2019.
11. Corey DC Heath, Tracey Heath, Troy McDaniel, Hemanth Venkateswara, and Sethuraman Panchanathan. Using participatory design to create a user interface for analyzing pivotal response treatment video probes. In *International Conference on Smart Multimedia*, pages 183–198. Springer, 2019.
12. Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4620–4628, 2019.
13. Y Raghavender Rao, Nikhil Prathapani, and E Nagabhooshanam. Application of normalized cross correlation to image registration. *International Journal of Research in Engineering and Technology*, 3(5):12–16, 2014.
14. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
15. Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009.

Figures

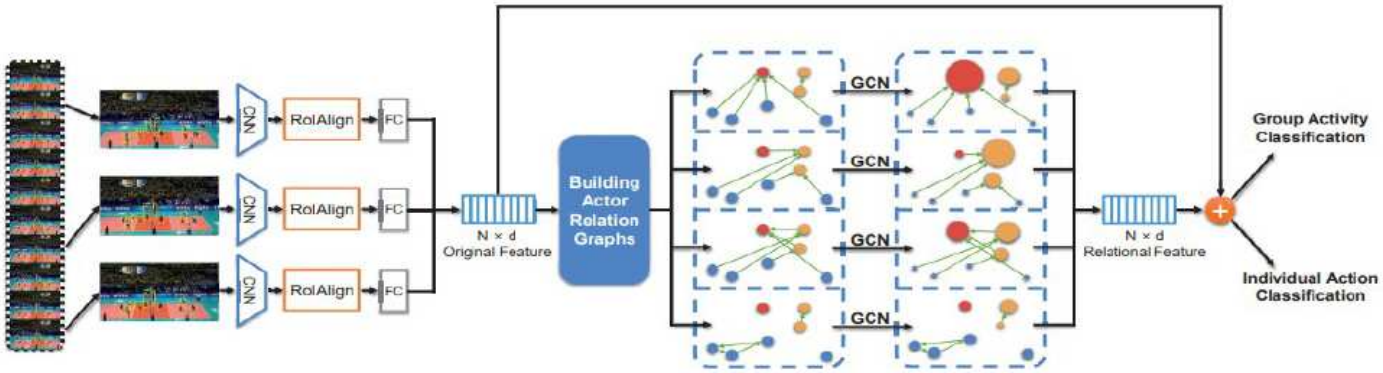

Figure 1

ARG-based group activity recognition framework.

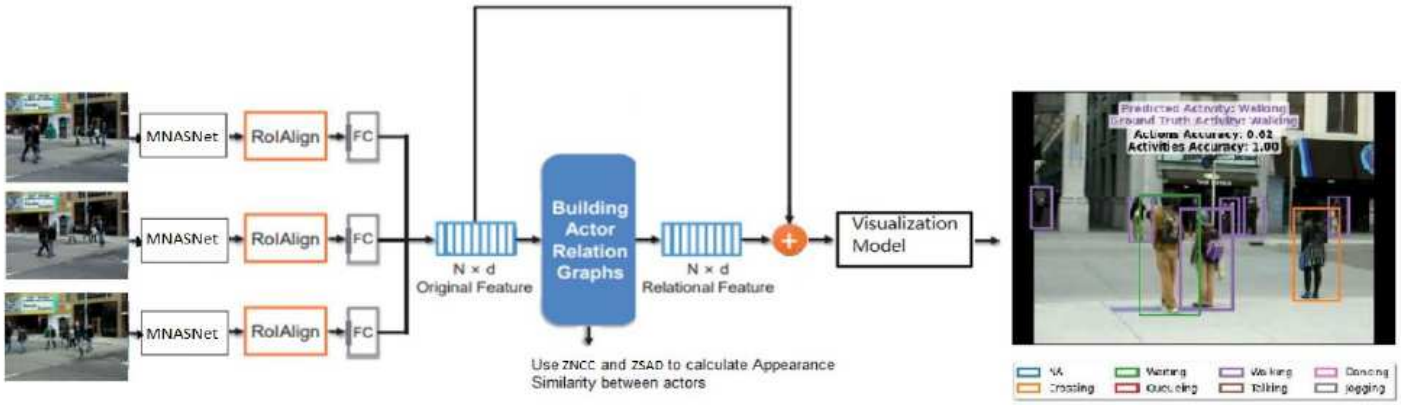

Figure 2

Optimized ARG-based human actions and group activity recognition

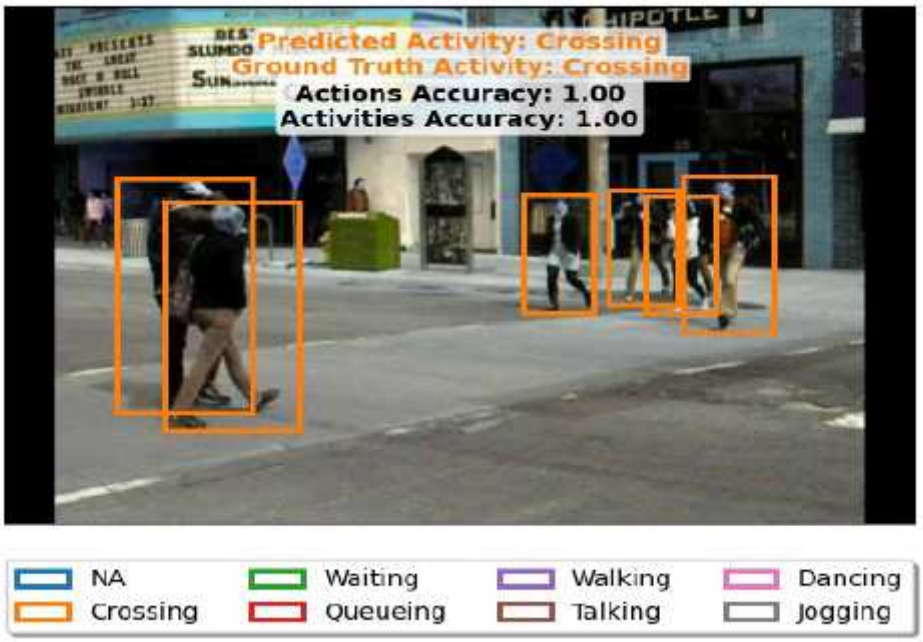

Figure 3

Optimized ARG-based human actions and group activity recognition

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HIghlights.docx