Indian Journal of Science and Technology, Vol 10(23), DOI: 10.17485/ijst/2017/v10i23/113996, June 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

# A Comprehensive Study of Group Activity Recognition Methods in Video

**S. A. Vahora[1]* and N. C. Chauhan[2]**

[1]Department of Computer Engineering, CHARUSAT University, CHARUSAT Campus, Changa – 388421, Gujarat, India; safvan465@gmail.com
[2]Department of Information Technology, A. D. Patel Institute of Technology, New V. V. Nagar – 388121, Gujarat, India; narendracchauhan@gmail.com

## Abstract

**Objectives:** To provide comprehensive review of different group activity recognition methods, categorize them and provide path to new researcher in this domain. **Methods/Statistical Analysis:** Different methods of group activity recognition categorized and analyzed according to hand-crafted and learned feature descriptors. Pros and cons of each method are presented. Methods are analyzed in detailed by finding its local level features to global level feature descriptors used along with performance on benchmark dataset. **Findings:** Different models of group activity recognition are characterized as per the capabilities of the defined model considering individual pose of person, atomic activity of person, person-person interaction, person-group interaction, group-group interaction, uses of temporal information, and recognition of group activity frame wise or video wise. This comprehensive review provides brief information about group activity recognition methods and can be used as brief literature review to the researcher seeking the facts and findings in the field of computer vision in group activity recognition. **Applications/Improvements:** This reviews help in different applications of human activity analysis, mainly in group activity recognition and the models described here can be used in different applications such running or walking on pathways, waiting at public places, queuing in line in group and many more group activity applications for further enhancement.

**Keywords:** Context Model, Convolution Neural Network, Group Activity Recognition, Group Descriptor, Interaction Model

## 1. Introduction

Vision based human activity analysis is one of the most scientific and practical importance, having most challenging problems and has been actively studied in the research field of computer vision. Many previous studies have revealed high attention in recognizing action performed by a single human or complex human activities in video.[1,2] However, in real-world applications, group activity recognition is a challenging and important due to its technical difficulties as well as practical requirements for applications in public places like airports, railway station, sub-ways, bus-stop etc. Group activity analysis characterized by analysis of individual human action within group, analysis of human actions in context of other members within group and analysis of pair-wise person-person interaction with in group. However, which people are in the group and an atomic activity of individual person in each group to be recognized for high accuracy.[3] Figure 1 shows the importance of contextual information as by analyzing contextual information of nearby people. From figure 1 looking at individual person as highlighted without context information both person looks same as standing and looking in same direction but by considering context information we conclude that left person is talking and the right person is queuing.

Group activity recognition based on each group member contribution, where each member have own role that is different than other members in the group. This approach used to recognize single group activity recognition where number of persons in the group are limited and the behavior of each member is not uniform.[2] For

*Author for correspondence

(a)

(b)

**Figure 1.** Importance of contextual information: **(a)** talking video frame **(b)** queuing video frame.

(a)

(b)

**Figure 2.** People involved in different group activities **(a)** Queuing with walking **(b)** talking with walking.

example a presentation session with fixed number of group members. There might be number of group exist in the scene and each group might shows different activity. Group activity recognition is not only having single activity, there might be more than one activity performed where different people involve in different leading activity in the video scene. As shown in Figure 2(a), the leading activity is queuing but some people are walking and in Figure 2(b) the leading activity is talking but some people are walking as highlighted in the figure.

## 2. Review of Methods

Collective activity recognition methods have been applied to different benchmark dataset like collective activity dataset and its extended version, choi's collective activity dataset, UCLA courtyard dataset, volleyball dataset etc.[4-7]

By considering type of feature descriptor used like hand crafted group feature descriptor and leaned feature descriptor, different methods of group activity recognition categorized. Furthermore categorization of main category, into sub-category by considering context model, person-person interaction model and hybrid approach of context and interaction model presented in this paper.

### 2.1 Handcrafted Group Feature Descriptor

In this category group feature descriptors are generated by using low-level feature descriptor such as Histogram of Oriented Gradient (HOG), deformable part model, Spatio-Temporal Local (STL) descriptor, Scale-Invariant Feature

Transform (SIFT), shape context descriptor, Principal Component Analysis (PCA) etc.[4,8-10] Furthermore these methods are categorized as they uses context model like Action Context (AC) descriptor, Relative Action context (RAC) descriptor, Spatio-Temporal Volume (STV).[11,12] Person-person interaction model like distance based attraction and repulsion model, pair-wise interaction model or combine approach.[1,13,14]

### 2.1.1 Group Context Model

In group context model feature descriptor of focal person, and feature descriptor of each person within surrounded region of that focal person are computed. This surrounded region is computed by preprocessing and defining group of person with reference to focal person by applying group detection method or by providing fixed static region surrounded to the focal person. Concentration or pooling operation applied on this feature descriptor of focal person and people with in group to compute context feature descriptor that used for group activity recognition. For N number of person in the group, and $X_1, X_2 \ldots X_n$ are feature of each individual person, including the feature of focal person $X_c$, then group descriptor is describes as $\Delta[X_c, X_1, X_2 \ldots X_n]$, where $\Delta$ is operation performed on feature of each person in group with reference to focal person C.

Context information model is widely used for group activity recognition. In the proposed model which takes action of focal person and action of person within the context region of the focal person, represented as Action Context (AC) descriptor for group activity recognition.[11] In the proposed variation in the action descriptor model by adding pose information.[12] In this model pose information is added along with action of the person. Furthermore instead of fixed pose of person, relative pose of person within the context is used along with action with referenced to focal person, represented as Relative Action Context (RAC) descriptor, which improves accuracy of group activity recognition. In both of this model action recognition is performed by low-level local feature descriptor Histogram of Oriented Gradient (HOG) proposed.[8] An alternative of human action recognition using multi-scale deformable part model proposed, that captures coarser and finer details of person image.[9] In the reported article provides results on benchmark collective activity dataset, also provided results on surveillance videos from nursing home environment.[11,12]

In the proposed a spatio-temporal descriptor to define the spatial distribution of person over time and pose of a person.[4] Histogram based Spatio-Temporal Local (STL) descriptor captures histograms of number of people with pose in different bins surrounded to focal person over time that is used to classify the group activity.

In the proposed a new video descriptor as Bag of the Right Detections (BORDs) to identify the people participating in group activity and remove noisy people from the group.[15] This descriptor fed to generative temporal chain model of group activity that warps the chain of BORDs in time and space. Proposed Maximum a Posteriori (MAP) inference algorithm maps BORDs to their estimated location and maximize the posteriori probability of the chain. Here, BORD descriptor is associated with spatio-temporal feature defined as histogram of person pose in space-time neighborhood centered at a particular point in video.

In the proposed graph based kernel function that captures similarity between graphs, where group of person represented as graph by considering spatial location.[16] Group descriptor accurately encode group appearance include mean and standard deviation of people velocity for each activity and arrangement of group around by quantizing space into four different areas. This descriptor also encodes people orientation as relative angle. Multi-class SVM is used to classify the behavior of the group. Deviation in graph based model proposed as in contrast with the previous model, people spatial orientation is captured using HOG in eight different direction as front, front-left, left, back-left, back, back-right, right and front-right and classified using multi-class Group Lasso.[17] Matrix build using mean, standard deviation, velocity, orientation as well as arrangement of people with in context used to measure similarity and to classify group activity.

### 2.1.2 Interaction Model

In interaction model, features each person is computed within group and compared with each other person's features within group. Compared to context model, interaction model is peer-to-peer feature comparison instead of providing single feature for all people within group. For N number of person in the group, and $X_1, X_2 \ldots X_n$ are feature descriptor of each individual person, then group descriptor is computed as $[X_1\Delta X_2, X_1\Delta X_3, \ldots, X_1\Delta X_n, \ldots, X_{n-1}\Delta X_n]$, where $\Delta$ is operation like correlation coefficient, tensor product, cross product etc. This generated feature descriptor to be used for group activity recognition.

In the proposed hierarchical interaction model that captures individual person action along with pose, person-person interaction and group-person interaction.[14] Individual person action potential is computed using action feature HOG, action label used to capture person-person interaction that jointly captures group activity.[8] Action potential uses action-pose information of individual person. Adaptive structure proposed here to define person-person interaction that decides interaction between two people to be considered or not and the result of this adaptive structure compared with no connection, minimum spanning tree and $\in$-neighborhood graph with different values of $\in$ (100, 200 and 300).

Graph based interacting group discovery algorithm proposed.[1] Here, Bag-of-word approach with motion feature used to represent group activity. Dominant set based clustering algorithm named Social Force Model (SFM) and Visual Focus of Attention model (VFoA) required distance and angle between individual people used to define interacting group from video sequence. For group activity recognition Local Group Activity (LGA) descriptor proposed that takes interacting group as input and encode by concatenating motion and pose information of person in the group along with person-person interaction weight. Magnitude of motion information of all people used in LGA descriptor as motion information.

Temporal interaction matrix of group motion pattern based group activity recognition proposed.[18] In this model, 4D interaction tensor generated for person p1 and p2 between time interval t1 and t2, and characterized in riemannian geometry using Discriminative Temporal Interaction Manifold (DTIM). This approach does not require any domain specific knowledge. For each class a multi-model density function on the DTIM leaned and MAP classifier is designed on DTIM for group activity recognition. Different result using two local feature HOG and Histogram of Optical Flow (HOF) and three different measure methods between histogram are $\chi 2$ distance, Euclidian distance and cosine distance used in probabilistic DTIM model. Furthermore, this model also applied for group activity recognition in form of point trajectories on Gatech Football Play Dataset.[19]

In the proposed distance based Group Interaction Zone (GIZ) to detect and update interactive groups in a scene that removes persons from the scene those were not participating in a group activity.[13] A novel Attraction Repulsion Feature (ARF) based on the relative distance over a specific time period is used to describe the group activity within GIZ. Along with these features additional features like mean and variance of the magnitudes and orientation of velocity used for group activity recognition. Group activity recognition performed on BEHAVE dataset having different ten group activities performed by two to five people.[20]

In the proposed a model to recognize mixed group activity using only spatial information of video frame and without considering the spatio-temporal information.[21] By assuming taking place of scene with activities, four level model consist of visible pose, standard pose, mixed group activity and scene. In this model, inter-level person interaction, intra-level person interaction, intra-level interaction between groups and intra-level interaction within the group is considered. By applying the top-down approach for a given scene category group activity are trained, and for each group activity standard pose of the person within the group trained that is derived from the trained visible pose of the person. By considering this, to label unknown image maximum likelihood estimation model applied. In this model Deformable Parts Model (DPM) used to person, and action descriptor is used to represents standard pose and visible pose.[9]

### *2.1.3 Combined Approach*

In this approach, group descriptor is computed using context model as well as interaction model, and combination of these used for group activity recognition. In this approach descriptor of both model used to generate new group descriptor or the output of both model approach as context model and interaction model used to recognize group activity by using combining multiple classifier approach.

In the proposed context model along with multi-scale relationship feature of person such as size (similar size or different size), position (far or near), movement (stay or move) and time sequence (far time or near time) in a single model.[22] In this model, human relationship is not constant, but described as a variable potential using unary potential and pair-wise potential. Unary potential computed, individual for each person as described using Action Context (AC) descriptor and pair-wise potential computed by encoding relationship between each person using multi-scale relationship features as described.[11] Fully Connected – Conditional Random Field (FC – CRF) model is used for group activity recognition by using these unary and pair-wise potential that is represented as Action + Context Fully Connected Conditional Random

Field (AC + FC CRF) model and compared with adjacency connected CRF. In the article includes pose information using Relative Action Context (RAC) descriptor along with action information, proposed Action Context – Relative Action Context Fully Connected CRF (AC-RAC + FC –CRF) and compared with different types of connection as connected per frame, adjacency connected and simple fully connected model.[3] In these model, motion is computed using mean optical flow by using approach presented.[23]

In the proposed a framework for tracking of multiple people, individual person action, context information, person-person interaction and group activity recognition.[5] For each individual person action two visual features computed are HOG of each person image and Bag of Video words (BOV) by computing histogram of video words within the spatio-temporal volume. Video words obtained by applying Principal Component Analysis (PCA) and k-means algorithm on the spatio-temporal volume and group activity descriptor computed using Spatio-Temporal Local (STL) descriptor. In this model, overall energy function divided into seven different local energy functions are $\Psi(C,I)$, $\Psi(C,O)$, $\Psi(I,A,T)$ , $\Psi(A,O)$, $\Psi(C)$, $\Psi(I)$ and $\Psi(A)$ defined as collective interaction, collective activity observation, correlation between individual activity and interaction, individual atomic activity observation, temporal relationship of group activity, interaction and atomic activity (with pose) across the adjacent frames respectively. In the proposed hybrid approach that combines bottom-up detection information to top-down proof.[24] Bottom-up approach automatically infer group activity label and top-down approach provides contextual information of group of individuals in spatio-temporal domain. In contrast with method proposed instead of using STL descriptor, Randomized Spatio-Temporal Volume (RSTV) is used that partition binning space around a person to maximize discrimination power.[5] Here, the feature is defined as number of people lying in spatio-temporal volume specified by pose, location, velocity and time.

In the proposed hierarchical hybrid model of feature level and structure level approaches that is extension of model proposed.[14] In structure level approach person-person interaction derived from individual person action, which is defined as adaptive structure model and in feature level approach Action Context (AC) descriptor is used.[25]

To model individual person action, participating objects and group activity jointly[6] proposed three layer AND-OR graph representation and this hierarchical structure express new inference algorithm. This model includes both context and interaction information as inference algorithm calculates three different processes are $\alpha$, $\beta$ and $\gamma$ process represented as direct inference based activity detection, bottom-up inference based on parts of activity detection and top-down inference on context of activity detection respectively. For person detector and individual person pose detection DPM and HOG are used with SVM classifiers respectively. For action recognition motion based feature as first spatiotemporal interest point (STIP) motion feature described using HOG feature.[26] For tracking Harris corners KLT tracks used and finally combination of Kanade–Lucas–Tomasi (KLT) feature tracks with STIP feature used to capture spatiotemporal relationship to capture human action. Space-time-volume descriptor of every people that encodes people count, location, pose and velocity in surrounded bins used for group activity recognition per frame. Cost-sensitive optimal inference sequence of Spatio-Temporal AND-OR Graph (ST-AOG) learned using Monte Carlo Tree Search (MCTS) proposed.[27] ST-AOG is expansion of temporal AND-OR graph (AOG), and non-temporal AOG.[6,28,29] MCTS estimates inference step by using an experimental average over training data. Here, group activities represented as highest level of ST-AOG, where at the bottom level individual human action defined, who may interact with different parts also known as children node of the group activity. Children nodes shared by multiple parents, AND nodes represents configuration of humans/parts and OR node represents alternative configuration. ST-AOG captures temporal edge link of group activity to model. In contrast to proposed approach in the article[6] inference algorithm identifies ST-AOG node and calculates four different process on that node as $\alpha$, $\beta$, $\gamma$ and $\omega$ where $\omega$ represents tracking of activity in time interval. For detection, tracking and group activity recognition descriptor proposed.[6]

In the proposed group contextual interaction model for group activity recognition using individual context and pair-wise interaction factor.[30] In this model, dominant set based clustering algorithm is used to define interacting group from video sequence. This model consists of two factors, singleton factor and pair-wise factor. Singleton factor defined as Group Context Activity (GCA) descriptor formed using Bag-of-Words representation of

people in the group. It encodes individual person behavior and person within context region of that focal person by taking motion boundary histogram as local descriptor. Pair-wise factor is computed pair-wise activity label of each person with another with in context region. The inference is performed by MAP inference on network.

To capture human motion interaction in group video, poselet Activation Pattern Over Time (TPOS) descriptor proposed.[31] Poselet is defined by dividing an image frame into a set of $N_h$ x $N_w$ grid cells and run poselete detector as a filter on each cell that provides detected poselet with their bounding box size over a time. This N temporal poselete create a codebook using clustering method like k-means clustering algorithm using cosine distance. Thus inference is performed by using bag-of-word for each video by assigning nearest cluster using cosine distance.

In the proposed hierarchical model for group localization and group activity recognition by taking advantage of relationship among participants.[32] By considering multiple group, those may exist in one video, each group is represented as tree structure and group-group relationship is also represented as fully connected graph. For each graph structure (h) with best group location (g), its activity label (a) and input to this each person (x) solved by providing training to combination of these three parameters (h,g,a) and optimizing these parameters for better result of group localization. Optimization of graph structure (h) performed by fixing group (g) and activity label (a) using spanning tree algorithm, optimization of activity (a) performed by fixing group (g) and graph structure (h) by all possible activities, and optimization of group (g) performed by fixing graph structure (h) and activity (a) by splitting and merging of tree. In this model, two types of potentials are computed as intra-group potential that encodes relationship among different person within that group and inter-group potential that encodes relationship between different groups in a video. For intra group potential pairwise person-person descriptor is computed using person descriptor that is defined, their element wise subtraction and element wise multiplication with bag-of-word representation, and for inter group potential group-group descriptor is computed by taking concatenation of pairwise potential along with relative deviations in the x and y direction over a time period. As a person descriptor HOG and Support Vector Machine (SVM) classifier is used in this model.

In the proposed structure prediction function with the help of Boosted Hidden Conditional Random Fields (HCRFs) for group activity.[33] This function learn over the inputs, outputs and the discrete variables between inputs and outputs also known as latent variables. The Action Context (AC) descriptor of each individual person provided as input to the potential function, processed as combination of multiple non-linear functions step by step that generated by using gradient ascent. Train this model using HCRF-Boost algorithm and continue to update potential functions for number of iteration as it converged or maximum number of iteration reached. In the reported article it reaches to measurable high accuracy compared to the state-of-the-art methods, by providing multi-instance kernel based on the cardinality relationship.[34] Each group activity classified by calculating actions of individual person in that group that reduces the change in clutter and noise effect to group activity recognition. Multi-instance learning uses positive and negative instance by representing positive bag that contains at least one positive instance and negative bag contains all negative instances. Kernel over the bags (video) defined to compute the cardinality similarity between the videos by measuring likely same number of counts of frame from the videos of same event of interest. This kernel fed to SVM for classification. The key point here is to define parameter of cardinality model that controls model. Cardinality kernel model also proposed for event detection on TRECVID MED11 dataset and provides promising results.

## 2.2 Learned Feature Descriptor

As the methods described above use hand crafted feature descriptor such as HOG, HOF, BoV, MBH, STIP, AC, RAC etc. In contrasts with hand crafted feature descriptor, learned features are obtained by providing training. Learned features gains high attention and providing measurable results in visual recognition and description applications like activity recognition, video summarization, object categorization from image or video etc. In the proposed recurrent convolutional architecture for large scale human activity recognition, video description and image captioning.[35] Long Term Recurrent Convolutional Network (LRCN) model is used for human activity recognition, where visual input as image is given to deep Convolutional Neural Network (CNN) that extract fixed length visual feature vector, and this visual feature vector fed to stack of recurrent sequence learning models known as Long Short-Term Memory (LSTM) as input that predict

final variable length output. Human activity recognition is performed on UCF101 videos using caffenet reference model proposed also known as caffe framework and uses pre-trained network model that used.[36-38]

In the proposed hierarchical learning approach based on person-person interaction model instead of hand-crafted feature descriptor model for group activity recognition without unified framework to learn human action, pose estimation, tracking etc.[39] In this model class specific interaction matrix is designed to measure person-person interaction in that group activity class and proposed model is computed by accumulating all these class specific interaction from a video sequence, known as Interaction Response (IR) model. IR matrix computed for each group activity class separately. In this model, person-person interaction matrix defined by inner product of two atomic activity and sum of such person-person interactions from video sequence is represented as global collective activity. Compared to majority of the state-of-the-art method, this method recognizing the video sequences as group activity instead of person wise or frame wise group activity recognition.

Group activity recognition using deep neural network proposed.[40] For a given video sequences, person detection applied and those bounding box of each person fed to deep neural network. Convolutional Neural Network (CNN) used to predict the score for action of individual person, pose of person and group activity. This score fed to the model that learns using multi step message passing parameters and infer using back-propagation neural network. In addition with this learned feature, Action Context descriptor (AC) is used as a global feature and average pooling function is used to combine AC descriptor of all persons of the group. The caffe library used for CNN implementation and pre-trained AlexNet network using the ImageNet data proposed by used in this model.[37,41]

**Table 1.** Comparative analysis of characteristic and capabilities of methods

| Models | Pose | Atomic Activity | Person-Person Interaction | Person-group Interaction | Group-Group Interaction | Temporal Information | Frame wise | Video wise | Hand Crafted / Learned Descriptor |
|---|---|---|---|---|---|---|---|---|---|
| 1, 30 | X | | X | | | X | X | | H |
| 15 | X | | | | | X | | X | H |
| 6, 27 | X | X | X | X | | X | X | X | H |
| 22 | | X | X | X | | X | X | | H |
| 5, 24 | X | X | X | X | | X | X | | H |
| 16 | | | | X | | X | X | | H |
| 17 | X | | | X | | X | X | | H |
| 18 | | | X | X | | X | | X | H |
| 4 | X | | | | | X | X | | H |
| 11 | | X | | | | X | X | | H |
| 12 | X | X | | | | X | X | | H |
| 14, 25 | X | | X | X | | | X | | H |
| 21 | X | X | X | | X | | X | | H |
| 32 | | X | X | | X | X | X | | H |
| 33 | | X | X | X | | | X | | H |
| 39 | X | | X | | | X | | X | L |
| 31 | X | | | X | | X | | X | H |
| 34 | | X | X | X | | | | X | H |
| 7 | | X | | X | X | X | X | | L |
| 40 | X | X | | X | | | X | | L |
| 42 | | X | X | X | | X | X | | L |

Features extracted from proposed model trained using RBF kernel SVM to predict the group activity. A unified framework for group activity recognition to combine graphical model and deep neural network model proposed.[42] In this model, with the help of caffe library each person's action score for a given video frame is computed using pre-trained AlexNet Convolutional Neural Network (CNN). This score is refines using number of iteration of message passing fed to Recurrent Neural Network (RNN) that iteratively applies pooling function to person's action score derived from CNN. For structure level learning, trainable gating function used to remove outliers from the video frame by turn on, turn off gating function between individual people from video frame and determines which persons are connected. Finally, each person in the group, which is connected, fed to prediction layer to predict each person's action. By observing individual person's action and their relationship with the others in the group, proposed deep model that captures temporal changes in individual human action and temporal changes in group activity.[7] In this multi-layer temporal deep model, each person image fed to deep CNN, pre-trained AlexNet. For person level temporal change collection, first layer of Long-Short Term Memory (LSTM) used, that take fc7 feature from AlexNet network of each person image, to describe person's action and temporal change in the person's action. By applying pooling function to each person's output of first layer of LSTM to aggregation function, resultant output fed to second level LSTM that captures group level temporal changes and softmax classification layer used to recognize

**Table 2.** Comparisons of methods with accuracy

| Methods | Collective Activity Dataset | Collective Activity Dataset Extended | Choi's Collective Activity Dataset | UCLA Courtyard Dataset |
|---|---|---|---|---|
| 30 | 82.90 % | | | 71.40% |
| 1 | 78.75 % | 80.77 % | | |
| 15 | | 81.50 % | | |
| 27 | 88.90 % | 84.20 % | | 79.30% |
| 22 | 72.20 % | | | |
| 3 | 74.70 % | | 70.70% | |
| 16 | 73.00 % | | | |
| 17 | 81.00 % | | | |
| 18 | 74.00 % | 85.00 % | | |
| 24 | 79.60 % | | 79.20 % | |
| 5 | 79.60 % | | 79.20 % | |
| 4 | 65.90 % | | | |
| 11 | 68.20 % | | | |
| 12 | 73.20 % | | | |
| 14 | 77.50 % | | | |
| 25 | 78.40 % | | | |
| 6 | 83.60 % | | | |
| 21 | 82.07 % | | | |
| 32 | 83.07 % | | | |
| 33 | 82.50 % | | | |
| 39 | 83.30 % | | 80.30 % | |
| 31 | 72.30 % | | | |
| 34 | 83.40 % | | | |
| 7 | 81.50 % | | | |
| 40 | 80.60 % | | | |
| 42 | 81.20 % | | | |

group activity. This model captures temporal change at person level and group level using LSTM that results in measurable accuracy compared to the state-of-the-art methods using hand-crafted feature descriptor as well as leaned feature descriptor.

## 3. Comparative Analysis of Methods

By analyzing each model presented in section 2 of this paper, based on the different characterizes and capabilities of each model to recognize group activity, Table 1 provides detail information with different perception like use of individual person pose, atomic activity of individual person, pair-wise person-person interaction, person feature contribution to direct group activity recognition mapping, interaction between multiple groups, temporal information used in the model, frame wise or video wise group activity recognition and hand-crafted or leaned feature descriptor used by the model.

## 4. Dataset Description and Results

To comprehensive analysis of group activity recognition methods, we analyzed method's performance on different benchmark dataset as collective activity dataset, that consist of 40 short video clips sequence of five group activities are walking, talking, queuing, crossing and waiting.[4] Choi's collective activity dataset consists of 32 video clip sequences of six group activities are walking together, dismissal, chasing, queuing, talking and gathering.[5] UCLA courtyard dataset consists of 106 minute video footage at 30fps consist of different group activities are standing-in-line, waiting, sitting-together, walking, discussion-in-group and guided-tour.[6] This dataset also includes individual human action and different objects in annotated form. Table 2 provides comparisons of all methods on different benchmark dataset.

## 5. Conclusion

We reviewed group activity recognition methods along with the group descriptor used by each method. The state-of-the-art methods widely used in computer vision, machine learning, video surveillance, human detections and tracking, action recognitions, etc. have been considered in this work. Our review is based on different methods by looking at hand-crafted and learned feature descriptors with different characteristics and capabilities of each method and experimental performance on benchmark dataset. The methods are analyzed in the direction from local level to global level feature descriptors. It has been observed that spatio-temporal feature descriptor provides measurable results. Furthermore, hand-crafted and learned feature both have their own advantages and gives reckonable results if they are chosen appropriately. The presented review can be helpful to researchers and developers to get the present status of work and further explore and contribute.

## 6. References


1. Tran KN, Gala A, Kakadiaris IA, Shah SK. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. Pattern Recognition Letters. 2014;44:49-57. Crossref
2. Aggarwal JK, Ryoo MS. Human activity analysis. ACM Computing Surveys. 2011;43(3):1-43. Crossref
3. Kaneko T, Shimosaka M, Odashima S, Fukui R, Sato T. A fully connected model for consistent collective activity recognition in videos. Pattern Recognition Letters. 2014;43:109-18. Crossref
4. Wongun C, Shahid K, Savarese S. What are they doing? : Collective activity classification using spatio-temporal relationship among people. International Conference on Computer Vision Workshops, ICCV Workshops; 2009/09: IEEE; 2009. Crossref
5. Choi W, Savarese S. A Unified Framework for Multi-target Tracking and Collective Activity Recognition. Computer Vision ECCV 2012: Springer Berlin Heidelberg; 2012;215-30. Crossref
6. Amer MR, Xie D, Zhao M, Todorovic S, Zhu S-C. Cost-Sensitive Top-Down/Bottom-Up Inference for Multiscale Activity Recognition. Computer Vision ECCV 2012: Springer Berlin Heidelberg; 2012;187-200. Crossref
7. Ibrahim MS, Muralidharan S, Deng Z, Vahdat A, Mori G. A Hierarchical Deep Temporal Model for Group Activity Recognition. Conference on Computer Vision and Pattern Recognition; 2016/06: IEEE; 2016. Crossref
8. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition: IEEE; 2005. Crossref
9. Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Conference on Computer Vision and Pattern Recognition; 2008/06: IEEE; 2008. Crossref

10. Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002;24(4):509-22. Crossref
11. Lan T, Wang Y, Mori G, Robinovitch SN. Retrieving Actions in Group Contexts. Trends and Topics in Computer Vision: Springer Berlin Heidelberg; 2012;181-94. Crossref
12. Kaneko T, Shimosaka M, Odashima S, Fukui R, Sato T. Viewpoint Invariant Collective Activity Recognition with Relative Action Context. Computer Vision ECCV 2012 Workshops and Demonstrations: Springer Berlin Heidelberg; 2012;253-62. Crossref
13. Kim Y-J, Cho N-G, Lee S-W. Group Activity Recognition with Group Interaction Zone. 2014 22nd International Conference on Pattern Recognition; 2014/08: IEEE; 2014. Crossref
14. Lan T, Wang Y, Yang W, Mori G, editors. Beyond actions: Discriminative models for contextual group activities. Advances in neural information processing systems; 2010.
15. Amer MR, Todorovic S. A chains model for localizing participants of group activities in videos. International Conference on Computer Vision; 2011/11: IEEE; 2011. Crossref
16. Noceti N, Odone F. A Spectral Graph Kernel and Its Application to Collective Activities Classification. 22nd International Conference on Pattern Recognition; 2014/08: IEEE; 2014. Crossref
17. Noceti N, Odone F. Humans in groups: The importance of contextual information for understanding collective activities. Pattern Recognition. 2014;47(11):3535-51. Crossref
18. Li R, Chellappa R, Zhou SK. Recognizing Interactive Group Activities Using Temporal Interaction Matrices and Their Riemannian Statistics. International Journal of Computer Vision. 2012;101(2):305-28. Crossref
19. Kihwan K, Dongryeol L, Essa I. Detecting regions of interest in dynamic scenes with camera motions. IEEE Conference on Computer Vision and Pattern Recognition; 2012/06: IEEE; 2012. Crossref
20. Blunsden S, Fisher R. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA. 2010;4(1-12):4.
21. Zhou Z, Li K, He X, Li M, editors. A Generative Model for Recognizing Mixed Group Activities in Still Images. 25th International Joint Conference on Artificial Intelligence; 2016: AAAI Press.
22. Kaneko T, Shimosaka M, Odashima S, Fukui R, Sato T, editors. Consistent collective activity recognition with fully connected CRFs. 21st International Conference on Pattern Recognition; 2012: IEEE.
23. Sun D, Roth S, Black MJ. Secrets of optical flow estimation and their principles. IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010/06: IEEE; 2010. Crossref
24. Choi W, Savarese S. Understanding Collective Activitiesof People from Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(6):1242-57. Crossref
25. Tian L, Yang W, Weilong Y, Robinovitch SN, Mori G. Discriminative Latent Models for Recognizing Contextual Group Activities. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012;34(8):1549-62. Crossref
26. Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. IEEE Conference on Computer Vision and Pattern Recognition; 2008/06: IEEE; 2008. Crossref
27. Amer MR, Todorovic S, Fern A, Zhu S-C. Monte Carlo Tree Search for Scheduling Activity Recognition. IEEE International Conference on Computer Vision; 2013/12: IEEE; 2013. Crossref
28. Si Z, Pei M, Yao B, Zhu S-C. Unsupervised learning of event AND-OR grammar and semantics from video. International Conference on Computer Vision; 2011/11: IEEE; 2011. Crossref
29. Pei M, Yunde J, Zhu S-C. Parsing video events with goal inference and intent prediction. International Conference on Computer Vision; 2011/11: IEEE; 2011. Crossref
30. Tran KN, Yan X, Kakadiaris IA, Shah SK. A Group Contextual Model for Activity Recognition in Crowded Scenes. Proceedings of the 10th International Conference on Computer Vision Theory and Applications: SCITEPRESS - Science and and Technology Publications; 2015. Crossref
31. Nabi M, Del Bue A, Murino V. Temporal Poselets for Collective Activity Detection and Recognition. IEEE International Conference on Computer Vision Workshops; 2013/12: IEEE; 2013. Crossref
32. Sun L, Ai H, Lao S. Activity Group Localization by Modeling the Relations among Participants. Computer Vision ECCV 2014: Springer International Publishing; 2014;741-55. Crossref
33. Hajimirsadeghi H, Mori G. Learning Ensembles of Potential Functions for Structured Prediction with Latent Variables. IEEE International Conference on Computer Vision; 2015/12: IEEE; 2015. Crossref
34. Hajimirsadeghi H, Wang Y, Vahdat A, Mori G. Visual recognition by counting instances: A multi-instance cardinality potential kernel. IEEE Conference on Computer Vision and Pattern Recognition; 2015/06: IEEE; 2015. Crossref
35. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Conference on Computer Vision and Pattern Recognition; 2015/06: IEEE; 2015. Crossref
36. Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402. 2012.

37. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe. Proceedings of the ACM International Conference on Multimedia - MM '14: ACM Press; 2014. Crossref
38. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer Vision ECCV 2014: Springer International Publishing; 2014;818-33. Crossref
39. Xiaobin C, Wei-Shi Z, Jianguo Z. Learning Person-Person Interaction in Collective Activity Recognition. IEEE Transactions on Image Processing. 2015;24(6):1905-18. Crossref
40. Deng Z, Zhai M, Chen L, Liu Y, Muralidharan S, Roshtkhari MJ, et al. Deep Structured Models For Group Activity Recognition. Procedings of the British Machine Vision Conference: British Machine Vision Association; 2015. Crossref
41. Krizhevsky A, Sutskever I, Hinton GE, editors. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.
42. Deng Z, Vahdat A, Hu H, Mori G. Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition. IEEE Conference on Computer Vision and Pattern Recognition; 2016/06: IEEE; 2016. Crossref