A survey on human activity recognition from videos

Conference Paper · February 2016

DOI: 10.1109/ICICES.2016.7518920

CITATIONS
59

READS
4,798

2 authors:

Subetha Thankaraj
Bvrit Hyderabad college for women
14 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Chitrakala Gopalan
Anna University, Chennai
122 PUBLICATIONS **416** CITATIONS
SEE PROFILE

A Survey on Human Activity Recognition from Videos

T.Subetha
Research Scholar, Department of CSE
Anna University, CEG, Chennai
subethathankaraj@gmail.com

Dr.S.Chitrakala
Associate Professor, Department of CSE
Anna University, CEG, Chennai
au.chitras@gmail.com

***Abstract*— Understanding the activities of human from videos is demanding task in Computer Vision. Identifying the actions being accomplished by the human in the video sequence automatically and tagging their actions is the prime functionality of intelligent video systems. The goal of activity recognition is to identify the actions and objectives of one or more objects from a series of examination on the action of object and their environmental condition. The major applications of Human Activity Recognition varies from Content-based Video Analytics, Robotics, Human-Computer Interaction, Human fall detection, Ambient Intelligence, Visual Surveillance, Video Indexing etc...This paper collectively summarizes and deciphers the various methodologies, challenges and issues of Human Activity Recognition systems. Variants of Human Activity Recognition systems such as Human Object Interactions and Human-Human Interactions are also explored. Various benchmarking datasets and their properties are being explored. The Experimental Evaluation of various papers are analyzed efficiently with the various performance metrics like Precision, Recall, and Accuracy.**

***Index terms*— Computer Vision; Human Activity Recognition; Human Object Interaction; Human Human Interaction:**

I. INTRODUCTION

Human Activity Recognition system gains its popularity due to the increase in number of surveillance cameras. The goal of activity recognition is to identify the actions and objectives of one or more objects from a series of examination on the action of object and their environmental condition. The major applications of this system are not limited to choreography, surveillance security, sports, and context-based retrieval. A single activity corresponds to many elementary actions.

Human Activity Recognition systems conventionally follows a hierarchical manner. Background subtraction, feature extraction, tracking and detection comes under the lower level. The action recognition module falls under mid-level approach followed by the reasoning engines on the high level that encode the context of the actions based on the units of lower level. The general framework of Human Activity Recognition system is depicted in figure 1.

In the lower-level, background subtraction is performed on the extracted frames either by pixel-based, block-based or the combination of both. Gaussian Modelling, Mixture of Guassians, ViBe, Kalman Filter, Hidden Markov Model are commonly used pixel-based background models. Block-based approach usually falls under Normalized Vector Distance, Histogram Similarity, Incremental PCA and Local Binary Pattern Histogram. After the detection of foreground, the feature extraction is performed by either of the two approaches. In Model based technique a human model is built for recognition of activity whereas in feature based technique either the local or global features or both features are extracted that aid in the numerical computation for activity detection.

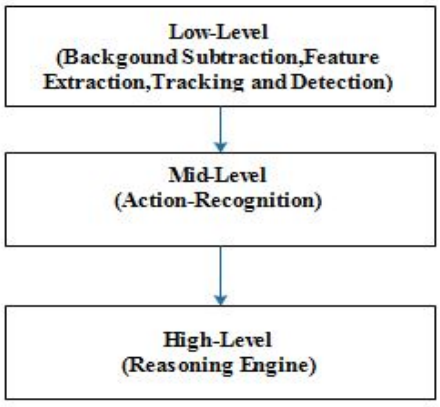

Fig. 1. General Framework of Human Activity Recognition

During mid level after the detection and tracking process, it is given to the classifier for the recognition of actions. After the identification of actions it is given to a high level that includes a reasoning engine that interprets the activity of humans. This paper focuses on providing an extensive survey on the high level approach that discusses how the activity of humans are interpreted from the assorted works of various authors .The challenges and their applications are also explored.

This paper is structured as follows. Section 2 depicts about the various challenges in Human Activity Recognition. Section 3 depicts about various Human Activity Recognition techniques. Section 4 depicts about Human-Object Interactions. Section 5 describes about the various existing interactions. Section 6 shows the datasets used and Section 7 describes about the experimental evaluation .Section 8

978-1-5090-2552-7

discusses about the applications and section 9 about the inferences made and future research direction followed by conclusion in Section 10.

REFERENCES	METHODOLOGY	ADVANTAGES	FUTURE WORK
[1]	multiple instance-SVM	Improves the local feature based activity recognition by using advanced machine learning techniques which are different from bag-of-features based representation	incorporates spatio-temporal informations and different descriptors
[2]	subspace clustering approach	can handle multi-dimensional data that are not possible with the typical clustering method	aims in fusing a large contextual information such as emotions, health conditions.
[3]	non-parametric comparison of trajectory data with the fusion of bayes net	ability to detect activity even in medium resolution videos	The commentary can be extended to security applications
[4]	Kinematics Model-Based	Takes only the prominent human poses that dispatches the information and eliminates the rest	handling of similar action recognition
[5]	Kinematics Model-Based	Depends on Contour points for learning keyposes.	This method shows high tolerance to inter actor variance handling of occlusion and view-invariance
[6]	Physics Model-Based	Dynamic features are computed and by using these features action classes are classified in terms torques	apply dynamic features to human-gait recognition
[7]	Kinematics Model-Based	Adaptive vision-based human action recognition method is proposed.	adaptive learning should be compared to other benchmarking incremental learning and continuous adaptation methods.
[8]	Kinematics Model-Based	A new skeletal representation that specifically models the 3D geometric relationships between various body parts using rotations and translations in 3D space	increment the system to model complex activities

TABLE 1: COMPARISON TABLE OF VARIOUS METHODOLOGIES OF ACTIVITY RECOGNITION ALGORITHMS

II. CHALLENGES IN HUMAN ACTIVITY RECOGNITION

The major applications of Human Activity Recognition system is discussed below.

A. Application Domain

The activities of interest and the importance of fine details will vary based on the application domain. For example, for a surveillance system, the main interest is typically in finding the unusual behavior (e.g. falling down, jumping over a fence, etc.).

B. Variations in Inter and Intra class

The performance of the system depends on the large variations in activity class. For example the activity walking and jogging will vary by only a small degree. A good human activity recognition should be able to differentiate the activities of one class with another class.

C. Learning Paradigm Usage

A learning based approach is used to recognize different human activities. The main advantage is robustness to intra class variations. The learning paradigm usage can be either supervised or unsupervised based on the type of training data available.

D. Occlusion

Occluded parts either self or due to objects have a major effect in recognition of actions. As the features from the Occluded body parts are lost often, the system can end up in recognizing an action wrong and yet the system would be correct as the features from the occluded parts have a little effect in the output.

E. Background and Recording settings

Identifying the activities of human with a cluttered or dynamic background is difficult. The quality of a video is also a prime factor in deciding the performance of the system. An efficient activity recognition system should recognize the human even in the varying quality of the video and the cluttered background.

III. ACTIVITY RECOGNITION

Human Activity is grouping of human/object movements with its semantics. Activity Recognition plays its role in finding the video segments that contains such movements .This section discusses about the various techniques adapted for an efficient feature activity recognition and the overview is shown in Figure 2 .A comparison table is prepared for various methodologies and shown in Table I.

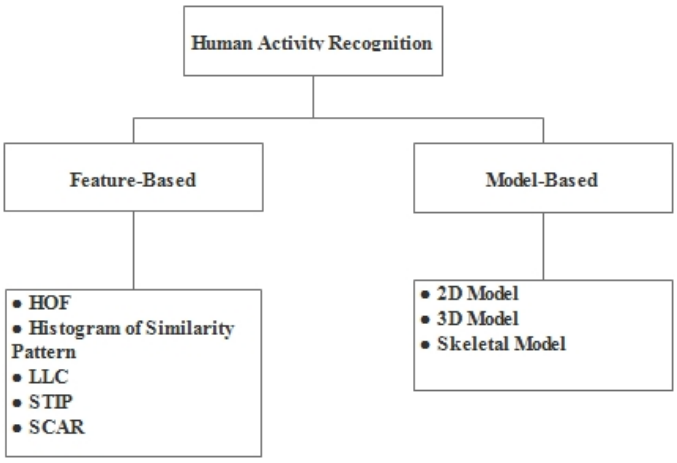

Fig. 2. Overview of Human Activity Recognition

A. Feature-based approaches

Sabanadesan et al [1] improves local feature based activity representation. The dense HOG (Histogram of oriented Gradients) and histogram of similarity patterns are extracted from videos. Then they utilized multiple instance- Support Vector Machine (mi-SVM) to build a separate a codebook for every activity class instead of a single class. The Locality constrained Linear Coding (LLC) is then used to exemplify each input feature with collective elements of codebook followed by pooling of spatio-temporal features. After building the dictionary, SVM is applied for classification. As the conventional clustering process couldn't handle the multi-dimensional data, *zhang et al* [2] improves the activity recognition by applying subspace clustering. SCAR (subspace clustering based approach) utilize a typical SUBCLU [9] which is a density-based clustering method that acquire clusters in axis-parallel subspaces. Data acquisition is performed by the sensors and features that are needed for recognizing the human activity are extracted and the clusters are obtained using SUBCLU. *Neil et al* [3] depicts a human activity recognition approach based on nonparametric analogous of trajectory data and simultaneous motion features that is combined with the bayesian networks. The data is acquired from a tracker that is color-based. The activity is inferred using Hidden Markov Model.

B. Model-based approaches

In Model-Based Approach, a human model is constructed for action recognition. *Li Liu et al* [4] constructs a model using kinematic approach that extracts features from a sequence of frames and projects human poses from them. *Alexandros et al* [5]constructs a kinematic model based on contour lines thereby learning sequence of poses. *Alexandro et al* [7] extends the HAR to incremental and adaptive learning. They extract features from sequence of frames thereby projecting the pose representation of humans from videos. The main use of keyposes are to make the system recognize the actions like normal human being. *Vemulapalli et al* [8] constructs a kinematic model by using skeletal representation. They clearly model the geometric relationships between various body parts using geometric functions like translation and rotation in 3D space. Though kinematic feature is efficient, they are high dimensional with small inter class variation. These models wont take into account about the environment during motion. Hence *Al et al* [6] constructs a physics based model that uses dynamic features. Physics-based model is more discriminative than kinematics model because they can capture gravity, ground contact and all other physical interaction with the ground.

IV. HUMAN-OBJECT INTERACTION

In human activity recognition, handling the interactions between the humans and object is a challenging task due to a variety of reasons like the size, shape, position, color of the object etc,..This section describes about how human object interactions are handled in still images and video. The sample human object interaction is shown in figure 3.

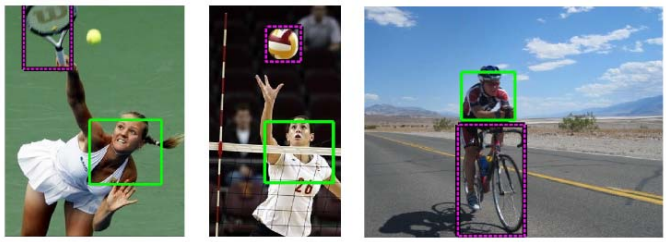

Fig. 3. Sample Human-Object interactions

A. Recognition in still images

Delaitre et al [10] design a human-object interaction recognition system in still images by constructing a co-occurrence model using the already trained parts of the body and object detectors. They use sparse SVM for classification. The performance of the system depends only on the object detectors as the objects size are very small and occlusion has a major effect in recognition. This is overcome by *Yao et al* [11]using the detection of human body, parts, poses and the objects simultaneously with the assumption that recognizing anyone will lead to the detection of other. Since all of these approaches are mainly based on detectors alone, incorporating the contextual information as prescribed by *Chaitanya et al* [12] can increase the recognition rate of the system.

B. Recognition in videos

Prest et al [13] comes up with an approach that extracts the frame and applies object and human detectors to identify both the humans and objects in the particular frame. Learning is done using the relative motion. But the system fails in their performance because of having many detectors .To overcome this *Mohsen et al* [14]utilizes a kernel function for calculating the similarity between two videos by measuring the similarity of the interactions between the human body parts and the objects. The main advantage of this system is that it is independent of labels. *Gupta et al* [15] use motion feature and they depend on the movement of the hand to reach the objects. The major prevalence of this system is the robustness of the velocity profile. But the system lacks in handling uncontrolled

video and it needs a longer training time. This problem is eliminated by *Prest et al* [13].Both the humans and objects are precisely tracked. Interaction is represented as the relative position and motion of object with respect to the humans. Most of the algorithms discussed above doesn't takes spatial information in account. Thus *Chen et al* [16] adds the spatial information but fails in capturing the temporal information that is resolved by *Victor et al* [17] utilizing the spatio-temporal cues. A low level correlation based tracker is used. The limitation of the system is adding the annotation manually due to the failure of the tracker in some videos.

V. HUMAN-HUMAN INTERACTION

Human-Human interaction is another challenging task in human activity recognition .In this section we divide the interactions into One-to-one interactions and group interactions. One-to-one interactions could be a reasonably action that happens as two humans have a control upon each other. Group interactions depends upon the detection of a group objects in the circumstances of social aggregation. For both situations, the motion features should be fused with psychological and sociological information that rule social relations. The Sample one-to-one interaction is depicted in figure 4.

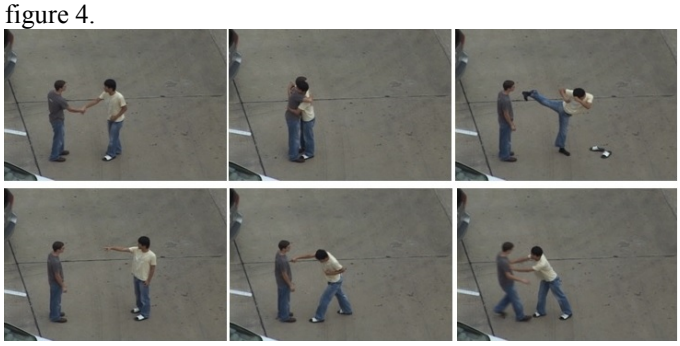

Fig. 4. Sample One-to-One Interactions

A. One-to-One Interactions

Kiwon et al [18] develops a human-human interaction dataset and also evaluates the system with many features such as joint distance, joint motion, plane and velocity. They performed the interaction recognition on the whole collection of frames using Multiple Instance Learning (MIL).MIL is a framework that is used to boost the incorrect frames in the training data. MILBoost developed by *Viola et al* [19] joins any boosting algorithm with MIL .Here not every object is labeled individually instead they are grouped into a bag of objects. On selection MILBoost seeks for the objects with higher weight. Thus MILBoost algorithm allots higher positive on the subspace of objects and these objects prevail subsequence learning. But the system fails in multiple view-points .Hence *Alazrai et al* [20] introduces a Motion Pose Geometric Descriptor (MPGD) that is view-invariant and used for human-human interaction classifier and prediction. A hierarchical system is developed with one representation layer and three classification layers. Image acquisition is performed on Kinect sensor. The 3D joint positions are captured for each human from the obtained RGBD images. In the first stage SVM is used to place the frame into particular states based on the spatio-temporal configuration of each person on the frame. The frames are combined to give the sub-activities of two persons in the second classification layer. The final classification is performed using Constrained Dynamic Time Warping (CDTW).This system can be applied in security systems because of its ability to predict the activity within a limited number of frames. But the system holds good only for two person interaction and fails for multiple interactions.

B. Group Interactions

Cristani et al. [21] identify the F-formations in the frames to interpret an interaction occurs between two or more persons is occurring. *Ni et al* [22] categorize the interactions between bunch of peoples can under three variants, namely, self-causality, pair causality, and group-causality .The features characterizing each category are detected using the trajectories. The assumption made in social interaction model depicted in [23] and [24] is people react habitually while walking in groups. The Trajectory rules are formulated for every person within a particular time window, and the best rule is generated with respect to social factors. This algorithm increases the tracking performance by adding additional conditions and an interaction energy potential is clipped to build the dependence between groups of people [25].

VI. DATASETS USED FOR EXPERIMENTATION

There are number of benchmarking datasets are used for Experimental Evaluation of Activity Recognition. These datasets vary from static backgrounds to dynamic backgrounds for handling clutter backgrounds. Various datasets like KTH, WEIZMANN, UCF sports, SBU Kinect Interaction, iLIDS, Gaming, VISOR, USC-SIPI Human Activity Dataset, 3Dpes, ELSA, IXMAS, HMDB 51,PETS, Hollywood, CASIA etc,.. that are used for recognizing both single and multi-person activity. Kinect Interaction Dataset [18] is a human-human interaction recognition dataset that contains eight interactions such as kicking, pushing, punching, approaching each other, exchanging an object, shaking hands, departing and hugging. Gaming Dataset [26] comprises twelve people split into six pairs. Every pair of humans are interacted through a gaming interface which highlights six sports activities like football, volleyball, table tennis, sprint, boxing and hurdles. The USC-SIPI Human Activity Dataset [27] contains the daily activities of humans for ubiquitous activity recognition. The sample SBU Kinect Interaction and Gaming dataset is shown in figure 5 and in figure 6.

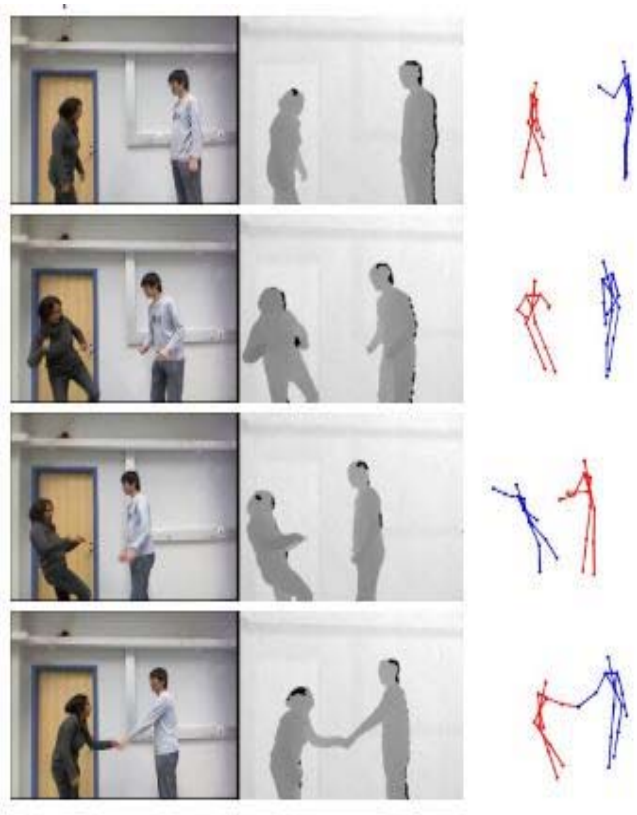

Fig. 5. SBU Kinect Interaction Sample

Fig. 6. Gaming Dataset Sample

VII. EXPERIMENTAL EVALUATION

There are number of parameters like precision, recall, recognition time, training time, accuracy, F-score are used to validate the activity recognition. F-score is performed after calculating precision and recall. Precision is calculated by finding the number of positive recognition to the total number of positive recognition. Recall is found by calculating the ratio of true positive with the actual positive. Their equation are given below respectively. A chart is prepared and given below in Fig. 7 with two parameters such as methodologies adapted and the accuracy achieved with the datasets.

$$\text{F-score} = ((2*P*R))/(P+R) \quad (1)$$

$$\text{Precision} = \frac{\text{No.of instances of correct positive recognition}}{\text{Total no.of positive recognition}} \quad (2)$$

$$\text{Recall} = \frac{\text{No.of instances of positive recognition found}}{\text{Total no.of relevant input instances}} \quad (3)$$

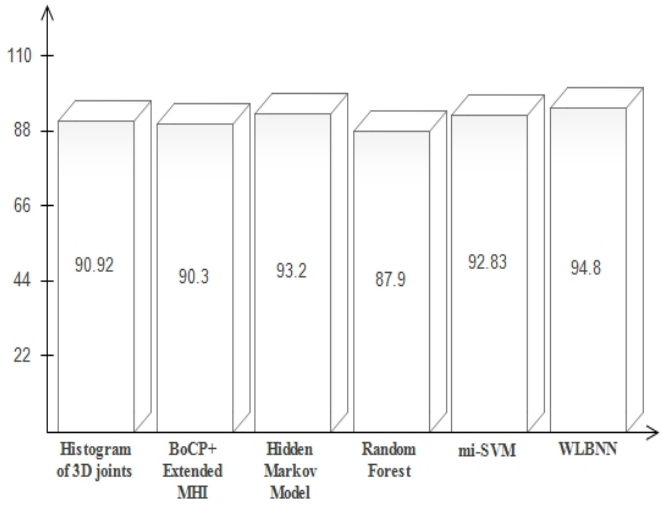

Fig. 7. Accuracy of different algorithms

VIII. APPLICATIONS OF HAR

The major applications of Human Activity Recognition varies from Content-based Video Analytics, Robotics, Human- Computer Interaction, Human fall detection, Ambient Intelligence, Visual Surveillance, Video Indexing etc...The Content based Video Analytics has gain its importance due to increase in the number of websites that share videos. Hence there is a need to develop an effective indexing and methods to store the videos efficiently. The vital applications needed for designing Human-Machine interfaces are helping the elderly people by developing context-aware computing, monitoring of health and fitness, building a smart rooms which would respond to the people gestures and so on. The automatic detection of abnormal activities in video surveillance is not limited to detecting unauthorized people entry, ATM fraud detection and abnormal crowd behavior. The activity recognition can be applied to even behavioral biometrics that involves in understanding methods and their algorithms to identify the human uniquely based on their behavioral cues.

IX. CHALLENGES AND FUTURE RESEARCH DIRECTIONS IN HUMAN ACTIVITY RECOGNITION

Many issues are still open and deserve further research in human activity recognition system. Some of them are discussed in this section. In Background Modeling the main issues are Gradual variations of the lighting conditions in the scene, Small movements of non-static objects such as the

branches of tree and bushes blowing in the wind noise image, due to a poor quality image source, Permanent variations of the objects in the place, Multiple objects moving in the scene both for long and short periods, Dynamic background, Camouflage, Lightning Variation, Intrinsic scale variation, Shadows and Bootstrapping. The recognition of human becomes difficult due to View invariance, change in style, change in anthropometry and change in dressing. Differentiating similar actions and handling Human Object interaction is still an open research topic. Detecting and recovering the missing limbs/Joints from the constructed models is computationally demanding. Tracking Multiple Objects are difficult and identifying the abnormal actions like abnormal crowd behavior, Fraud detection within a limited number of training data is a cumbersome task.

X. CONCLUSION

This survey has showcased the various Human Activity Recognition algorithms, various methodologies adapted for human object interaction in both still images and videos, different human-human interaction techniques and types of Classifications used for all these recognition. The goal is to provide an extensive survey and comparison of different techniques and approaches of Human Activity Recognition. Using this survey, some of the challenges and future directions are also highlighted. In summary, the literature in Human Activity Recognition depicts major progresses in various aspects. However these works still have not addressed the various challenges of activity recognition like incorporating the context of the scene, human-object interaction in videos, real-time activity prediction to its full extent.

REFERENCES

[1] S. Umakanthan, S. Denman, C. Fookes, and S. Sridharan, "Multiple instance dictionary learning for activity representation," in Pattern Recognition (ICPR), 2014 22nd International Conference on, Aug 2014, pp. 1377–1382.

[2] H. Zhang and O. Yoshie, "Improving human activity recognition using subspace clustering," in Machine Learning and Cybernetics (ICMLC), 2012 International Conference on, vol. 3, July 2012, pp. 1058–1063.

[3] N. Robertson and I. Reid, "A general method for human activity recognition in video," Computer Vision and Image Understanding, vol. 104, no. 2, pp. 232–248, 2006.

[4] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," Cybernetics, IEEE Transactions on, vol. 43, no. 6, pp. 1860–1870, Dec 2013.

[5] A. A. Chaaraoui, P. Climent-Pacrez, and F. Flarez-Revuelta, "Silhouette based human action recognition using sequences of key poses," Pattern Recognition Letters, vol. 34, no. 15, pp. 1799 – 1807, 2013, smart Approaches for Human Action Recognition.

[6] A. Mansur, Y. Makihara, and Y. Yagi, "Inverse dynamics for action recognition," Cybernetics, IEEE Transactions on, vol. 43, no. 4, pp. 1226–1236, Aug 2013.

[7] A. Chaaraoui and F. Florez-Revuelta, "Adaptive human action recognition with an evolving bag of key poses," Autonomous Mental Development, IEEE Transactions on, vol. 6, no. 2, pp. 139–152, June 2014.

[8] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, June 2014,pp. 588–595.

[9] K. Kailing, H.-P. Kriegel, and P. Kröger, "Density-connected subspace clustering for high-dimensional data," in Proc. SDM, vol. 4. SIAM, 2004.

[10] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in Advances in neural information processing systems, 2011, pp. 1503–1511.

[11] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human Poses," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 9, pp. 1691–1703, 2012.

[12] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," in Computer vision and pattern recognition workshops (CVPRW), 2010 IEEE computer society conference on. IEEE, 2010, pp. 9–16.

[13] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human object interactions in realistic videos," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 4, pp. 835–848, 2013.

[14] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, and V. Leung, "A similarity measure for analyzing human activities using human-object interaction context," in Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014, pp. 2368–2372.

[15] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 10, pp. 1775–1789, 2009.

[16] C.-Y. Chen and K. Grauman, "Efficient activity detection with max subgraph search," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 1274–1281.

[17].V. Escorcia and J. C. Niebles, "Spatio-temporal human-object interactions for action recognition in videos," in Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. IEEE, 2013, pp. 508–514.

[18] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. IEEE, 2012.

[19] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in Advances in neural information processing systems, 2005, pp. 1417–1424.

[20] R. Alazrai, Y. Mowafi, and C. G. Lee, "Anatomical-plane-based representation for human–human interactions analysis," Pattern Recognition, vol. 48, no. 8, pp. 2346–2363, 2015.

[21] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, "Towards computational proxemics: Inferring social relations from interpersonal distances," in Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011, pp. 290–297.

[22] B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 1470–1477.

[23] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in Computer Vision–ECCV 2010. Springer, 2010, pp. 452–465.

[24] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp.261–268.
[25] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 3161– 3167.
[26] V. Bloom, V. Argyriou, and D. Makris, "G3di: A gaming interaction dataset with a real time detection and evaluation framework," in Computer Vision-ECCV 2014 Workshops. Springer, 2014, pp. 698–712.
[27] M. Zhang and A. A. Sawchuk, "Usc-had: A daily activity dataset For ubiquitous activity recognition using wearable sensors," in ACM International Conference on Ubiquitous Computing (Ubicomp) Workshop on Situation, Activity and Goal Awareness (SAGAware), Pittsburgh, Pennsylvania, USA, September 2012.

978-1-5090-2552-7