International Journal of Advanced Technology and Engineering Exploration, Vol 5(47)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2018.547006

Review Article

A survey on human activity prediction techniques

Manju D.[1*] and Radha V.[2]
Ph. D Research Scholar, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore[1]
Professor and Head, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore[2]

Abstract
Nowadays, in order to prevent criminal behaviors or traffic accidents, video surveillance systems have become more and more popular in both outdoor and indoor places such as offices, departmental stores, public places, railway stations, and airports, etc. So, there is a great demand for an intelligent system to detect abnormal events in videos. In the surveillance tasks, people are generally the main objects of interest. Even though, recognition of human action is an emerging topic in the field of computer vision, detection of abnormal event is recently attracting more research attention. Abnormal behaviors can be identified as irregular behavior from the normal ones. Certainly, various techniques and approaches are proposed in order to ensure human safety. This paper presents a survey on different human activity prediction techniques in video surveillance system. Initially, different techniques developed by previous researchers are studied in detail. Then, the limitations in those techniques are also addressed to suggest further improvement on human activity prediction in videos using advanced techniques. The efficiency of the different human activity prediction techniques is proved by comparing their parameters. The comparison results show the best human activity prediction technique among them.

Keywords
Video surveillance system, Human activity prediction, Human behavior detection, Abnormal behavior detection, Human action recognition.

1.Introduction
Initially, video surveillance systems [1] have been used to capture, store and distribute video. Almost always, analysis of these videos has been carried out by operators manually. Smart surveillance systems are being increasingly installed in order to raise alarms in potentially dangerous circumstances. The main objective of video surveillance system requires detection of abnormal and suspicious activities as opposed to normal activities. It should be able to alert the human operator of an existing suspicious activity like "stole a purse in a crowded area". Such abnormal and suspicious activities in videos can be predicted by going through the basic steps namely motion detection, interesting objects tracking and behavior analysis [2, 3]. Lower level image processing techniques correspond to motion detection and interesting object tracking. At a higher level, machine learning techniques are used to perform behavior analysis.

Human activity prediction [4] is one of the important processes of abnormal behavior prediction. Various intelligent systems can benefit from human activity prediction. In a smart room, people's intention of activity can be predicted by video surveillance system, so that the system will adaptively provide services, even help if necessary. In the crowded scenes of videos, it is very difficult to find out the abnormal events. The detection of abnormal events can be differentiated as global abnormal event and local abnormal event. When behavior of the group, in the global scene is abnormal, it is referred as global abnormal event. When behavior of an individual is different from their neighbor's behavior, local abnormal event is identified. The main objective of this paper is analyzing various human activity prediction techniques to predict abnormal events in video surveillance system. The analysis is carried out based on the merits and demerits and each human abnormal activity prediction techniques compared in terms of equal error rate, area under curve, true positive, false positive, detection rate and area under ROC to select the best technique for human activity prediction.

*Author for correspondence

2.Survey on prediction of human activity in video

A temporally-weighted generalized time warping (TGTW) [5] was proposed for human activity prediction. Either a complete or incomplete activity video was decomposed into a sequence of short video segments. Then, classical bag-of-visual-words model was employed to represent each segment by the local spatial-temporal statistics. By doing this, the comparison between a reference sequence and query sequence boiled down to the problem of aligning their corresponding segment sequence. Finally the similarity derived from the TGTW was combined with the k-nearest neighbors' algorithm to predict the activity class of an input sequence.

A joint action and interaction learning approach [6] was proposed to understand the human activities in videos. In this approach, a novel conditional random field model was proposed which enabled the estimation of human actions and interactions in videos jointly and simultaneously without introducing latent variables. Hence it is called as supervised learning method. It derived a new algorithm which solved the interference problem in the joint learning framework and compared its effectiveness against the original inference. Moreover, two optimization algorithms called alternating search and belief propagation were proposed to solve the relevant inference problem.

An optimization model [7] was proposed for human activity recognition inspired by information on human-object interaction. This model utilized insight to integrate spatiotemporal features with information on human-object interaction to predict human activity categories in realistic videos. It provided a long-term prediction based on spatiotemporal features which were extracted by using a deep 3-dimensional conventional network. Then extended the spatiotemporal features by integrate with information on human-object interaction created from an object detection model.

A robust and human activity recognition scheme called ReHAR [8] was proposed for human activity prediction. This scheme was used to handle single person activities and group activities prediction. Initially, an optimal flow image for each video frame was created and then both video frames and their corresponding optimal flow images were given as input to a Single Frame Representation Model. It generated representation for images. Finally, based on the generated representations a long short term memory (LSTM) model was used to predict the final activities.

A deep neural network (DNN) [9] model was presented for prediction of human activity. This model was used for human activity estimation using multi-view sequences of raw images. It was a single model which incorporated feature discriminator and extractor. It consisted of three parts are convolutional neural network (CNN) BLOCK, multiple stacked long short-term memory residual (MSLSTMRes) and a dense layer. These parts enabled discrimination of human activity such as sit down and walk by merely using sequences of raw images.

A Deep learning framework [10] was proposed to analysis uncertainty related to dyadic human activities at a small temporal granularity. This framework reported at what degree of certainty each activity was occurred from definitely not occurring to definitely occurring. For this process, CNN based unary probabilities and pairwise relations between body joints were extracted. From each frame in a time slice, the features were extracted and examined different temporal aggregation scheme to create a descriptor for the whole time slice.

Context-associative hierarchical memory model [11] was proposed for human activity recognition and prediction. It recognized human activities based on the incoming visual content of previous experienced activities. The high-level activity was parsed into consecutive sub-activities and a context cluster was built to model the temporal relations. Then the semantic attributes of the sub-activity was organized by a concept hierarchy. A series of similarity functions were defined based on the hierarchy which was turn into the recognition computing into intervals over the contextual memory.

A mem-long short-term memory (mem-LSTM) [12] model was proposed for prediction of action in videos. It predicted action in the early stage in which a memory module was introduced to record several hard-to-predict samples and a variety of early observations. It utilized CNN and LSTM to model partial observed video input. The LSTM was augmented with a memory module to achieve better performance.

A probabilistic framework [13] was proposed for prediction of dynamical evolution of human activities from a single image. It inferred dynamic information related with a human pose. The inference problem

was posed as a non-parametric density problem on a non-Euclidean manifold of linear dynamical models. The proposed framework is a direct modeling which was intractable. It estimated the density for the test sample under consideration. Statistical inference on the estimated density provided the quantities of interest like the most probable future motion of human and the amount of motion information conveyed by a pose.

Dominant sets framework [14] was used to detect abnormal behavior in a tracking scenario. Dominant sets framework was an unsupervised learning framework which modeled normal behaviors to detect anomalous behaviors more easily. The problem was casted as one of outlier detection in tracks. Instead of directly extracting meaningful normal behavior class, it allowed to over segmenting the behavior space and hence the above approach works as a robust non parametric density estimation approach. Behavior falling in the low-density area of the phase space is termed as abnormal behavior. This framework has worked reliably even on low-level motion features as opposed to the high level actions.
An anomaly detector [15] was proposed for detection and localization of anomalies in crowded scenes. It had span spatial, time and space scale using a joint representation of video appearance, globally and dynamics consistent of inference. For this purpose, the crowded scenes was modeled with a hierarchy of mixture of dynamic texture (MDT) models, equated temporal anomalies to discriminant saliency, and combined the scores of anomalies across space, time and scale with a conditional random field (CRF). The spatial saliency score was produced by a center surround discriminant saliency detector and the temporal saliency scores was produced by a model of normal behavior that was learned from training data. The multi scale scores were used in CRF that ensures global consistency of the anomaly judgments.

A Normal Behavior Model [16] was proposed for behavior modeling and abnormal event detection by using low level features. This method was directly processed with event characterization and behavior modeling using low level features. Initially, a normal behavior model called co-occurrence matrix was built through learning statistics about co-occurring events in a spatio-temporal volume. By the mutual information between motion label sequences, the concept of co-occurring events was defined. After that, the co-occurrence matrix was used as a potential function in a Markov random field framework to define, as the video streams in the probability of observing new volumes of activity. The behavior of objects was observed in the training phase. It was used to detect the moving objects which differ from the objects observed in the training phase. It was achieved with the help of co-occurrence matrix.

A novel approach AMC [17] was proposed for detection of abnormal events of crowds. This approach was started with estimating optical flows. Then adjacency-matrix based clustering (AMC) was applied in the human crowd scene which clustered the human crowds into groups in unsupervised manner. A model of force field was used to characterize the behaviors of the crowd with crowd size, attributes, position and orientation when the clusters of human crowds were obtained. Behaviors of human crowds were detected from this information. Moreover, anomalies of human crowd(s) present in the scene were also detected using AMC.

A novel criterion called sparse reconstruction cost (SRC) [18] for detection of abnormal events in the crowded scene. In order to calculate the normality of testing samples, SRC was proposed over the normal dictionary. The prior weight of each basis was introduced during sparse representation. Therefore, it became a more robust method compared to the other outlier detection criteria. In addition to this, the over completed normal bases were condensed into a compact dictionary through designing a novel dictionary selection with group sparsity constraint. As the group sparsity involved a low rank structure, the problem was reformulated as using matrix decomposition that handled huge volume of training samples by minimizing the memory requirement. The column wise coordinate descent was used to solve the matrix decomposition. Hence the SRC method was able to detect both the local and global abnormal events in crowded scenes.

A novel particle entropy approach Gaussian mixture model (GMM) [19] was proposed to represent the crowd distribution information and to detect abnormal behavior. This approach avoided unstable foreground extraction and also reduced the computational complexity for abnormal behavior detection. The approach used GMM to detect the abnormal crowd behaviors. Crowd distribution information along with crowd speed information were used together to compute the parameters of GMM and abnormal behaviors of crowd were predicted.
A generalized activity prediction framework [20] was proposed for prediction of human activity. This

framework modeled three key aspects of activity are causality, context-cue and predictability. The causality was modeled by probabilistic suffix tree (PST). Both large and small order Markov dependencies between action units were represented. The context-cue was modeled by using sequential pattern mining (SPM). This utilized interactive objects as cues for predicting human activity. Then, the predictability was modeled by predictive accumulative function. This learned the predictability pattern of each kind of activity automatically from the data. These three key aspects were particularly beneficial for prediction of various kinds of human activity in diverse environment.

3.Discussions

This section listed merits and demerits of different human activity prediction techniques for abnormal event detection in video surveillance system those are discussed in the previous section. From the analysis of literature survey on human activity prediction techniques for abnormal event prediction the following limitations are observed.

- Because of searching for the corresponding part between activity videos, [5] still has prediction error.
- For the belief propagation process in [6], the cost for computation is high.
- The [7] cannot fetch semantic information in activity scenes.
- For only small datasets, [8] is applicable.
- The temporal evaluation should improve the performance of [9].
- When the activity is performed in a highly different way, errors may occur in [10].
- The accuracy of [11] is same as the conventional methods accuracy.
- The performance of [12] is relatively low.
- The accuracy of [13] is low.
- In certain situations, the error rate of [14] is high.
- Because of sparsity of scenes, the performance of [15] is quite low.
- When dealing with moving objects in videos, [16] used a threshold value which highly influences the performance.
- The [17] concentrated only on global events instead of local ones which lead to worse detection rate.
- Memory cost is high in [18].
- The [19] is applicable only for simple dataset.
- The activities with shallow structure are not suited for [20].

From the following *Table 1*, the most challenging issues in human activity prediction for abnormal event detection in video surveillance system are observed and an ideal solution is identified to overcome those issues for abnormal event detection.

Table 1 Comparison of different human activity prediction techniques

References	Methods	Merits	Demerits	Performance Metrics
[5]	Temporally-Weighted Generalized Time Warping, k-nearest neighbor	More effective to recognize partially observed activity	Still have prediction error due to searching for the corresponding part between activity videos	Recognition rate (UT-Interaction dataset) = 0.81 Recognition rate (DARPA-Y1 dataset) = 0.63 Recognition rate (UCF sports dataset) = 0.815
[6]	Joint learning and interaction approach	Joint training approach outperforms the methods using heuristic interactions	Computation cost is high for process of belief propagation	Precision Recall (PR) curve (TVHI dataset) = 72.53% PR curve (UT dataset) = 79.56% PR curve (BIT dataset) = 84.97% Recognition Accuracy (TVHI dataset) = 65.8% Recognition Accuracy (UT dataset) = 90.3% Recognition Accuracy (BIT dataset) = 83.0%
[7]	Optimization Model	High computational efficiency	It cannot fetch semantic information in activity scenes	Recognition accuracy (UCF101 dataset) = 92.5%

References	Methods	Merits	Demerits	Performance Metrics
[8]	Robust and efficient human activity recognition	High activity recognition accuracy	Applicable for small dataset	Mean Average Precision (NCAA Basketball dataset) = 0.589 Mean Average Precision (UCF sport action dataset) = 0.928
[9]	Deep Neural Network, MSLSTMRes	High recognition rate	Temporal evaluation in online scenario can improve performance	Recognition rate (IXMAS dataset) = 90.15%
[10]	Deep learning framework	Analysis uncertainty in human activities effectively	Errors may occur corresponds to the cases when the activity is performed in a highly different way	Mean Average Precision (Time-Slice Activity Prediction dataset) = 0.729
[11]	Context-associative Hierarchical Memory Model	Less time cost	Achieves similar accuracy as the conventional methods	Activity prediction accuracy (sensor dataset) = 0.86
[12]	mem-Long Short-Term Memory	Memory of this method still memorize some of the challenging samples to be classified in the middle or late stage of videos	Performance is relatively low	Prediction Accuracy (UCF-101 dataset) = 51.02% Prediction Accuracy (Sports-1M dataset) = 57.60%
[13]	Probabilistic framework	More robust	Less accuracy	Recognition accuracy (UFC dataset) = 57.9% Recognition accuracy (Cross dataset) = 50.9% Recognition accuracy (CMU dataset) = 50.5%
[14]	Dominant sets	More robust method	Error rate is high in special situations	Equal Error Rate = 26%
[15]	Anomaly detector	Guarantee globally consistent inference	Due to sparsity of scenes the performance of anomaly detector is quite weak	AUC (UMN dataset) = 99.5 AUC (Subway entrance dataset) = 89.7 AUC (Subway exist dataset) = 90.8 AUC (U-turn dataset) = 95.2
[16]	Normal behavior model	Avoid the risk of error propagation, Robust to noise	Threshold value influence the performance while dealing with multiple moving objects	True Positive = 90.5% False Positive = 9.5%
[17]	Adjacency- Matrix based Clustering	More effective to detect unusual events for uncontrolled environments of surveillance videos	Focused only on global events instead of local ones will leads to worse detection rate	Detection Rate (UMN dataset) = 94%
[18]	Sparse Reconstruction Cost	Detects both local and global abnormal events	Memory cost is high	Equal Error Rate (UCSD ped1 dataset) = 19% Rate of Detection (UCSD ped1 dataset) = 46% Area Under Curve (UCSD ped1 dataset) = 46.1%
[19]	Gaussian Mixture Model	Low computational complexity	Not applicable for complex dataset	Area under ROC (UMN dataset) = 0.9911
[20]	Generalized activity prediction framework	Achieves superior performance for predicting global activity classes and local action units	Activities with shallow structure are not suited for this framework	Prediction accuracy (MPII-cooking dataset) = 0.88 Prediction accuracy (UCI-OPPORTUNITY dataset) = 0.95

4.Conclusion

In this paper, a detailed survey on human activity prediction for abnormal event detection in video surveillance system is presented. It is obvious all researchers have tried in different techniques or approaches to predict the human activity effectively for better detection of abnormal event in videos. The discussion on human activity prediction techniques provides the recent developments in abnormal event detection. They are analyzed by describing the novel ideas incorporated in them. The analysis of these techniques provides better understanding of steps involved in each process thus increasing the scope for finding the efficient techniques to achieve better performance. Based on the analysis, generalized activity prediction framework has better performance than the other techniques. Moreover, in the future contribution, this framework can be further improved by,

- Along with the temporal information, spatial information, size, and motion correlation among objects should be considered for human activity prediction.
- Instead of considering single video to predict the human activity, multiple video surveillance cameras should be considered for human activity prediction.
- In addition to human activity, faces should be recognized to prevent the abnormal events.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Tsakanikas V, Dagiuklas T. Video surveillance systems-current status and future trends. Computers & Electrical Engineering. 2018; 70:736-53.

[2] Hu W, Tan T, Wang L, Maybank S. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2004; 34(3):334-52.

[3] Taha A, Zayed HH, Khalifa ME, El-sayed M. Exploring behavior analysis in video surveillance applications. International Journal of Computer Applications. 2014; 93(14):22-32.

[4] Wang H, Yuan C, Shen J, Yang W, Ling H. Action unit detection and key frame selection for human activity prediction. Neurocomputing. 2018; 318:109-19.

[5] Wang H, Yang W, Yuan C, Ling H, Hu W. Human activity prediction using temporally-weighted generalized time warping. Neurocomputing. 2017; 225:139-47.

[6] Wang Z, Jin J, Liu T, Liu S, Zhang J, Chen S, et al. Understanding human activities in videos: a joint action and interaction learning approach. Neurocomputing. 2018; 321:216-26.

[7] Liu X, You T, Ma X, Kuang H. An optimization model for human activity recognition inspired by information on human-object interaction. In international conference on measuring technology and mechatronics automation 2018 (pp. 519-23). IEEE.

[8] Li X, Chuah MC. ReHAR: robust and efficient human activity recognition. Winter conference on applications of computer vision 2018 (pp. 362-71). IEEE.

[9] Putra PU, Shima K, Shimatani K. Markerless human activity recognition method based on deep neural network model using multiple cameras. In international conference on control, decision and information technologies 2018 (pp. 13-18). IEEE.

[10] Ziaeefard M, Bergevin R, Lalonde JF. Deep uncertainty interpretation in dyadic human activity prediction. In international conference on machine learning and applications 2017 (pp. 822-5). IEEE.

[11] Wang L, Zhao X, Si Y, Cao L, Liu Y. Context-associative hierarchical memory model for human activity recognition and prediction. IEEE Transactions on Multimedia. 2017; 19(3):646-59.

[12] Kong Y, Gao S, Sun B, Fu Y. Action prediction from videos via memorizing hard-to-predict samples. In AAAI 2018.

[13] Lohit S, Bansal A, Shroff N, Pillai J, Turaga P, Chellappa R. Predicting dynamical evolution of human activities from a single image. In proceedings of the conference on computer vision and pattern recognition workshops 2018 (pp. 496-505). IEEE.

[14] Alvar M, Torsello A, Sanchez-Miralles A, Armingol JM. Abnormal behavior detection using dominant sets. Machine Vision and Applications. 2014; 25(5):1351-68.

[15] Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014; 36(1):18-32.

[16] Benezeth Y, Jodoin PM, Saligrama V. Abnormality detection using low-level co-occurring events. Pattern Recognition Letters. 2011; 32(3):423-31.

[17] Chen DY, Huang PC. Motion-based unusual event detection in human crowds. Journal of Visual Communication and Image Representation. 2011; 22(2):178-86.

[18] Cong Y, Yuan J, Liu J. Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition. 2013; 46(7):1851-64.

[19] Gu X, Cui J, Zhu Q. Abnormal crowd behavior detection by using the particle entropy. Optik-International Journal for Light and Electron Optics. 2014; 125(14):3428-33.

[20] Li K, Fu Y. Prediction of human activity by discovering temporal sequence patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014; 36(8):1644-57.

Manju D. has completed B. Com, M.C.A., M. Com.(CA), M.Phil(CS). Currently, she is pursuing Ph.D. degree in Computer Science in Avinashilingam Institute of Home Science and Higher Education for Women. She has more than fifteen years of experience in the academia. She is currently Assistant Professor at Department of Computing, Coimbatore Institute of Technology, Coimbatore. Her research interest includes Data Analytics, Image Processing and Machine Learning. She has presented two papers in International Conferences and published a paper in International Journal.
Email: manjuphd2018@gmail.com

Radha V. has completed M.Sc., PGDOR., PGDCA., B.Ed., M.Phil., and Ph.D. Currently, she is heading the Department of Computer Science at Avinashilingam Institute of Home Science and Higher Education for Women. She has more than twenty-nine years of experience in the academia. Her research interest includes Signal and Image Processing, Data Mining, Query Optimization. She has guided 20 M.Phil. Scholars and 10 Ph.D. Scholars. She is currently guiding 7 Ph.D. scholars. She has published 72 articles in international journals, 3 articles in in-house journals, presented papers in 12 National Conferences and has contributed 6 books, 11 Book chapters and edited a book. She has organized 23 conferences / workshops / seminars and has delivered invited talks twice. She has been recognized with awards on three occasions.