International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

A Survey on Context Driven Activity Recognition and Analysis in Wide Area Surveillance

Anjana B H
Dept. of CSE, DSI, Bangalore
anjanahuliraj@gmail.com

Rashmi S R
Asst. Professor
Dept. of CSE, DSI, Bangalore

Dr. R Krishnan
Professor and Head
Dept. of ISE, DSI, Bangalore

***Abstract*— Activity recognition and analysis is an important area of computer vision research. Its applications include surveillance systems, patient monitoring systems, security and defense applications and other such fields. Most of these applications require an automated recognition of high-level activities, composed of multiple simple (or atomic) actions of persons. In many cases of activity recognition, contextual information learnt from the surroundings in the video plays an important role in increasing the accuracy of recognition. This paper provides an analysis of state-of-the-art research on activity recognition using context, giving a comparison of the advantages and limitations of each approach. In any activity recognition algorithm the method used to represent the relationships among activities plays an important role. To represent relationships, activity recognition algorithms usually use graphical models. The graphical model can be grid based, feature graphs, tree structured or hierarchical representation. Different methodologies are discussed under each classification. This review tries to identify the best methodology for activity recognition and analysis and will also provide the motivation for future research in more productive areas. The future direction of research is obviously encouraged and dictated by pressing applications like the surveillance and monitoring of public facilities like train stations, underground subways or airports, monitoring patients in a hospital environment, and other similar applications.**

***Index Terms*-Context-aware activity recognition, wide area activity analysis, spatiotemporal relationships.**

I. INTRODUCTION

Activity analysis or activity recognition is an important area in computer vision research today. The ability to recognize complex activities from videos leads to several important applications. Streamlined surveillance systems in public places like airports, railway stations, shopping malls and other such places require detection of abnormal and suspicious activities, rather than recognizing normal activities. Recognition of activities also enables the real-time monitoring of patients, children, and elderly persons. Building an intelligent traffic monitoring system also becomes possible with the use of real time activity recognition systems. Such recognition systems can also be used in security and defense applications.

The goal of activity recognition is to automatically analyze (or recognize) activities occurring within a known or unknown video sequence. At the simplest level, where a video contains only one occurrence of an activity, the objective of the system is to correctly classify the video as belonging to a particular activity category. In more realistic cases, the continuous recognition of activities must be performed by detecting starting and ending times of all occurring activities from an input video.

Activity recognition in realistic environments is a challenging task because surveillance typically consists of a number of people entering and exiting the scene over an interval of time. Therefore it is nearly always impossible to estimate the number of activities occurring in the scene and the number of people involved. This variability is termed as unconstrained environments. Most of the activity recognition algorithms focus on the region of activity alone ignoring the surroundings. These methods assume the number of objects, scale and view point of the scene and are not effective in more challenging environments. Often, by examining the surroundings of an activity under consideration provides useful clues about the activity. This information obtained from the surroundings is termed as “Context” of the activity. The usage of context in activity recognition algorithm is one of the many bases of classifications.

Activity recognition algorithms can be classified on this basis into two categories, first, the algorithms that do not make use of contextual information and, second, the algorithms that use contextual information.

The objective of this article is to provide an overview of the methodologies that use contextual information for activity recognition. Before going into the details of the methodologies, we need to know the advantages of using contextual information. Considering a wide area surveillance video which is unconstrained against a sports video which is governed by a definite set of rules, it is observed that in the long term surveillance video the activities may be related but do not unfold according to set rules. In such videos, several activities may influence each other casually while others may occur independently. However, theses casualties are not trivial due to the presence of multiple actors in the video sequence. Also, tracking and activity recognition in these videos becomes challenging due to the presence of clutter and occlusion. Contextual information can overcome these limitations to a great extent. Also, the use of contextual information along with a traditional activity recognition system improves the recognition rates considerably.

For easy reading, this survey is organized into 5 sections. Section 1 gives an introduction on activity recognition and analysis and its various applications. It also introduces the concept of contextual information and the advantages of using it along with recognition or analysis algorithms. Section 2 overlays the literature survey carried out on the various previous surveys and describes the uniqueness of this survey. Section 3 gives a brief overview about each of the approach selected for the survey. Section 4 overlays the result analysis in terms of various performance metrics. It also addresses the various advantages and disadvantages of each approach and arrives at the most suitable approach for activity recognition and

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

analysis in wide area surveillance. This is followed by section 5 which gives a conclusion for the survey and the future direction in which this survey can be taken.

II. RELATED WORK

There has been other related survey on activity recognition algorithms. Several previous reviews have focused on human action detection. Cedras and Shah in [1], Gavrila in [2], Aggarwal and Cai in [3] discussed human action recognition approaches as a part of their review. In [1], the paper provides a review of the developments in the computer vision aspect of motion based recognition during that time. In [2], the survey identifies a number of promising applications and provides an overview of recent developments in this domain. The scope of this survey is limited to work on whole-body or hand motion; it does not include work on human faces. In [3], the paper gives an overview of the various tasks involved in motion analysis of the human body.

Thomas B. Moeslund and Erik Granum in [4] present a comprehensive survey of computer vision-based human motion capture literature and methodologies. The survey focuses on a general overview based on taxonomy of system functionalities and covers research work over the period 1981-2001.

Zhao et al in [5] present an up-to-date critical survey of still- and video-based face recognition research. To provide a comprehensive survey, the authors not only categorize existing recognition techniques but also present detailed descriptions of representative methods within each category. In addition, relevant topics such as psychophysical studies, system evaluation, and issues of illumination and pose variation are covered.

Visual surveillance in dynamic scenes, especially for humans and vehicles, is currently one of the most active research topics in computer vision. It has a wide area of promising applications, such as access control in special areas, human identification at a distance, crowd flux statistics and congestion analysis, detection of anomalous behaviors, and interactive surveillance using multiple cameras, etc. In general, the processing framework of visual surveillance in dynamic scenes includes the following stages: modeling of environments, detection of motion, classification of moving objects, tracking, understanding and description of behaviors, human identification, and fusion of data from multiple cameras. Weiming Hu et al in [6] review recent developments and general strategies of all these stages.

Kruger et al. in [7] reviewed human action recognition approaches while classifying them on the basis of the complexity of features involved in the action recognition process. In their survey, the authors analyze the different approaches taken to-date within the computer vision, the robotics and the artificial intelligence community for the representation, recognition, synthesis and understanding of action. Their review focused especially on the planning aspect of human action recognitions, considering their potential application to robotics.

The past few years has witnessed a rapid addition of video cameras in all walks of life and has resulted in a sudden increase of video content. Several applications such as content-based video annotation and retrieval, highlight extraction and video summarization require recognition of the activities occurring in the video. The analysis of human activities in videos is an area with increasingly important consequences from security and surveillance to entertainment and personal archiving. Several challenges at various levels of processing—robustness against errors in low-level processing, view and rate-invariant representations at midlevel processing and semantic representation of human activities at higher level processing—make this problem hard to solve. Turaga et al in [8] present a comprehensive survey of efforts in the past couple of decades to address the problems of representation, recognition, and learning of human activities from video and related applications. In their paper, approaches are first categorized based on the complexity of the activities that they want to recognize, and are then classified in terms of the recognition methodologies they use.

Ronald Poppe in his survey, [9], provides a detailed overview of the current advances in the field of vision based human action recognition. The author also discusses the limitations of the state of the art and outline promising directions of research.

Joshua et al in their survey, [10], describe the current state-of-the-art image-processing methods for automatic-behavior-recognition techniques, with focus on the surveillance of human activities in the context of transit applications. The main goal of this survey is to provide researchers in the field with a summary of progress achieved to date and to help identify areas where further research is needed. This survey provides a thorough description of the research on relevant human behavior-recognition methods for transit surveillance.

However, most of these reviews focused on the introduction and summarization of activity recognition methodologies, but do not provide a means to compare different types of activity recognition approaches. Aggarwal and Ryoo in their survey, [11], present an interclass and intraclass comparisons between methodologies, while providing an overview of human activity recognition approaches which are categorized on the approach-based taxonomy. The survey enables a reader (even one from a different field) to understand the context of the development of human activity recognition and understand the advantages and disadvantages of the different categories of recognition methodologies.

The previous surveys mentioned above deal directly or indirectly with detection and analysis of human activities. Our survey focuses on identifying the recent activity recognition algorithms that use contextual information. The results and the accuracy of each methodology are discussed. The discussion enables a reader to understand the performance of each approach and choose the approach that best suits their application.

III. APPROACHES TO ACTIVITY RECOGNITION AND ANALYSIS

A major drive for research in complex activity recognition has been in the selection of features and their representations, most of which has dealt with single activity clips. Different representations have been used in activity recognition such as Space Time Interest Points (STIP) and histogram of optical flow. But, with advance in computer vision, long duration videos have to be dealt with and thus the need to explore the contextual information between different activities in the video to arrive at a representation of the scene.

Graphical models are the methods that are commonly used to encode relationships in video and activity analysis. The following figure, fig 3.1, shows a tabulation of the different methods used for

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

graphical representation of context and the various approaches under each representation method.

A grid based belief propagation method has been introduced in [12] for pose estimation. Tracking human body poses in any type of video has many important applications. However, this is challenging in realistic scenes due to background clutter, variation in human appearance, and self-occlusion. The complexity of pose tracking is further increased when there are multiple people whose bodies may inter-occlude. The proposed approach is a three stage approach with multilevel state representation that enables a hierarchical estimation of 3D body poses. This method addresses various issues including automatic initialization, data association, self and inter-occlusion. At the first stage, humans are tracked as foreground blobs, and their positions and sizes are coarsely estimated. In the second stage, parts such as face, shoulders, and limbs are detected using various cues, and the results are combined by a grid-based belief propagation algorithm to infer 2D joint positions. The derived belief maps are used as proposal functions in the third stage to infer the 3D pose using data-driven Markov chain Monte Carlo. Experimental results on several realistic indoor video sequences show that the method is able to track multiple persons during complex movement including sitting and turning movements with self and inter-occlusion.

Co-occurring activities and their dependencies have been studied using Dependent Dirichlet Process- Hidden Markov Models (DDP-HMMs) [13]. Daniel Kuettel, Michael D. Breitenstein, Luc Van Gool, Vittorio Ferrari in "What's going on? Discovering Spatio-Temporal Dependencies in Dynamic Scenes", Computer Vision and Pattern Recognition, 2010, present two novel methods to automatically learn spatio-temporal dependencies of moving agents

Spatio-temporal relationships have played an important role in the recognition of complex activities. Methods such as [14] and [15] explore spatio-temporal relationships at a feature level. M. S. Ryoo and J. K. Aggarwal in "Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities", International Conference on Computer vision, 2009, [14], introduce a novel matching, spatio-temporal relationship match, which is designed to measure structural similarity between sets of features extracted from two videos. This match hierarchically considers spatio-temporal relationships among feature points, thereby enabling detection and localization of complex non-periodic activities. In contrast to previous approaches to 'classify' videos, this approach is designed to 'detect and localize' all occurring activities from continuous videos where multiple actors and pedestrians are present. This paper implements and tests the methodology on a newly-introduced dataset containing videos of multiple interacting persons and individual pedestrians. The results confirm that this system is able to recognize complex non-periodic activities (e.g. 'push' and 'hug') from sets of spatio-temporal features even when multiple activities are present in the scene.

Yimeng Zhangy, Xiaoming Liuz, Ming-Ching Changz, Weina Gez, and Tsuhan Cheny in "Spatio-Temporal Phrases for Activity Recognition", [15], propose an approach that efficiently identifies both local and long-range motion interactions; taking the "push" activity as an example, this approach can capture the combination of the hand movement of one person and the foot response of another person, the local features of which are both spatially and temporally far away from each other. The computational complexity is in linear time to the number of local features in a video. The extensive experiments show that this approach is generically effective for recognizing a wide variety of activities and activities spanning a long term, compared to a number of state-of-the-art methods.

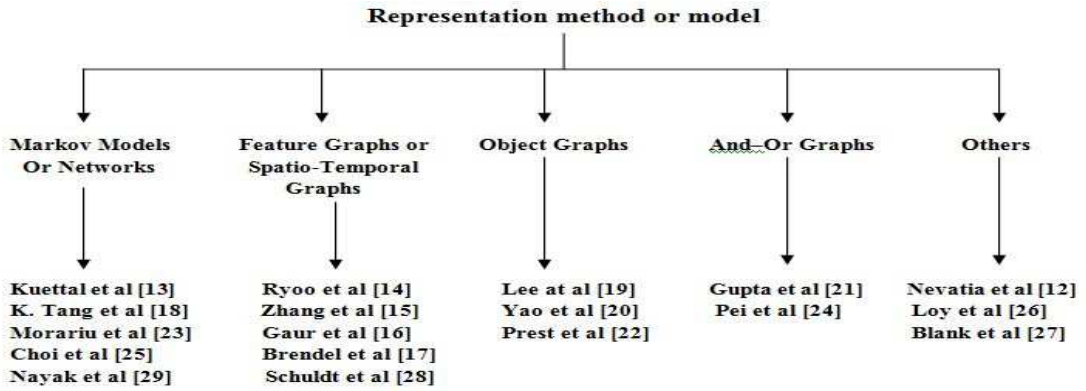

Fig. 3.1 **Representation methods or models**

in complex dynamic scenes. They allow discovering temporal rules, such as the right of way between different lanes or typical traffic light sequences. To extract them, sequences of activities need to be learned. While the first method extracts rules based on a learned topic model, the second model called DDP-HMM jointly learns co-occurring activities and their time dependencies. To this end Dependent Dirichlet Processes is applied to learn an arbitrary number of infinite Hidden Markov Models. In contrast to previous work, we build on state-of-the-art topic models that allow to automatically infer all parameters such as the optimal number of HMMs necessary to explain the rules governing a scene. The models are trained offline by Gibbs Sampling using unlabeled training data.

The spatial and temporal relationships between space time interest points have been encoded as "feature graphs" in [16]. This paper proposes a new feature model based on a string representation of the video which respects the spatio-temporal ordering. This ordered arrangement of local collections of features (e.g., cuboids, STIP), which are the characters in the string, are initially matched using graph-based spectral techniques. Final recognition is obtained by matching the string representations of the query and the test videos in a dynamic programming framework which allows for variability in sampling rates and speed of activity execution. The method does not require tracking or recognition of body parts, is able to identify the region of interest in a cluttered scene, and gives reasonable performance with even a single query example. The paper tests the approach in an example-based video retrieval framework

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

with two publicly available complex activity datasets and provides comparisons against other methods that have studied this problem.

Complex activities were represented as spatio-temporal graphs representing multi-scale video segments and their hierarchical relationships in [17]. This paper advances prior work by learning what activity parts and their spatiotemporal relations should be captured to represent the activity, and how relevant they are for enabling efficient inference in realistic videos. Videos are represented by spatiotemporal graphs, where nodes correspond to multiscale video segments, and edges capture their hierarchical, temporal, and spatial relationships. Access to video segments is provided by this new, multiscale segmenter. Given a set of training spatiotemporal graphs, their archetype graph is learnt, and pdf's associated with model nodes and edges. The model adaptively learns from data relevant video segments and their relations, addressing the "what" and "how." Inference and learning are formulated within the same framework – that of a robust, least-squares optimization – which is invariant to arbitrary permutations of nodes in spatiotemporal graphs. The model is used for parsing new videos in terms of detecting and localizing relevant activity parts. This method outperform the state of the art on benchmark Olympic and UT human-interaction datasets, under a favorable complexity-accuracy trade-off.

Variable length Hidden Markov models are used to identify activities with high level of intra class variabilities in [18]. This paper tackles the problem of understanding the temporal structure of complex events in highly varying videos obtained from the Internet. Towards this goal, a conditional model trained in a max-margin framework is utilized that is able to automatically discover discriminative and interesting segments of video, while simultaneously achieving competitive accuracies on difficult detection and recognition tasks. This method introduces latent variables over the frames of a video, and allows the algorithm to discover and assign sequences of states that are most discriminative for the event. The model is based on the variable-duration hidden Markov model, and models durations of states in addition to the transitions between states. The simplicity of the model allows it to perform fast, exact inference using dynamic programming, which is extremely important when being able to process a very large number of videos quickly and efficiently.

Spatial relationships between objects have been modeled using graphs for new object discovery in [19]. This paper proposes to leverage knowledge about previously learned categories to enable more accurate discovery. The article introduces a novel object-graph descriptor to encode the layout of object-level co-occurrence patterns relative to an unfamiliar region, and show that by using it to model the interaction between an image's known and unknown objects new visual categories can be better detected. Rather than mine for all categories from scratch, this method can continually identify new objects while drawing on useful cues from familiar ones.

The authors Bangpeng Yao and Li Fei-Fei in [20] use objects as context for activities and vice versa. This paper proposes a new random field model to encode the mutual context of objects and human poses in human-object interaction activities. This approach casts the model learning task as a structure learning problem, of which the structural connectivity between the object, the overall human pose, and different body parts are estimated through a structure search approach, and the parameters of the model are estimated by a new max-margin algorithm.

Association of tracks with activities and the generation of a high level storyline model using AND-OR graphs has been performed in an EM framework in [21]. This paper presents an approach to learn a visually grounded storyline model of videos directly from weakly labeled data. The storyline model is represented as an AND-OR graph, a structure that can compactly encode storyline variation across videos. The edges in the AND-OR graph correspond to causal relationships which are represented in terms of spatio-temporal constraints. The Paper formulates an Integer Programming framework for action recognition and storyline extraction using the storyline model and visual groundings learned from training data.

Spatial context between objects and activities has been modeled in [22]. This paper introduces a weakly supervised approach for learning human actions modeled as interactions between humans and objects. This approach is human-centric: first localize a human in the image and then determine the object relevant for the action and its spatial relation with the human. The model is learned automatically from a set of still images annotated only with the action label. This approach relies on a human detector to initialize the model learning. For robustness to various degrees of visibility, a detector is built that learns to combine a set of existing part detectors. Starting from humans detected in a set of images depicting the action, this approach determines the action object and its spatial relation to the human. Its final output is a probabilistic model of the human-object interaction, i.e. the spatial relation between the human and the object.

Spatio-temporal context in structured videos with manually defined rules has been modeled using Markov Logic Networks in [23]. This paper presents a framework for the automatic recognition of complex multi-agent events in settings where structure is imposed by rules that agents must follow while performing activities. Given semantic spatio-temporal descriptions of what generally happens (i.e., rules, event descriptions, physical constraints), and based on video analysis, the events that occurred is determined. Knowledge about spatiotemporal structure is encoded using first-order logic using an approach based on Allen's Interval Logic, and robustness to low-level observation uncertainty is provided by Markov Logic Networks (MLN). The article's main contribution is that it integrates interval-based temporal reasoning with probabilistic logical inference, relying on an efficient bottom-up grounding scheme to avoid combinatorial explosion. Applied to one-on-one basketball, this framework detects and tracks players, their hands and feet, and the ball, generates event observations from the resulting trajectories, and performs probabilistic logical inference to determine the most consistent sequence of events.

Models such as the AND-OR graphs or other tree structures have been suggested in [24] for modeling sports sequences and office environments. These models however, are more suited for structured environments where there are a set of rules governing the behavior of people such as in sports, or where the number of objects/activities or the combinations of sub-activities are limited as in an office environment. The authors in [24] present an event parsing algorithm based on Stochastic Context Sensitive Grammar (SCSG) for understanding events, inferring the goal of agents, and predicting their plausible intended actions. The SCSG represents the hierarchical compositions of events and the temporal relations between the sub-events. The alphabets of the SCSG are atomic actions which are defined by the poses of agents and their interactions with objects in the scene. The temporal relations are used to distinguish events with similar structures, interpolate missing portions of events, and are learned from the training data.

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

Methods like those described in [25] use context to infer a collective activity using single person activities. The approach described is a framework to recognize collective human activities using the crowd context and introduce a new scheme for learning it automatically. This scheme is constructed upon a Random Forest structure which randomly samples variable volume spatio-temporal regions to pick the most discriminating attributes for classification. Unlike previous approaches, this algorithm automatically finds the optimal configuration of spatio-temporal bins, over which to sample the evidence, by randomization. This enables a methodology for modeling crowd context. 3D Markov Random Field is employed to regularize the classification and localize collective activities in the scene. The paper also demonstrates the flexibility and scalability of the proposed framework in a number of experiments over other methods.

The method described in [26] deals with recognizing a single activity over multiple cameras by topology inference, person re-identification and global activity interpretation. A novel approach is proposed to understand activities from their partial observations monitored through multiple non-overlapping cameras separated by unknown time gaps. In this approach, each camera view is first decomposed automatically into regions based on the correlation of object dynamics across different spatial locations in all camera views. A new Cross Canonical Correlation Analysis is then formulated to discover and quantify the time delayed correlations of regional activities observed within and across multiple camera views in a single common reference space. In this work it is showed that learning the time delayed activity correlations offers important contextual information for (i) spatial and temporal topology inference of a camera network; (ii) robust person re-identification and (iii) global activity interpretation and video temporal segmentation. Crucially, in contrast to conventional methods, this approach does not rely on either intra-camera or inter-camera object tracking; it thus can be applied to low-quality surveillance videos featured with severe inter-object occlusions.

M. Blank et al in their work, [27], regard human actions as three dimensional shapes induced by the silhouettes in the space time volume. We adopt a recent approach for analyzing 2D shapes and generalize it to deal with volumetric space time action shapes. This method utilizes properties of the solution to the Poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure and orientation. It is shown that these features are useful for action recognition, detection and clustering. The method is fast, does not require video alignment and is applicable in (but not limited to) many scenarios where the background is known. Moreover, the robustness of this method to partial occlusions, non-rigid deformations, significant changes in scale and viewpoint is demonstrated.

C. Schuldt et al proposed a local approach for recognizing human actions in their work [28]. Local space-time features capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. In this paper it is demonstrated how such features can be used for recognizing complex motion patterns. Video representations in terms of local space-time features are constructed and then such representations are integrated with SVM classification schemes for recognition. For the purpose of evaluation a new video database is introduced, containing sequences of six human actions performed by 25 people in four different scenarios. The presented results of action recognition justify the proposed method and demonstrate its advantage compared to other relative approaches for action recognition.

A novel method for capturing spatio-temporal context between activities using a Markov Random Field (MRF) has been proposed by Nandita M Nayak et al in [29]. The structure of the MRF is improvised upon during test time and not pre-defined, unlike many approaches that model the contextual relationships between activities. Given a collection of videos and a set of weak classifiers for individual activities, the spatio-temporal relationships between activities are represented as probabilistic edge weights in the MRF. This model provides a generic representation for an activity sequence that can extend to any number of objects and interactions in a video. A greedy hill combination method to construct the MRF is proposed. An inference on the MRF gives the estimate of the activities in the video sequence.

The next section deals briefly with the advantages and disadvantages of the methodologies and tries to identify the best approach. It is followed by the result analysis of the best approach.

IV. DISCUSSION OF PUBLISHED RESULTS

One of the many advantages of using contextual information for activity recognition or activity analysis, as discussed in section 1, is that it improves accuracy of the method employed for activity detection or analysis. Hence, the main parameter or performance metric used for result analysis and discussion is accuracy. Accuracy in activity recognition or analysis can be defined as the ratio of the correctly classified activity to the total number of activities used. The following table, Table 4.1, gives the accuracy of each method either independently or as compared with baseline methodologies. Three other metrics have been used; weighted average error or WE give the average error of classification, average precision or AP is the ratio of the correctly classified activities to correctly detected activities, recall ratio or RR.

The discussion so far has dealt with individual analysis of the methods or approaches. However, regardless of the individual accuracy the suitable approach for any application is itself application dependent and may vary from application to application. For example, a method suited for a common surveillance video may not be suitable for sports video analysis. Hence, before choosing any method of implementation a complete understanding of the application is necessary. The following discussion tries to arrive at the best method that is suitable for versatile video types.

Though methods proposed in [12] and [13] deal with complex and co-occurring activities, realistic videos usually consists of multiple actors and actions over a period of time. Hence, for analysis of long term wide area analysis of surveillance video requires representations like that described in [25] and [29] which use a Markov Random Field for modeling relationships between activities.

Methods such as [14], [15] and [16] explore spatio-temporal relationships at a feature level and have been encoded as “feature graphs”. Although such methods have been applied to multiple activities occurring simultaneously, it may not be practical to construct such graphs over long term video sequences and they do not explore the relationships across activities.

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

Method	Average Accuracy
Nevatia et al [12]	23.54 (WE)
Kuettal et al [13]	73%
Ryoo et al [14]	0.9
Zhang et al [15]	Increases accuracy by 7-10%
Gaur et al [16]	65%
Brendel et al [17]	68%
K. Tang et al [18]	11.25% (AP)
Lee at al [19]	78.55%
Yao et al [20]	83.3%
Gupta et al [21]	70%
Prest et al [22]	80%
Morariu et al [23]	0.92 (RR)
Pei et al [24]	90%
Choi et al [25]	70.9%
Loy et al [26]	99.7%
Blank et al [27]	93.9%
Schuldt et al [28]	70%
Nayak et al [29]	77.9%

AP - Average Precision **RR – Recall Ratio**
WE - Weighted average Error

Table 4.1 Average accuracy of the discussed approaches

Complex activities were represented as spatio-temporal graphs representing multi-scale video segments and their hierarchical relationships in [17]. Most of these papers focus on the modeling of low level features for recognition. Variable length Hidden Markov models are used to identify activities with high amount of intra-class variabilities in [18]. However, in [29] the authors have modeled the spatio-temporal relationships between different activities which form a higher level representation.

Spatial relationships between objects have been modeled using graphs for new object category discovery in [19]. The authors in [20] use objects as context for activities and vice versa. Similarly, association of tracks with activities and the generation of a high level storyline model using AND-OR graphs has been performed in an EM framework in [21]. Spatial context between objects and activities has been modeled in [22]. Spatio-temporal context in structured videos with manually defined rules has been modeled using Markov Logic Networks in [23]. However, the authors in [25] and [29] propose a generalized formulation for context modeling that is suited for unconstrained video sequences such as outdoor surveillance videos. The key aspects of such sequences are that there is no constraint on the number of actors in the scene or the number of activities in the sequence. Also, different actors in the scene may act independently or interact with each other if they choose to do so.

Models such as the AND-OR graphs or other tree structures have been suggested in the past [21], [24] for modeling sports sequences and office environments. These models however, are more suited for structured environments where there are a set of rules governing the behavior of people such as in sports, or where the number of objects/activities or the combinations of sub-activities are limited as in an office environment. Applying such models to unconstrained sequences can be laborious due to the exponential number of combinations of activities which have to be learnt here to construct such models.

Similarly, papers such as [25] use context to infer a collective activity using single person activities. In such sequences however, it is assumed that all participating persons/objects contribute to the collective activity. Whereas, in a typical surveillance scenario, different actors may or may not be interacting with each other, therefore such models cannot be directly applied in such scenarios.

The authors in [26] deal with recognizing a single activity over multiple cameras by topology inference, person re-identification and global activity interpretation. In a typical surveillance scenario however, we are dealing with a set of different activities which may or may not be correlated, therefore a Markov Random Field as proposed in [29] is a more suited model to capture these complex spatio-temporal relationships.

Most of the works discussed so far deals with short duration videos or with videos with a predefined structure such as sports videos. However, in [29] the authors propose to define the structure of the graph on the test sequence rather than use a predefined structure.

Hence, we can safely conclude that for wide area activity analysis in unconstrained environments which consists of multiple actors and actions, where it is hard to predict the number of activities occurring in the scene and the number of people involved in those activities, the method as proposed in [29] suits best either individually or combined with other methodologies.

V. CONCLUSION

Computer recognition of activities is an important area of research in computer vision with applications in many diverse fields. The application to surveillance is natural in today's environment, where the tracking and monitoring people is becoming an integral part of everyday life. Other applications include human-computer interaction, biometrics based on gait or face, and hand and face gesture recognition. Another important application is in traffic monitoring, where automatic monitoring of vehicles is becoming more popular. Computer recognition of activities is also becoming an integral part of security applications in various fields. We have provided an overview of the current approaches to activity recognition. The approaches are diverse and yield a spectrum of results.

In this review we have summarized the methodologies previously explored for the recognition of human activity, and discussed the advantages and disadvantages of those approaches. And at the end of the discussion, we have arrived at an approach which may suit best for typical surveillance videos.

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 2, Issue 5, May 2015)

The future direction of research is obviously encouraged and dictated by applications. The pressing applications are the surveillance and monitoring of public facilities like train stations, underground subways or airports, monitoring patients in a hospital environment or other health care facilities, monitoring activities in the context of UAV surveillance, and other similar applications. All of these applications are trying to understand the activities of an individual or the activities of a crowd as a whole and as subgroups.

The preceding areas of research, the space-time feature-based approaches, manifold learning, rigid/non rigid motion analysis, and hierarchical approaches briefly mentioned are but a small glimpse into the large number of methodologies being pursued today. Hopefully, a review in another ten years will document significant progress in human activity recognition, to the extent that off-the-shelf systems will be readily available.

REFERENCES

[1] Claudette Cedras and Mubarak Shah, "Motion-based recognition: a survey," Image and Vision Computing, Volume 13, Number 2, March 1995.
[2] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey," Computer Vision and Image Understanding, Vol. 73, No. 1, January, pp. 82–98, 1999.
[3] J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," Computer Vision and Image Understanding, Vol. 73, No. 3, March, pp. 428–440, 1999.
[4] Thomas B. Moeslund and Erik Granum, "A Survey of Computer Vision-Based Human Motion Capture," Computer Vision and Image Understanding, Vol. 81, pp. 231–268, 2001.
[5] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, "Face Recognition: A Literature Survey," ACM Computing Surveys, Vol. 35, No. 4, December 2003, pp. 399–458.
[6] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank, "A Survey on Visual Surveillance of Object Motion and Behaviors," IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 34, No. 3, August 2004.
[7] Volker Kr uger, Danica Kragic, Ale˘s Ude and Christopher Geib, "The Meaning of Action: A review on action recognition and mapping," Advances in Computer Vision and Machine Intelligence CVM| 2007.
[8] Pavan Turaga, Rama Chellappa, V. S. Subrahmanian, and Octavian Udrea, "Machine Recognition of Human Activities: A Survey," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 18, No. 11, November 2008.
[9] Ronald Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, Vol. 28, pp. 976–990, 2010.
[10] Joshua Candamo, Matthew Shreve, Dmitry B. Goldgof, Deborah B. Sapper, and Rangachar Kasturi, "Understanding Transit Scenes: A Survey on Human Behavior-Recognition Algorithms," IEEE Transactions on Intelligent Transportation Systems, Vol. 11, No. 1, March 2010.
[11] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," ACM Computing Surveys, vol. 43, no. 3, 2011, Article 16.
[12] M. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 27–38, Jan. 2009.
[13] D. Kuettel, M. Breitenstein, L. V. Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in Proc. Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 1951–1958.
[14] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in Proc. Int. Conf. Computer Vision, Kyoto, Japan, 2009, pp. 1593–1600.
[15] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spation-temporal phrases for activity recognition," in Proc. Eur. Conf. Computer Vision Part III, Florence, Italy, 2012, pp. 707–721.
[16] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, "String of feature graphs analysis of complex activities," in Proc. Int. Conf. Computer Vision, Barcelona, Spain, 2011, pp. 2595–2602.
[17] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," in Proc. Int. Conf. Computer Vision, Barcelona, Spain, 2011, pp. 778–785.
[18] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in Proc. Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 1250–1257.
[19] Y. J. Lee and K. Grauman, "Object graphs for context aware category discovery," in Proc. Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 1–8.
[20] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in Proc. Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 17–24.
[21] A. Gupta, P. Srinivasan, J. Shi, and L. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in Proc. Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 2012–2019.
[22] A. Prest, C. Schmid, and V. Ferrari, Weakly Supervised Learning of Interactions Between Humans and Objects, INRIA, Tech. Rep., 2010.
[23] V. I. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in Proc. Computer Vision and Pattern Recognition, Providence, USA, 2011, pp. 3289–3296.
[24] M. Pei,Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in Proc. Int. Conf. Computer Vision, Barcelona, Spain, 2011, pp. 487–494.
[25] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in Proc. Computer Vision and Pattern Recognition, Providence, USA, 2011, pp. 3273–3280.
[26] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," Int. J. Comput. Vis., vol. 90, no. 1, pp. 106–129, Oct. 2010.
[27] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Proc. Int. Conf. Computer Vision, Beijing, China, 2005, vol. 2, pp. 1395–1402.
[28] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in Proc. Int. Conf. Pattern Recognition, Cambridge, U.K., 2004, vol. 3, pp. 32–36.
[29] Nandita M. Nayak, Yingying Zhu, and Amit K. Roy-Chowdhury, "Exploiting Spatio-Temporal Scene Structure for Wide-Area Activity Analysis in Unconstrained Environments," IEEE Transactions on Information Forensics and Security, Vol. 8, No. 10, October 2013.