Anticipating Human Activities from Surveillance Videos – A Survey

Shanmughapriya Murugesan[1] Ranjith Balakrishnan[2]
[1]PG Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Velammal Engineering College, Surapet, Chennai-600066, India

Abstract— An important aspect of human vision is anticipation (prediction), which humans use extensively in day to day activities through interactions. Predicting which activities will be happening next, in unconstrained surveillance videos will help in monitoring and pin pointing the malicious activity prone areas. Anticipating activity is a probabilistic process - from the videos containing frames of onset of activity; the goal is to predict all possible futures. Prediction also helps in precise classification of activities in learning environments. Major work includes: (i) Action Recognition. (ii) Activities Classification. (iii) Developing the learning algorithms. And (iv) Predictions. Activities are modeled by Integral Histogram of Spatio Temporal Features and how the feature distribution change over time is captured. This paper focuses on the methods used in Person identification from video frames, such as Histogram of Oriented Gradients and Motion Boundary Histograms and how the feature extraction helps in action recognition and classification. The Bag of Words approach is studied to gain knowledge on prediction algorithms.

Key words: Action Recognition, HOG, MBH, SVM, Bag of Words, Bag of Visual Words, Activity Prediction

I. INTRODUCTION

An automated technique to recognize and classify, the events and actions performed by humans from video data – Human Action Recognition and Classification, is a significant field in computer vision. Smart home systems, quality of care service for elderly and needy, automated surveillance system applications are possible outcomes of Action Recognition Automation. Action recognition has evolved from simple human action such as walking and running [1, 2, and 3] to more complex recognitions involving multiple persons and objects [4, 5, and 6].

Action recognition approaches can be classified based on the approaches [Fig 1].

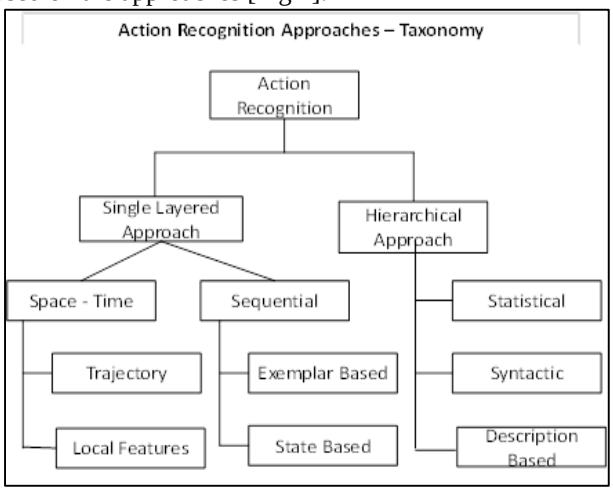

Fig. 1: Taxonomy of Approaches to Human action recognition from video data

In the past five years approach using Spatio Temporal features obtained success in recognizing activities [7].

The bag of words paradigm ignores the location of features during feature extraction and attained success in action classification [videos in 7, 5].

Most research is based on the classification of the activities after the complete observation of entire video sequence. Four levels of video understanding provides the complete details needed for classification, levels include

- Object Level
- Tracking Level
- Pose Level
- Activity Level

This classification helps in annotation of activities with labels for persons, objects, events and actions. Based on these Labels and the classification, it is possible to classify the unlabeled video whose feature can be added to the supervised learning.

Most recognition approaches misses the important aspect of activity analysis that is anticipating what is going to happen next.

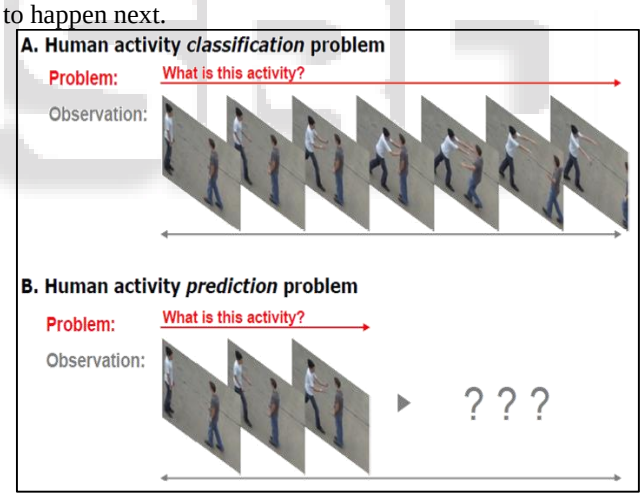

Fig. 2: Activity Classifications Vs. Action Prediction. In prediction problem the action to be anticipated on the onset of activities compared to the classification where activities are annotated after watching the whole sequence.

A. Action Classification

The work involves in categorizing the videos into a limited number of classes. Let's assume a video observation O, with frames 1 to t. Then the system must be automated to assign a class A_i to the video, based on the fact that system found the class through algorithms defined.

K Means Clustering [8], SVM [9] are used for classification purposes. Localizing the activities from continuous video frames is prioritized in action classification. The length of the video (time t) is neglected since the posterior probability is only taken in classification and the time t is independent of the activity.

Let, P (A_i / O, t) is the Probability of activity class Ai from video O of length t.

R^* be the variable describing the progress of activity.

At final calculations the Activity class with maximum P (A_i / O, t) will be chosen as activity happening in O.

$$P(A_i / O, t) = P(A_i R^* / O)$$

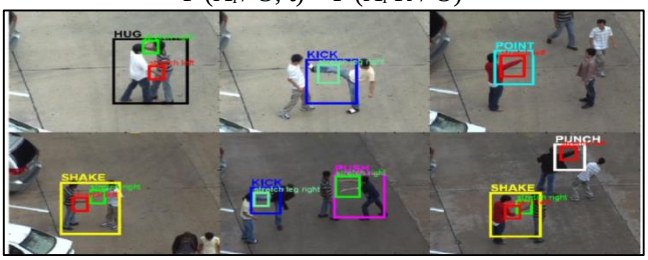

Fig. 3: The Action Classes are Shake, Hug, Punch, Kick, Push and Point. In each frame the critical frame deciding the class is highlighted. After the completion of video sequence the respective action class will be having the highest posterior probability, which will be selected for classification.

Prior to classification, the system is trained with the actions to be classified. The training video data set is used and SVM is trained with full video sequence containing the full execution of a single activity.

B. Anticipating Activities

Incomplete or onset of video containing the activity is taken and the system should decide on the 'activity that will be happening in next frames' of the video. Here it is not presumable that the ongoing activity is finished in video sequence. The progress of activity is to be anticipated in video and the most probable action to be predicted.

Let

P (A_j / O, t) is the Probability of activity class Ai from video O of length t.

R be the variable describing the progress of activity.

P (A_p/O, t) is the Action to be anticipated from Video O of length t.

$$P(A_p / O,t) = \sum P(A_p, R / O, t)$$

For example in the above equation R=50 indicates that prediction progresses from 0^{th} frame to 50^{th} frame. The maximum likelihood estimation is used to anticipate the activity.

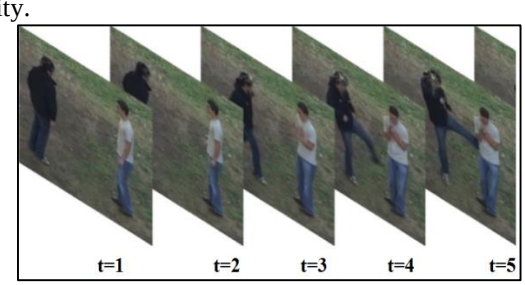

Fig. 4: AP calculated will be the maximum likelihood estimator at t=5, thus at t<=3 and t>=4, the system should be able to predict the Kick activity. The onset of activities only when t is 1 to 4, is provided to the test system to predict the Kick activity.

II. FEATURE EXTRACTION AND CLASSIFICATION

A. HoG Descriptor

The histogram of oriented gradients (HoG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. Navneet Dalal and Bill Triggs [9] first described HOG descriptors at the 2005 Conference on Computer Vision and Pattern Recognition (CVPR). In this work they focused on pedestrian detection in static images, although since then they expanded their tests to include human detection in videos, as well as to a variety of common animals and vehicles in static imagery.

B. Motion Boundary Histogram MBH Descriptor

Optical flow represents the absolute motion between two frames, which contains motion from many sources, i.e., foreground object motion and background camera motion. If camera motion is considered as action motion, it may corrupt the action classification. Various types of camera motion can be observed in realistic videos, e.g., zooming, tilting, rotation, etc. In many cases, camera motion is locally translational and varies smoothly across the image plane.

Since MBH represents the gradient of the optical flow, locally constant camera motion is removed and information about changes in the flow field (i.e., motion boundaries) is kept. MBH is more robust to camera motion than optical flow, and thus more discriminative for action recognition.

C. Bag of Words

The bag-of-words model (BoW model) can be applied to image classification, by treating image features as words. Content Based Image Retrieval (CBIR) systems are used to find images that are visually similar to a query image. The application of CBIR systems can be found in many areas such as a web-based product search, surveillance, and visual place identification.

A common technique used to implement a CBIR system is bag of visual words, also known as bag of features. Bag of features is a technique adapted to image retrieval from the world of document retrieval. Instead of using actual words as in document retrieval, bag of features uses image features as the visual words that describe an image.

Image features are an important part of CBIR systems. These image features are used to gauge similarity between images and can include global image features such as color, texture, and shape. Image features can also be local image features such as speeded up robust features (SURF), histogram of gradients (HOG), or local binary patterns (LBP).

The benefit of the bag-10 of-features approach is that the type of features used to create the visual word vocabulary can be customized to fit the application.

III. LITERATURE REVIEW

M. S. Ryoo [10] provided the solution for activity predictions. An important contribution of this paper is the systematic formulation of the concept of activity prediction, which has not been studied in depth in previous research. This paper presents novel methodologies that reliably

identify unfinished activities from video streams by analyzing their on- sets. The experiments confirm that the approach is able to correctly predict ongoing activities even when the videos containing less than the first half of the activity is provided, in contrast to the previous systems.

Konrad Schindler, Luc Van Gool[11] discusses the factual number of frames required to recognize actions with ease and precision. The paper provides information on; Human vision proves that simple actions can be recognized almost instantaneously. A system for action recognition from very short sequences ("snip- pets") of 1–10 frames. Even local shape and optic flow for a single frame are enough to achieve ≈90% correct recognitions. Snippets of 5-7 frames (0.3-0.5 seconds of video) are enough to achieve a performance similar to the one obtainable with the entire video sequence.

Chenxia Wu, Jiemi Zhang, Bart Selman, Silvio Savarese and Ashutosh Saxena [12] elaborates a new system of watchbot which focuses on real time action recognition. The system uses k-means to cluster the human-skeleton-trajectories/interactive-object-trajectories from all the clips to form a human-dictionary and an object-dictionary, where the system uses the cluster centers as human-words and object- words. In real robotic applications, people perform a very wide variety of actions. These are hard to learn from existing videos on the Internet and there are few with annotations of actions or objects. So the work proposes a probabilistic learning model in a completely unsupervised setting, which can learn actions and relations directly from the data without any annotations, only given the input RGB-D frames with tracked skeletons from Kinect v2 sensor.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid Michael J. Black [13] paper explains about: Both HOG and HOF descriptors are created out of blocks. By choosing the sampling rate identically to the size of a single block, one can reuse these blocks. Once responses per block are computed, descriptors can be formed by concatenating adjacent blocks. This system use descriptors of 3 by 3 blocks in the spatial domain and 2 blocks in the temporal domain, but these parameters can be easily changed. Hence each block is reused 18 times (except for the blocks on the borders of the video volume).

Work mainly focuses on HOF and HOG Descriptors for Action Recognition Focuses on the trade off – Computational Efficiency and Accuracy k means, Hierarchical k means for video representation and classification.

Mohamed H. Elhoseiny, H.M. Faheem, T.M. Nazmy, Eman Shaaban[14] NVidia presents an architecture known as CUDA provided with a SDK, accessible to software developers through standard programming languages. In the same direction, OpenCL, initially developed by Apple Incorporation, became a standard framework for writing Kernels on heterogeneous platforms that consists of CPUs and GPUs. As a result, advances in GPU industry and development have been achieved and a lot of work has been done to get GPU applied in many high performance fields. Examples include Bioinformatics, Image processing and computer vision. For building a channel between GPUs and activity recognition systems, this work investigated many major techniques used in action recognition and computer vision, with the objective of making this connection results in an initial framework, on which many action recognition systems can be built to gain speedup. Figure 2.7 presents the workflow of the proposed GPU framework. The input is raw images are captured of moving camera, and the output is the recognized activity.

IV. SYSTEM DESIGN

The system developed to demonstrate that action recognition can be achieved using local measurements in terms of spatiotemporal interest points. Such features capture local motion events in video and can be adapted to the size, the frequency and the velocity of moving patterns, hence, resulting in video representations that are stable with respect to corresponding transformations.

The ability to perform early detection of activities is tested with the public video dataset containing high level person interactions is also to be done.

A. DataSet

For experiments, the segmented version of the KTH and UT-Interaction dataset containing videos of six types of human activities: (walking, jogging, running, boxing, hand waving and hand clapping) and interactions like, punching and pushing. The UT-Interaction dataset is a public video dataset containing high level human activities of multiple actors. The dataset is composed of two different sets with different environments containing a total of 120 videos of six types of human- human interactions. Each set is composed of 10 sequences, and each sequence contains one execution per activity.

B. Experimental Setup

Global motion of subjects in the database is a strong cue for discriminating between the leg and the arm actions when using histograms of spatio-temporal gradients (HistSTG). This information, however, is (at least partly) canceled when representing the actions in terms of velocity adapted local features.

The integral bag-of-words (BoW) method will be implemented but here the system adopted the cuboid feature descriptors as spatiotemporal features. In principle, the proposed approaches are able to cope with any spatiotemporal feature extractors as long as they provide XYT locations of the features being detected. Extracted features were then clustered into 800 visual words (i.e. k = 800), and integral histograms were constructed.

Standard SVMs, which are designed to classify videos assuming that they contain full activity executions. The testing was performed by applying the learned classifiers to the videos containing ongoing activities.

All sequences were divided with respect to the subjects into a training set (8 persons), a validation set (8 persons) and a test set (9 persons). The classifiers were trained on a training set while the validation set was used to optimize the parameters of each method. The presented recognition results were obtained on the test set.

That is, for each round, videos in one sequence were selected as the testing videos, and videos in the other sequences were used for the training. Integral histograms were constructed from the training videos to recognize activities in the testing videos. This testing/training video

selection process was repeated for 10 rounds, measuring the average recognition accuracy.

C. Work Flow Diagram

The work flow diagram provides a way to understand the system functions at ease.

Work flow for Action Recognition/ Classification and Action Anticipation remains same till some levels, since anticipation can only be done after careful classification of actions.

The motivation was to enable the early detection of unfinished activities from initial observations. The problem has been formulated probabilistically, and presented novel recognition methodologies designed for the efficient prediction of human activities. The experimental results confirmed that the proposed approaches are able to recognize ongoing human-human interactions. The system is able to predict human actions from the onset of video.

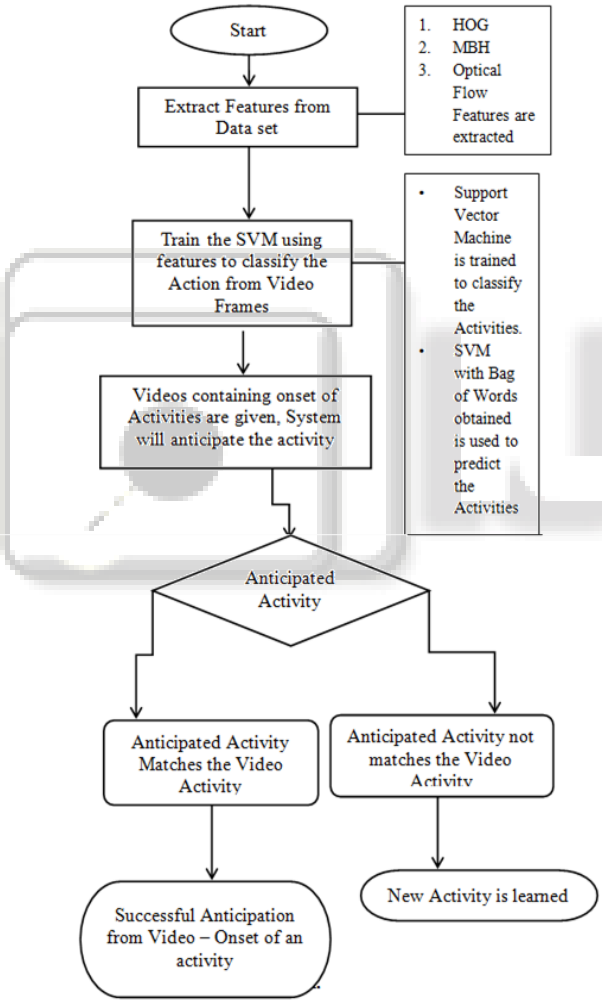

Fig. 6: Experimental Work Flow

V. CONCLUSION

In this paper, we have studied various paradigms of human activity prediction. The motivation was to enable the early detection of unfinished activities from initial observations. We formulated the problem probabilistically, and presented two novel recognition methodologies designed for the efficient prediction of human activities.

REFERENCES

[1] Xiaojiang Peng, Limin Wang, Xing xing Wang, Yu Qiao: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. 1077-3142/© 2016 Elsevier Inc.
[2] Chirag I Patel,Sanjay Garg ,Tanish Zaveri ,Asim Banerjee , Ripal Patel: Human action recognition using fusion of features for unconstrained video
[3] Sequences 0045-7906/© 2016 Elsevier Ltd.
[4] M. S. Ryoo: Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos. 2011 IEEE
[5] Konrad Schindler , Luc Van Gool: Action Snippets: How many frames does human action recognition require? 2011 IEEE
[6] Mohamed H. Elhoseiny, H.M. Faheem, T.M. Nazmy, Eman Shaaban: GPU-Framework for Teamwork Action Recognition. IEEE 2013
[7] Alexander Artikis , Marek Sergot and Georgios Paliouras: A Logic Programming Approach to Activity Recognition. ACM 2013
[8] Chenxia Wu, Jiemi Zhang, Bart Selman, Silvio Savarese and Ashutosh Saxena: Watch-Bot: Unsupervised Learning for Reminding Humans of Forgotten Actions. IEEE 2016 Conference DOI 10.1109/ICRA 2016.7487401
[9] Anitha Edison & Jiji C.V: HSGA: A Novel Acceleration Descriptor for Human Action Recognition. 978-1-4673-8564 IEEE 2015
[10] J.R.R. Uijlings, N. Rostamzadeh, I.C.Duta, N.Sebe: Realtime Video Classification using Dense HOF/HOG. ACM 978-1-4503-2782 April 2014
[11] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid Michael J. Black: Towards understanding action recognition. IEEE Explore 2013
[12] Kuanhong Xu, Ya Lu, Hongwei Zhang, Xuetao Feng, Wonjun Kim, And Jae-Joon Han: Combining Nonuniform Sampling, Hybrid Super Vector, And Random Forest With Discriminative Decision Trees For Action Recognition. 978-1-4799-8339 ©2015 IEEE

