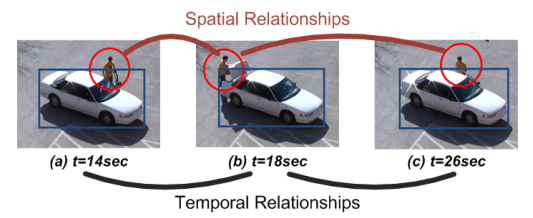
Context Based Activity Recognition in Videos

# Introduction

Activity recognition in a video is the process of locating a moving object in a video. It has variety of uses, some of which are human computer interaction, security and surveillance, video communication and compression, traffic control etc. Adding further to the complexity is the possible need to use object recognition (context) techniques for tracking, a challenging problem in its own right. Visual event recognition is attracting and growing interest, it can be defined as recognition of spatio-temporal visual pattern in videos. Much research work has been done on basic human activity recognition (like walking, turning etc.).

Activities in natural scenes and videos rarely happen independently i.e., activities takes place along with other objects in the scene. The spatial layout of activities and also the sequence in which activities take place also provide useful cues for their understanding. Suppose the activities that took place in same space-time region in Figure 1: the presence of the nearby car gives information about the person’s (bounded by red circle) action, and the relative position of the person and the car tells that activity (a) and (c) are different from (b). But it is difficult to tell what the person is doing in (a) and (c) – getting in or out of the vehicle. If we knew that these activities took place around same time then it would have become immediately clear. This example shows the importance of spatial and temporal relationships for activity recognition.



**Figure 1** [3]

So, rather than modeling activities in video individually, we need to design a framework which models activity using motion and context features. There is no definition of context in computer vision, we consider all the objects detected or other motions present in the video as providing the contextual information about each other. Contextual information can be referred to as additional information about target objects. Also due to various problems such as high intra-class variations and low image resolution, it has been very difficult to detect event in surveillance videos with a good accuracy, here the contextual information plays a more and more important role for results with higher accuracy.

## Motivation

Because of the real time applications of context based activity recognition in video we are highly motivated to do the project and also not much work has been done in this domain. The recognition of activities related to human has recently become a relevant area of research by the variety of applications which shows signs of future success such as anomaly detection in video–surveillance [1, 2], human computer interaction and intelligent monitoring [4]. So being able to recognize activity in a video could be of great use.

# Problem Definition and Objective

## Problem

With the advent of technology everyday new data is being added to the database which has no useful information inside it. For example the surveillance cameras which are planted all various locations to keep the places safe or to keep a record if some unusual activity takes place. But we don’t have any efficient methodology to get the activity of interest from that video. Scanning the entire video for locating some activity is not only a tiresome process but also a waste of time. Unfortunately, the current surveillance system heavily rely on human observers. This limits the capability of these systems.

So some efficient technique is needed to the entire work such that it gives not only the activity region but also the activity which took place at that moment of time. It could reduce tremendous efforts.

## Objective

To implement an approach that recognizes the activities in the video and also can make use of the contextual information to capture activities with higher accuracy. Towards this goal, we utilize a structural model that jointly models the underlying activities which are related in space and time. The learned model is used to optimally label the activities in testing videos.

# Literature Survey

## Related Work

Many existing work using context focus on interaction among objects and their actions [5], environmental conditions such as space location of certain activities in scene[6], and temporal relationship of activities. The Spatio-temporal constraints across activities in a wide area scene are rarely considered.

In [7], a complex activity is modeled by a variable – duration Hidden Markov Model (HMM) on equal length temporal segments. It decomposes a complex activity in sequential actions, which are context of each other. But it considers, only temporal relationships, while neglecting the relationships which exist in space. However, the learning and inference processes of AND-OR graphs become more complex as the graph grows large and more and more activities are learned.

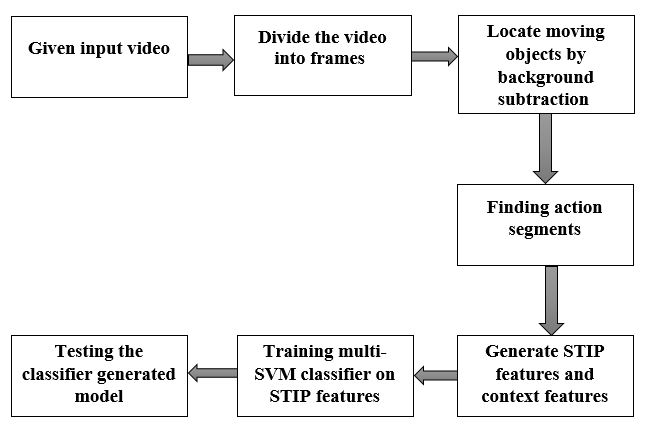
In [9] structural model is proposed to learn both feature level and action level interactions of group members. This method labels each image with a group activity label. How to smooth the labeling results along time is a problem and is not addressed in the paper. Also, these methods aim to recognize group activities and are not suitable in our scenario where activities cannot be considered as the parts of larger activities. In [3], a structural model is used to integrate motion features and context features in and between activities. However, there was no activity segmentation or modeling of the activity duration; only the regions with activity were detected. We propose an alternative method that explicitly models the durations, motion, intra-activity context and the spatio-temporal relationships between the activities and use them in the inference stage for recognition.

# Proposed Methodology

Given a continuous video, we need to recognize activity going on in the sequence of frames. There are various modules that are to be implemented which are shown in block diagram. First of all, we divide the video into different frames. Then we do background subtraction to locate the moving objects. The bounding boxes of moving persons are used as starting of tracking method developed to obtain local trajectories of moving object. Then Spatio-temporal Interest Points are generated only for these motion regions. Thus, STIPs select the features from the motion segments which further aid in classification of the activities.

Then using those generated STIP features we train the multi SVM classifier to generate a model that can generate a model that can classify the testing data with high accuracy.

## Block Diagram Of Proposed Methodology



**Figure 2.** Steps to be used in Activity Recognition

### Converting Video to frames

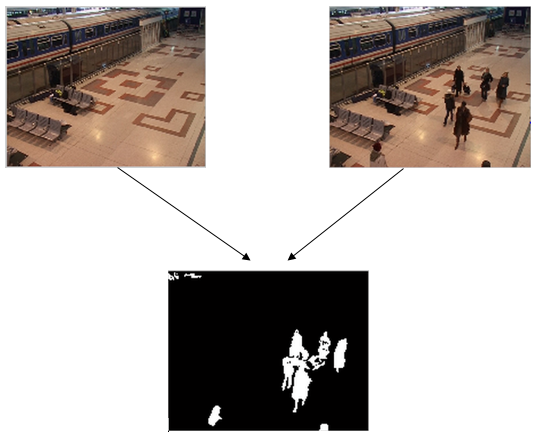
First step is to divide the video into frames. Frames are images that are displayed per second i.e. is also the frame rate. We have taken frames at an interval of 1second. To reduce the complexity of the feature extraction.

### Background Subtraction

Background subtraction is a computational vision process of extracting foreground objects in a particular scene. A foreground object can be described as an object of attention which helps in reducing the amount of data to be processed as well as provide important information to the task under consideration. Background Subtraction is used to locate the moving objects. It is done by **Frame Differencing** subtraction of two successive images. The difference image sequence is calculated by subtracting the pixel value at the same position (x, y) of adjacent frames of the current image.



The difference image contains positive, negative and zero pixel values. Also, the background and static part of the body get eliminated in the difference image. So, it is clear that the difference image has proved to be more suitable for person and background independent activity recognition. It will be clearer by **Figure 3**.



**Figure 3.** Two subsequent frames and their resulting difference image using background subtraction.

### Feature Extraction

Feature Extraction is basically interest point detection, in computer vision it refers to the detection of interest points which could be used for subsequent processing. Characteristics or interest points must be – it has a clear, well found definition ; it has well defined position in space ; it is stable under local and global perturbations ; they should include an attribute to scale, to make it possible to compute interest points from real-life images as well as under scale changes. Types of image features: Edges, Corners/interest points, Blobs/region of interest or interest points, ridges etc.

Based on the output of previous module we will now have the motion regions in the frames then we need to extract features from the resultant image using spatio-temporal interest points STIP [10] features. Also the context features are extracted and stored.

Local image features or interest points dense and abstract representation of patterns in an image. To detect spatio-temporal events, we used the idea of Harris and Forstner interest point operators and detect local structures in space which have more local dissimilarity in both space and time.

The Harris operator is not invariant to scale and correlation is not invariant to rotation. Goal was to develop an interest operator that is invariant to scale and rotation.

***Idea of SIFT***

Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters

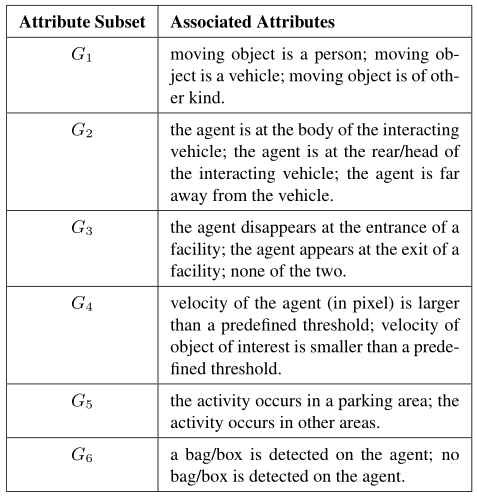
1. *Scale-space extrema detection*: Search over multiple scales and image locations.
2. *Keypoint localization*: Fit a model to detrmine location and scale. Select keypoints based on a measure of stability.
3. *Orientation assignment* : Compute best orientation(s) for each keypoint region.
4. *Keypointdescription* : Use local image gradients at selected scale and rotation to describe each keypoint region.

*Motion and Context feature Descriptors*: If there are M + 1 classes of activities in the scene, including a background class with label 0 and M classes of interest which are labelled 1,…, M. An activity is a 3D region which consists of single or multiple consecutive action segments. An agent is a moving person along a trajectory. **Motion region** at nth frame is a circular region surrounding the moving objects of interest in that frame of activity. **Activity region** is the smallest rectangular region that encapsulates the motion regions over all frames of activity. Based on this we can now encode the motion and contextual information into feature descriptors.

*Intra-activity motion feature descriptor*: Features of an activity that encode motion information extracted from motion features such as STIP features are defined as intra-activity motion features.

*Intra-activity context feature descriptor:* Features that capture the relationships between the agents and other interacting objects, are defined as intra-activity context features. Objects including vehicles, opening/closing entrance/exits doors of facilities, which overlap with the motion regions, are detected.

We define a set G of attributes related to the scene and the involved objects in activities of interest. G consists of NG subsets of attributes that are exclusively related to certain image level features. Since we work on VIRAT dataset with individual person activities and person-object interactions, we use NG = 6 subsets of attributes for the development of intra-activity context features in the experiments as shown in **Figure 4**



**Figure 4.**

For a given activity, the above attributes are found from image-level detection. For frame n of activity we obtain gi(n) = I(Gi(n)), where I(.) is an indicator function. gi(n) is then normalized so that its element sum to 1.



**Figure 5:** The image shows one frame of person unloading an object from a vehicle, and the person is in the rear of the vehicle. So for this frame, g1(n) = [1 0 0] and g2(n) = [0 1 0], where n is the frame number of this image in activity.



Let . whereNf is the total number of frames associated with the activity. The concatenation of all g over Ng will be the intra-activity context feature vector of the activity.

*Inter-activity context feature descriptor:* Features that capture the special and temporal relationships of an activity relatively are defined as inter-activity context feature.

### Multi – SVM Classifier

By using those STIP and SIFT features we train a multi-SVM [8] classifier upon the detected action segments to generate the normalized confidence scores si,0, …, si,M of classifying the action segment I as activity classes 0, 1, … , M. Such that ∑j=0Msi,j = 1.

Support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Given some training data \mathcal{D}, a set of n points of the form

\mathcal{D} = \left\{ (\mathbf{x}_i, y_i)\mid\mathbf{x}_i \in \mathbb{R}^p,\, y_i \in \{-1,1\}\right\}_{i=1}^n

where the y_i is either 1 or −1, indicating the class to which the point \mathbf{x}_i  belongs. Each  \mathbf{x}_i is a p-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having y_i=1from those having y_i=-1. Any hyperplane can be written as the set of points \mathbf{x}satisfying

\mathbf{w}\cdot\mathbf{x} - b=0,\,

where\cdot denotes the dot product and {\mathbf{w}} the (not necessarily normalized) normal vector to the hyperplane. The parameter \tfrac{b}{\|\mathbf{w}\|}determines the offset of the hyperplane from the origin along the normal vector {\mathbf{w}}.

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

\mathbf{w}\cdot\mathbf{x} - b=1\,

And

\mathbf{w}\cdot\mathbf{x} - b=-1.\,

By using geometry, we find the distance between these two hyperplanes is\tfrac{2}{\|\mathbf{w}\|}, so we want

to minimize \|\mathbf{w}\|. As we also have to prevent data points from falling into the margin, we add the following constraint: for each ieither

\mathbf{w}\cdot\mathbf{x}_i - b \ge 1\qquad\text{ for }\mathbf{x}_i of the first class

or

\mathbf{w}\cdot\mathbf{x}_i - b \le -1\qquad\text{ for }\mathbf{x}_i of the second.

This can be rewritten as:

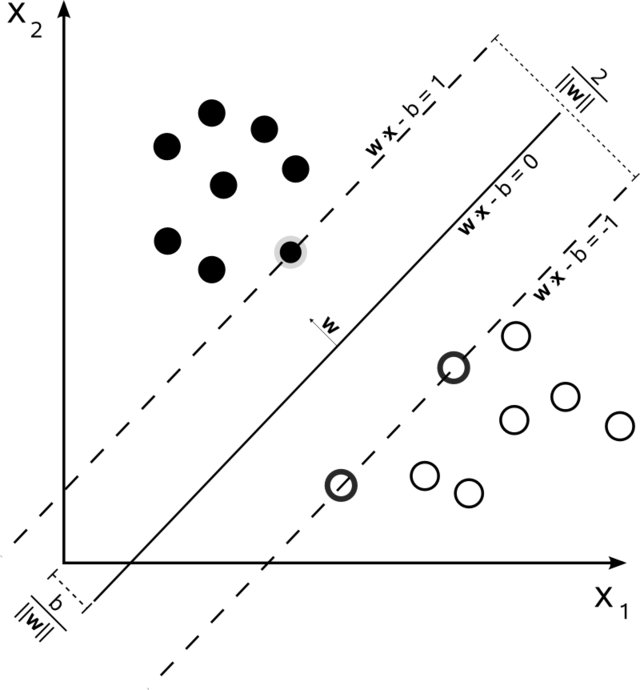
y_i(\mathbf{w}\cdot\mathbf{x}_i - b) \ge 1, \quad \text{ for all } 1 \le i \le n.\qquad\qquad(1)

We can put this together to get the optimization problem:

Minimize \|\mathbf{w}\|

subject to (for any i = 1, \dots, n)

y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \ge 1. \, 



**Figure 6.**Showing the decision boundary.

We will have M classes of activities and the multi Support Vector Machine SVM classifier will build a model that will build by the training dataset. The model will be tested and its accuracy will be computed.

This method is also called winner-take-all classification. Suppose the dataset is to be classified into M classes. Therefore, M binary SVM classifiers may be created where each classifier is trained to distinguish one class from the remaining M-1 classes. For example, class one binary classifier is designed to discriminate between class one data vectors and the data vectors of the remaining classes. Other SVM classifiers are constructed in the same manner. During the testing or application phase, data vectors are classified by finding margin from the linear separating hyperplane. The final output is the class that corresponds to the SVM with the largest margin. However, if the outputs corresponding to two or more classes are very close to each other, those points are labeled as unclassified, and a subjective decision may have to be made by the analyst. Otherwise, a reject decision (Schölkopf and Smola, 2002) may also be applied using a threshold to decide the class label. This multiclass method has an advantage in the sense that the number of binary classifiers to construct equals the number of classes. However, there are some drawbacks. First, during the training phase,

the memory requirement is very high and amounts to at the square of the total number of training samples. This may cause problems for large training data sets and may lead to computer memory problems. Second, suppose there are M classes and each has an equal number of training samples. During the training phase, the ratio of training samples of one class to rest of the classes will be 1:(M −1). This ratio, therefore, shows that training sample sizes will be unbalanced. Because of these limitations, the one against one approach of multiclass classification has been proposed.

# Dataset

* We will be using VIRAT video surveillance dataset [11].
* In the dataset we use 5 different kinds of activities to be recognized which are:

1. Person exiting a facility
2. Person entering a facility
3. Person opening a vehicle trunk
4. Person getting into a vehicle
5. Person getting out of a vehicle

# Results

We have worked on VIRAT dataset to find the effectiveness of our methodology. We find the recognition result on the dataset which are shown in the ***Table 2***. We can see that performance of our method is comparable to the other methods. However the other methods work on the video, which contains an activity of interest with additional 10 seconds occurring before or after the target activity instance, while we have worked on continuous video.

**Precision**: Precision is the ratio TP/D, where D is the total number of detections (correct and incorrect); and TP is the number of correct detections.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2. Comparison of Precision Values for the activities as in VIRAT dataset | | | |
| Activity Class | BOW + SVM[3] | NDM + SVM[3] | SIFT + SVM |
| Getting-into-vehicle | 40.4 | 32.7 | 41.5 |
| Getting-out-of-vehicle | 42.2 | 32.1 | 43.1 |
| Opening-trunk | 47.2 | 40.6 | 39.9 |
| Exiting-a-facility | 45.3 | 44.5 | 46.3 |
| Entering-a-facility | 41.6 | 43.1 | 42.2 |

# Conclusion and Future Scope

In the project we have modeled an approach that could work with other baseline classifiers and promise to give a better accuracy than the state-of-art methods.

Extending the scope of the project we can use it as a preprocessing step in video surveillance for detection of unusual activity. The algorithm can also be used to work in real time. So it has wide future scope.

# References

1. Wiliem, Arnold, Vamsi Madasu, Wageeh Boles, and Prasad Yarlagadda. "A Context Space Model for Detecting Anomalous Behaviour in Video Surveillance." In *Information Technology: New Generations (ITNG), 2012 Ninth International Conference on*, pp. 18-24. IEEE, 2012.
2. Wang, Xiaoyang, and Qiang Ji. "A hierarchical context model for event recognition in surveillance video." In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 2561-2568. IEEE, 2014.
3. Zhu, Yingying, Nandita M. Nayak, and Amit K. Roy-Chowdhury. "Context-aware modeling and recognition of activities in video." In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2491-2498. IEEE, 2013.
4. Chaaraoui, Alexandros Andre, José Ramón Padilla-López, Francisco Javier Ferrández-Pastor, Mario Nieto-Hidalgo, and Francisco Flórez-Revuelta. "A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context." *Sensors* 14, no. 5 (2014): 8895-8925.
5. Yao, Bangpeng, and Li Fei-Fei. "Modeling mutual context of object and human pose in human-object interaction activities." In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 17-24. IEEE, 2010.
6. Marszalek, Michael, Ivan Laptev, and Cordelia Schmid. "Actions in context." In*Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929-2936. IEEE, 2009.
7. Tang, Kevin, Li Fei-Fei, and Daphne Koller. "Learning latent temporal structure for complex event detection." In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1250-1257. IEEE, 2012.
8. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: A library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)*2, no. 3 (2011): 27.
9. Lan, Tian, Yang Wang, Weilong Yang, and Greg Mori. "Beyond actions: Discriminative models for contextual group activities." In *Advances in neural information processing systems*, pp. 1216-1224. 2010.
10. Laptev, Ivan. "On space-time interest points." *International Journal of Computer Vision* 64, no. 2-3 (2005): 107-123.
11. <http://midas.kitware.com/collection/view/165> Last accessed on: September, 2015