# Automatic Action Discovery in Psychological videos

by :
Diwakar Chauhan

diwakarc@iitk.ac.in

Guide :
Prof. Amitabha Mukerjee

November 7, 2012

## Abstract :

Activity recognition is an essential part of several applications. Robotics, smart homes, monitoring are the major areas which involve activity recognition. Lot's of work has been done on computer vision based activity recognition. It involves tracking the agents and understanding the behaviour of agents and their interaction. One more aspect of activity recognition is to assign linguistic to the activities. This process is very closely related to human learning process. An infant learns the nouns in first place than verbs. The reason behind this may be that verbs therefore activities require more information and information of different kind. Naigles(1990)[4] demonstrated that while learning meaning of verbs, infants use the syntactic information. She proved that given a illegal verbs in a sentence the children use the syntax to guess the meaning of the verb.

However difficult the learning process of infants may be, they do not analyse large amount of data in order to learn the meanings of verbs or recognize the activities. They learn from the competent speakers who know relate the words to the event's objects in the environment.[3] And the children can extend their learning from one event to another.

## Introduction :

Actions can be based on the the variations in the shape and size change , poses[1] or orientation of the object. For example a person walking can be identified by motion of his hands and legs relative to his body. Similar can be said about the actions involving multiple agents. But what if the action which either don't need any shape changes to be expressed or the agent is an abstract object. These type of actions are represented either by specific motion of single object or particular type of motions and interaction between the agents. In this project we analyse the second type of actions. We use psychological videos[6]. All these videos are goal driven. Most of these animations contain only two agents which are triangles. The videos represent a particular transitive action. We take some of the videos and apply HMM to on the feature vectors extracted from the frames of video. We take frames in small consecutive groups. Then we compare the HMM of each group with group using mutual acceptance metric. And based on this comparison we create hierarchical clustering of the frames of a single video.

## Related Work :

Activity recognition has been a very much worked upon topic by researchers after 1980. Most of the work has been done on human activity recognition. E. Tapia [8] have used different censors to collect data from home and based on that, they recognise activities. Vail [9] have formulated activity recognition problem as temporal classification. They use CRF for recognising activity of robots. After that they compare the results with HMM classification.

Considering of psychological videos, Heider-Simmel Videos [2] have been major attention for noun, verb recognition and linguistic mapping. Mukerjee and Satish(2008)[7] have used unsupervised approach to cluster visual events into action classes. According to them, in some visual, the objects in the focus are more likely to be the part of the action happening in the videos. They have used merge neural gas algorithm to cluster the events. But there has been very less work done on activity recognition on psychological videos.

## Algorithm :

We extract feature vector $F = \{f_1, f_2, \cdots f_s\}$ of length $s$ from each of the frame in the video. The feature vector passed to HMM is described in the section of feature extraction. We take feature

vectors of 20 frames at a time and feed it to HMM. Below is the formulation of the HMM.

### Hidden Marcov Models :

Hidden Marcov Models dynamic Bayesian Networks. These are very powerful tool for voice, action recognition[5], parts of speech tagging etc. A HMM is modeled as -

$$\lambda = \{A, B, \pi\}$$

Where

$A$ : State transition probability distribution,
$B$ : Observation symbol probability distribution and
$\pi$ : Initial probability distribution.

Here we compute the log likelihood$P(S|\lambda)$ of the data given the HMM. We use the Mixture of gaussians HMM to compute the log likelihood.
Now we divide the feature vectors of all the frames into equal sizes of 20 taken at distance of 10 frames. And we calculate he log likelihood for each of the set.

### Hierarchical Clustering :

Once we have calculated the log likelihood, we create hierarchical clustering of the this data. The each element of data represents 20 frames in the video.
Now we define mutual acceptance of two HMM as the distance between them. Mutual acceptance is mathematically defined as -

$$dist(S_1, S_2) = \frac{|logP(S_1|\lambda_1)|}{T_1} + \frac{|logP(S_2|\lambda_2)|}{T_2}$$

Where

$T_i$ is the length of the sequence $S_i$ which is 20 here.

Based on this metric we create hierarchical clustering of the data obtained from HMM. We use in built matlab function linkage for this type of clustering. We tested for different methods of merging two clusters. Out of the 'ward' method resulted into best merging method. This is inner squared distance method

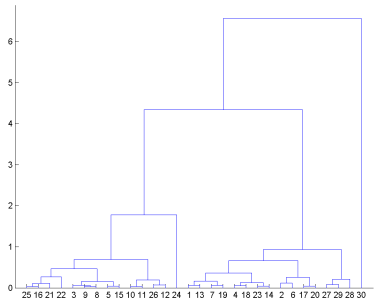


Figure 1: Dendrogram plot for coaxing video with nodes

In the hierarchical clustering the data is recursively classified into increasing number of clusters. And thus a hierarchy tree is generated. The leafs represent the individual points in the dataset and the rest of the nodes represent cluster of these points. The points merging at close to leaf nodes are more similar compared to those merging relatively near the root of the tree. So in our dataset the nodes near the leaves represent the set of frames which are more similar is some way than those

in other nodes. The root of the tree contains all the points in the data. Therefore it represents all the frames of the video. The benefit of hierarchical clustering is that to classify the data in any number or cluster, we need only to cut the tree at appropriate height.

## Preprocessing of Data and Feature Extraction :

We selected 4 of the videos namely Coaxing, Chase, Mocking, Fighting to use in our project. In the Chase and coaxing video there were three objects, a red triangle, a relatively smaller blue triangle and and a rectangular box. In the rest of the two videos the rectangular box is absent.
The frames are converted to gray scale and following steps are done :

**Determination of triangles :**

In the first step, we determine the two triangles in the video and track them. To do so one frame of the video is taken as the base frame $frame_0$ and for all the images we calculate the difference of frames :

$$difference_i = frame_i - frame_0$$

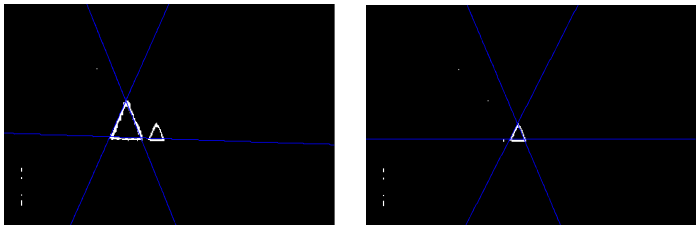

Figure 2: Two triangles as detected by Hough Transform

Then we apply erosion and dilation to remove the noise in the image. After that Hough transform is applied once on the image to detect the edges of the larger triangle and then on the remaining image to detect the edges of the smaller triangle. Once the edges are determined, centroid if each triangle is calculated.

**Determination of rectangular box :**

For the frames which contain the rectangular box, we detect the all sides by Hough Transform based on those, calculate the four vertices of the rectangle. We use midpoint approximation to determine the coordinates of the opening points of the triangles.

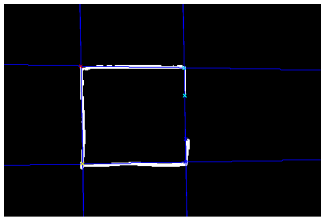

Figure 3: Vertices of the rectangle detected by Hough Transform

### Feature Extraction :

Since in all these videos an action is represented by the motion of the triangles as well as the interaction between the triangle, we need to know the absolute position and absolute orientation of both the triangles. Apart from that the videos in which the rectangular box is present, it has a lot effect on the action representation of the triangles. Therefore there should be a parameter representing presence of the box in between the triangle. We use visibility of triangles [0 1] as the measure of this factor.

**Orientation of the triangles :**

To determine the orientation of the triangle, we calculate the direction of motion of the triangle. The relative position of the vertex lying around the direction of the motion with respect to the centroid of the triangle will be the orientation of the triangle

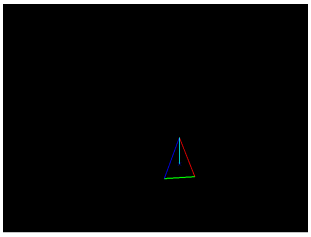

Figure 4: Orientation of triangle represented by the cyan line

**Visibility of triangles :**

This feature is added to the feature vector only if the rectangle is present in the video. The value of measure of visibility ranges from 0 to 1. Visibility of the triangles is calculated by calculating the fraction of the distance between the the supporting lines of the triangles which is hindered by the rectangle.

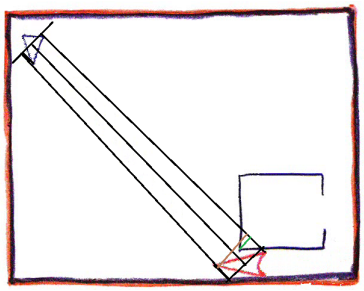

Figure 5: The ratio of length of green and brown light is the measure of visibility of the triangles

we pass the feature vector consisting of coordinates of the centroids of the triangles, their orientation vector and the visibility of the triangles in the videos where the rectangle is present. So for each frame in the videos containing rectangle, the length of the feature vector is 7 while those not containing the rectangle have feature vector of length 6. The results shown here do not use visibility feature.

## Database :

In this project we have used the animations[6] created at University of cognitive Neuroscience, UK. These animations were created to study autism in human. Autism creates Theory of Mind deficiency in children and in adults. People suffering from this are not able to predict the actions and thoughts of the other people or objects or in other way they are not able to make a theory about the reaction or interpretation of their surroundings. Traditionally autism was tested by false-belief test. False belief is an attribute in human, which makes them able to believe that other people in world can have different belief than him about same phenomena. This quality is developed in the childhood. But adults with autism were able to pass the false-belief test.

These animations named Frith-Happe animations had better success in capturing autism in adults. There are three major classes of animations. "Random" ,"Goal-Desired", and "Theory of Mind". Each class is made to test different psychological aspect of mind. Each of these class have 4 videos in it.

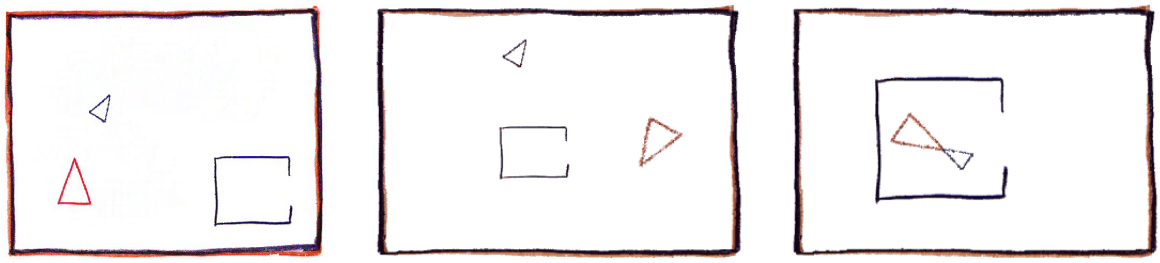

Figure 6: Stills from videos. The leftmost image is from chase video which is in Goal-Directed category, second one is from Drifting which is in random category and third one is from Coaxing video, which is in Theory of Mind Category

We used these videos because these videos represent intentionality in them. Generate such video automatically with those is a very difficult problem.

## Commentaries:

We recorded commentaries on two videos Chasing and Coaxing from two different users. The process of recording the commentaries was as follows:-

These two videos were shown to the user 2-3 times and then each user was given certain instruction about the content type of the commentaries. Then the commentaries were taken. Further these commentaries are time-stamped according to the videos. These commentaries will be used in further works on this topic.

## Results :

In the hierarchical clustering obtained, we created tool to visualize the image sets contained in each of the node. By analysing those nodes of the tree, we found that the nodes merging at the lowest level were very similar.

Below are the dendrograms plots of some videos. In each of the dendrogram only merging of 30 nodes is displayed :
First using the acceptance metric and 'ward' merging method the results were -

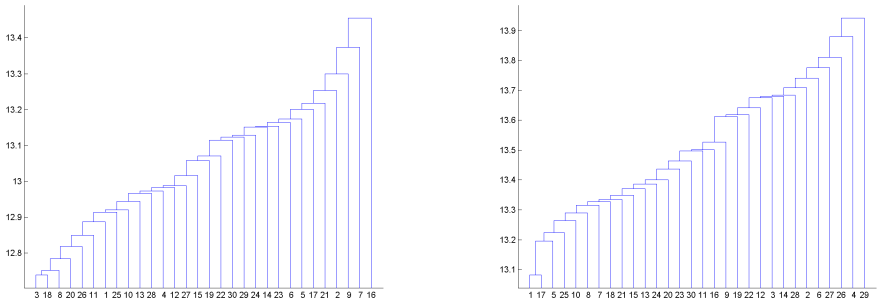

Figure 7: The dendrogram plot of the hierarchical clustering mutual acceptance and 'ward' merging

Similar results were for other metric and merging methods. The best result was found with 'ward' merging and 'euclidean' distance metric which was as follows :-

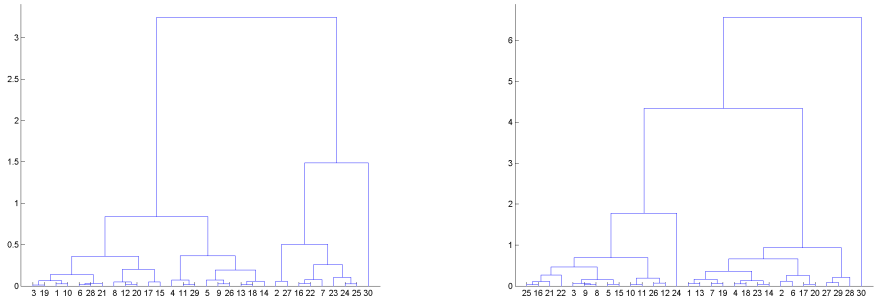


Figure 8: The dendrogram plot of the hierarchical clustering with 30 nodes visible

## Conclusion and Further Work:

By hierarchical clustering we get information about which set of frames are more similar to each other and which are less similar. If we take a few of levels from top, these nodes represent one specific kind of interaction or motion of objects. Therefore in the next step we use the commentaries on the same videos and try to find linguistics assigned to the activities represented by the triangles.

# Bibliography


[1] Jezekiel Ben-Arie, Zhiqian Wang, Purvin Pandit, and ShyamSundar Rajaram. Human activity recognition using multidimensional indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1091–1104, 2002.

[2] Fritz Heider and Marinne Simmel. An experimental study of apparant behaviour. 1940.

[3] W. Kerr, P. Cohen, and Y.H. Chang. Learning and playing in wubble world. In *Proceedings of the Fourth Artificial Intelligence for Interactive Digital Entertainment Conference (AIIDE)*, 2008.

[4] Latitia Naigles. Children use syntax to learn verb meanings. *Journal of Child Language*, 1990.

[5] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[6] Rosannagh Rogers Sarah J. White, Devorah Coniston and Uta Frith. Developing the frith-happe animations: A quick and objective test of theory of mind for adults with autism. Technical report, Instute of Cognitive Neuroscience.

[7] G. Satish and A. Mukerjee. Acquiring linguistic argument structure from multimodal input using attentive focus. In *7th IEEE International Conference on Development and Learning (ICDL 2008)*, pages 43–48, 2008.

[8] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors. In *Pervasive*, pages 158–175, 2004.

[9] Douglas L. Vail, Manuela M. Veloso, and John D. Lafferty. Conditional random fields for activity recognition. In *AAMAS*, page 235, 2007.