IIT Kanpur

B.Tech Project

Action Discovery in Psychological Videos and Hindi Verb Modelling

Author:
Diwakar Chauhan

Supervisor:
Prof. Amit Mukerjee

April 19, 2013

Acknowledgements

I am very thankful to people who have helped me with this project directly or indirectly. Many-many thanks to Dr. Uta Frith, Institute of Cognitive Neuroscience,University College London, and her Ph.D student Sarah White for providing me with psychological videos.

Contents

Abstract

*Human learning process involves learning the identity of objects(nouns) the relationship between the object(preposition) and the interaction of objects. An infant learns the identities of objects in first place than verbs. The reason behind this may be that verbs therefore activities, require more information. Naigles(1990)[**?**] 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.[**?**] And the children can extend their learning from one event to another. Here we are trying to understand the leaning process of infant and based on that learning process, try to discover actions in psychological videos and map Hindi verbs to them.*

Chapter 1

Introduction

Actions can be based on the the variations in the shape and size change , poses[**?**] 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[**?**].

Most of these animations contain only two agents which are triangles. The videos represent a particular transitive action.The videos,each of them represents certain action either complex or simple. By complex action we mean that the action constitutes smaller actions, e.g. the "Coxing" in the video consists of many instances of "Pushing" and "Rotating Jointly". 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. We create HMM for each of this groups and evaluate the mutual acceptance measure for each of the groups. Based on this measure, hierarchical clustering is created. This cluster tree is later cut at some point to produce certain number of clusters. At the same time commentary is taken for the video. This commentary is processed to remove less important words and association measures are calculated for each of the remaining words. Based on this association measure, the nouns are identified.

Unnumbered paragraph heading The text of this paragraph.

Chapter 2

Prelinguistic Image Schemata

Before mapping the language to visual sequences, we need to get the properties of the visual sequences.

2.1 Agents in the visual sequences

A sentence in a language consists of small units of words, phrases appearing as different parts of speech for that language. Theses different words are mapped to different properties of visual sequences. For example the nouns are used in context of the shapes occurring in the video. The prepositions are used for the relative positions of objects or similar things. The verbs are mapped to the action sequences. Action sequences are expressed by involvement of one or more agents.

In this project we use Frith-Happe [**?**] animations. In these videos the objects are red triangle(also big triangle), blue triangle(also small triangle) and a rectangular box. The rectangle is static at some position in the video. The two triangles move to represent actions.

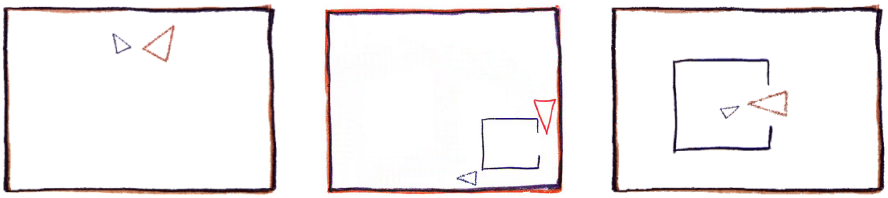

Figure 2.1: Stills from animations (a) Dancing Video, (b) Chase Video and (c) Coaxing Video

2.2 Extracting Objects:

We selected 4 of the videos namely Coaxing and Chase to use in our project. In these videos there were three objects, a red triangle, a 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 :

In the first step, we determine the two triangles in the video and track them by background subtraction. 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 :

$$difference_i = frame_i - frame_0$$

Then we apply erosion and dilation to remove the noise in the image. At this step the triangles are not separated from each other but they are separated from rest of the static objects in the video. We

use various image processing techniques to separate the triangles which are described below. The triangles are made of lines segments. Therefore the most obvious way of determining a triangle is to determine it's edges. Hough transform is a very effective way of identifying the line segments in a data.

2.2.1 Separation of the triangles

In the video the sizes of both the triangles are significantly different. We take the benefit of f this difference. The line segments corresponding to the bigger triangle will have high intensity in the (R, Θ) space compared to that of the smaller triangle.

We take top three peaks in the (R, Θ) space and get the points corresponding to them in (X, Y) space. These points form the bigger triangle. Once the line of a single triangle has been extracted, we can easily get its vertices by calculating the intersection points of the lines.

After separating the larger triangle, we apply same Hough Transform technique in the remaining points to get the edges of the smaller triangle and vertices subsequently.

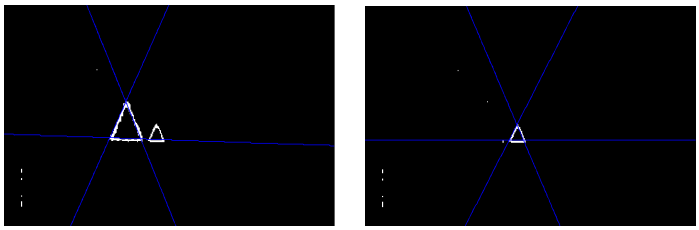

Figure 2.2: Two triangles as detected by Hough Transform

2.2.2 Determination of the Rectangle

Since the rectangle is static in the video. We can determine it in just one frame. After the triangles are identified and separated from the frame, we remove the white background and the black-red boundary. Then we are left with the points corresponding rectangle. On these points, applying hough transform, the four edges of the rectangle are determined. The four vertices of the rectangle are determined by intersection of the edges. And the vertices of the opening in the rectangle is determined by linear interpolation.

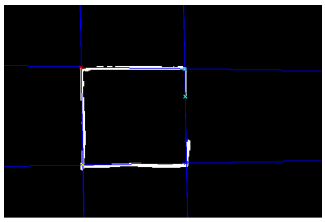

Figure 2.3: The detected rectangle and its vertices in Coaxing Video

2.3 Feature Extraction

Feature extraction is one of the most important parts of this project. The actions in a image sequence are represented bye the motion of the objects in the video. Therefore the motion of the objects in the video is highly dependent on the other object. To discover actions in the video, we need to get the features which are relevant to their interaction and their individual actions.

Let the feature vector of a frame is $F = \{f_1, f_2, \cdots, f_n\}$. Here are the features which we calculated and used different subsets of these in further works :-

2.3.1 Position of the Triangles :

The interaction between the objects is highly dependent on the position of the objects. The individual position of the objects also infer their relative positions.

In object identification, we calculated the coordinates of the vertices of the triangles. From them we calculate the coordinates of the centroid of each of the triangle which is also assumed to be the position of the triangle.

2.3.2 Orientation of the Triangles :

The extent of interaction between the triangles is dependent on their relative orientation, e.g. if both the triangles are facing each other(exactly opposite orientation), then they are more likely to interact or express an action.

The orientation of the triangle is calculated by tracking the motion of the triangle. The vertex relatively aligned in the direction of the motion with respect to the centroid of the triangle is the head of the triangle. The made by the line joining the centroid with the head of the triangle is the orientation of the triangle.

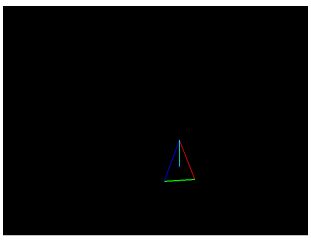

Figure 2.4: The detected rectangle and its vertices in Coaxing Video

2.3.3 Visibility of the triangle :

In the videos where the rectangular box is there, the fact that whether the two triangles are able to see each other or not and if they can then to what extent. Therefore we chose visibility too as a feature vector in the triangle.

The visibility of two triangle is a non-negative number $v \in [0, 1]$. Here 0 means that the two triangles are completely occluded with the rectangle and 1 means that they are completely visible to each other and any value between 0 and 1 is measure of extent they are visible.

The visibility of the triangles is calculated as follows :-

1− Draw the supporting tangents for the two triangles. Now part of the rectangle which falls between these two lines is determined. We take the projection of this part of rectangle in directions of perpendicular to the supported tangents is calculated. The fraction of this projection with total distance between supporting tangents is the measure of the visibility of the triangles. The figure below explains feature :-

2.3.4 Feature vector :

The feature vector of a frame is a subset or derived of the previously calculated features. Let the centroid is represented by C and orientation is represented byθ and the visibility between the triangles t_1 and t_2 is v. Then Following are some feature vector used :-

1- Ignore the visibility factor in both the triangles and take only the centroids and orientation of the triangles

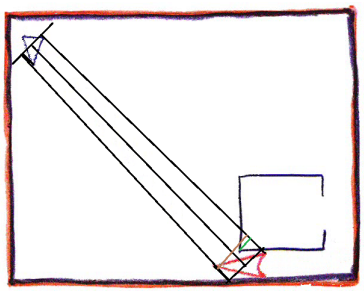

Figure 2.5: Visibility of triangles : The brown line between the two supporting tangents is the distance between the lines and the green line is measure of occlusion and the ratio of lengths of green and brown lines is the measure of visibility

$[C_{t_1}\ C_{t_2}\ \theta_1\ \theta_2]$

2- Include the visibility factor with other features

$[C_{t_1}\ C_{t_2}\ \theta_1\ \theta_2\ v]$

3- Since θ is in S_1 topology, i.e it is periodic value with period 360. It's values of 0 and 360 is overlapping. Therefore to remove this shortcoming, we can use the *sin* and *cos* of the θ.

$[C_{t_1}\ C_{t_2}\ sin\theta_1\ cos\theta_1\ sin\theta_2\ cos\theta_2]$

2.4 Actions in the Visual Sequence :

Actions in visual sequences are result of complicated motions of the objects in it. This complicated motion can be learnt by HMMs and appropriate feature vectors. We follow two approaches for learning the actions in the video. One is completely unsupervised hierarchical cluster based method and other is user labelled unsupervised method. These two methods differ on how we provide the input to the HMMs and how to use the created HMHs.

A. Completely Unsupervised Method

2.4.1 Learning the HMM :

We break the whole video into frame sequences each having N frames and each of sequences overlap by M frames. The choice of N and M is dependent on the approximate length of the action sequences and the time interval in which the action appears. For each of these sequence we learn HMM on that.

2.4.2 Hierarchical Clustering :

Once we get the HMMs on all the image sequences, we calculate the distance between these HMMs. Here we take the Mutual Acceptance as the distance between to HMM. Mutual acceptance is defined as :-

$$dist(S_1, S_2) = \frac{|logP(S_2|\lambda_1)|}{N_1} + \frac{|logP(S_1|\lambda_2)|}{N_2}$$

Where :

λ_1 and λ_2 are HMMs trained on data S_1 and S_2 respectively.
The length of S_1 is N_1 and that of S_2 is N_2.

Based on this measure a hierarchical clustering is created. We used in-built function of Matlab for such clustering. A hierarchical clustering is recursively merging of clusters or points to form a single cluster in the end. Thus we get tree in which different clusters are merged at different levels based on the distance measure provided and a merging method. We used Mutual acceptance as calculated above as the distance measure and 'ward' method for merging clusters.

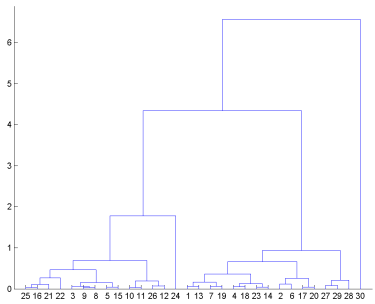

Figure 2.6: The hierarchical clustering created for Coaxing video, top 30 nodes are shown

The leafs in the tree represent the individual points in the dataset while the internal nodes represent the clusters of points. The points merging near the leaves are more similar to points merging far from leaves. This hierarchical clustering helps us to classify the data into specific number of clusters. We just need to cut the tree at appropriate level for this.

B. Used labelled Unsupervised Method

2.4.3 Labelling the actions :

In the Coaxing video, there are 4-5 types of actions namely घुम् , टक्कर् मारान , खेल , लर. Here we label all the actions in the video manually.

Appendix A

Hough Transform

Given a point (x, y) on a line L. If the distance of line from origin is r and the normal to the line from origin makes θ angle from the positive $x-axis$ then, the equation of the line can be written as :-

$$r = xcos\theta + ysin\theta$$

For each point (x, y) in the combined data of triangle we get the corresponding (r, θ) pairs for $\theta \in (-180, 180]$ at the intervals of θ_0. In this way (R, Θ) space is created. In this space the local maxima correspond to line segments. Higher the intensity of the maxima, larger the line segment corresponding to the point.

Appendix B

Hidden Markov Models

Hidden Markov Models(HMM) are dynamic Bayesian network. HMM are modelled as certain number of hidden states and some visible state. All of these states have probabilistic dependencies and these probabilistic dependencies are the characteristics of a particular HMM. Mathematically, these are modelled as follows :-

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

Where : A : State Transition Probabilities B : Observation Symbol Probabilities π : Initial Probability Distribution Like all dynamic Bayesian networks, the value of a state in HMM depends on values of previous k states. In this project, we take k as 1. So here the value of a state depends only on the previous state

B.1 Learning an HMM

We initialise the probabilities of HMM with selective random values i.e π. Now given the data, these probabilities are learnt by iteration. The values of these probabilities after these iterations are the characteristics of the HMM of the data.

B.2 Loglikelihood :

Given a learnt HMM λ on some data S, and given a test S_0 data we can measure the similarity of this data to the HMM. This value is represented as :-

$$logLik = log(P(S_0|\lambda))$$

This value is used to measure the distances between two HMMs.

B.3 Distances Between HMMs

Let a HMM learnt on Data S_1 of length T_1 is λ_1 and that on data S_2 of length T_2 is λ_2. The distance between the HMMs is defined as :-

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

Bibliography

Author, I. (Year). *Book Title*, Publisher; Place of publication.

Lamport, L. (1986), *LaTeX: A Document Preparation System*, Addison-Wesley; Reading, MA.

Author, I. (Year). 'Journal article title', *Journal*, **Vol**, pp.first–last.

Smith, A.D.A.C. and Wand, M.P. (2008). 'Streamlined variance calculations for semiparametric mixed models', *Statistics in Medicine*, **27**, pp.435–48.