IIT Kanpur

B.Tech Project

Language Independent Noun and Verb Acquisition from Psychological Videos

Author:
Diwakar Chauhan

Supervisor:
Prof. Amit Mukerjee

April 19, 2013

Acknowledgements

This project is completed direct or indirect help of many people. I want to thank them all for their support.

I am thankful to my project supervisor for guiding me through this project.Due to his sincere and timely guidance, I was able to do this project.

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 which were the core of the project.

I am thankful to Mr. Sushobhan Nayak for his work on language acquisition. This helped me a lot in the project.

I am also thankful to Mr. Prashant Jalan for doing syllable based noun discovery on the same video. Syllable based analysis helped to understand more about the co-relation of objects with utterances. He helped me in collecting the commentaries and transcribing to text.

Contents

Abstract

In Natural language processing and learning, the process of language acquisition can be modeled as similar to the learning process of an infant. The reason of this assumption is that, in both of the processes, there is very less or nothing prior knowledge about language or any other thing. The children across the world start learning their native language at almost same age[9]. They start from simple linguistic units and then learn complex ones later on. In all this development process, the children learning different languages are have very less age difference between them. Therefore modeling language acquisition as similar process to the child learning, gives us freedom to generalize for multiple languages. We apply this process two major language spoken in India i.e. Hindi and English. These two languages represent two very different classes of languages. Hindi is much more inflected language as compared to English and English has large set of lexemes. In Hindi there are much more variations of lexemes in order to represent different parts of speech, gender and tense. Therefore learning identification usage of a single word is a non trivial task. In such situations syllable based results may give better results compared to lexeme based. While English has larger set of words mapping to same image schema. Therefore identifying words corresponding to a image schema is also a non trivial task.

Here we try to get the mapping of words with image schemas in both of the languages.

Chapter 1

Introduction

1.1 Human Learning Process :

In the language learning process, humans start with learning identity of the objects i.e. nouns. Further they learn, the relationship between the object i.e. prepositions and interaction of objects i.e. verbs. The language mapping to the visual sequences can be formalized[3] as :

1 - Language learning objective is to map S to I.
2 - Set of points in S and I are the training set.

where

S is the language space
I is the image sequence space

Nouns are learned before the verbs because objects require very less information as compared to the actions. The reason behind this may be that verbs therefore activities, require more information. The infants have no prior knowledge about the environment. They see shapes, faces and moving objects in in front of them. At the same time they hear utterances of the people. The utterances spoken around the child are usually proper sentences. These contain diverse range of words. But there are some words which are specific to that particular object or action. Initially the kids don't know the word boundaries. So they try to map the phonetics of utterances with the visual sequences. Once they know the word boundaries, they map words to the visual sequences. So in both ways they filter out the words specific to visual sequences out of all the utterance and ignore the common or irrelevant words. Naigles(1990)[5] 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 analyze large amount of data in order to learn the meanings of verbs or recognize the activities. They learn from the competent speakers who know to relate the words to the event's objects in the environment.[4] And the children can extend their learning from one event to another.

Here we are trying to understand the learning process of infant and based on that learning process, try to discover actions in psychological videos and map Hindi and English verbs to them. The application of this process of multiple language shows that this process is independent of the language being learned.

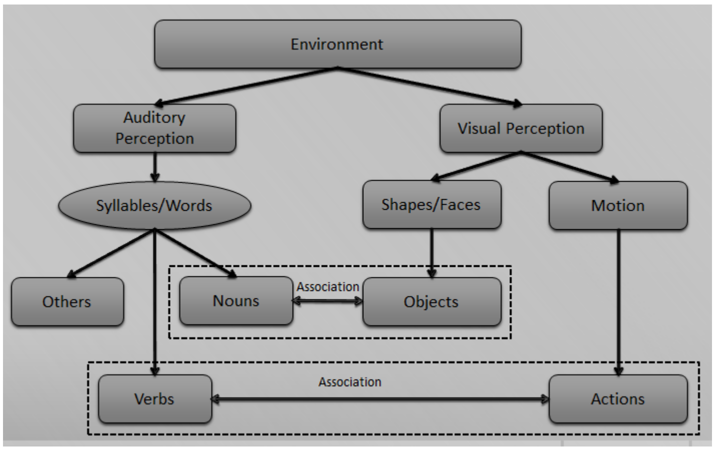


Figure 1.1: The schematic of learning process of child

1.2 Actions

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 analyze the second type of actions.

1.3 Psychological Videos :

Psychological videos are most used videos for study cognitive learning. These type of videos are manually created. The position of the objects in each of the frame of the video is result of the complex cognitive process of human mind. Each frame is designed to be a part of representation of some goal. Therefore the motion of the objects in these videos have some intent. This property of being simplistic in representation and complex in aim makes them very useful for machine learning and natural language processing purposes.
One of the very well known and widely studied psychological videos is Heider Simmel [2] video. In this video, there is a big box with a door, a small circle and two square shapes of different sizes. The actions represented in this video are **Chase**, **Go Away** and **Come Closer**.

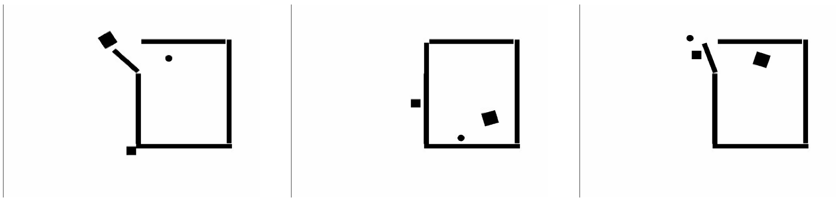


Figure 1.2: Stills from Heider Simmel video

With this video, only the above described actions can be learned. For other type of actions and more complex actions, we needed some dataset which contains variety of actions.

One appropriate choice for this purpose is animations from Frith-Happe[7]

These animations contain only two dynamic agents which are triangles. Each of the the videos represents a particular simple or complex transitive action. By complex action we mean that the action constitutes smaller actions, e.g. the ”Coxing” in the video consists of many instances of धक्का देना and घूमना. Then action discovery is done in two ways:-

1.3.1 Completely Unsupervised Action Discovery :

Here we start with zero information about the video or the actions in it. We take one videos and apply HMM 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 done and cluster tree is generated. This cluster tree is later cut at some point to produce some clusters. These clusters are action classes.

1.3.2 User labeled Action Discovery :

We label the video with actions occurring in video. Not we learn HMM on each of the user labeled action cluster. Then we merge the most similar clusters using mutual acceptance of HMM (section B.3) . On theses clusters, the associated commentaries are calculated. We calculate the relative frequency measure for words in a cluster. After that remove nouns and common words to get the verbs for the clusters.

1.3.3 Noun Discovery :

At the same time commentary is taken for the video. This commentary is processed to remove less important words. Then we take the part of commentaries which are more relevant to the triangles using attention model. And then we calculate association measures for each of the remaining words. Based on this association measure, the nouns are identified.

Below is the schematic of the whole process used in the project :-

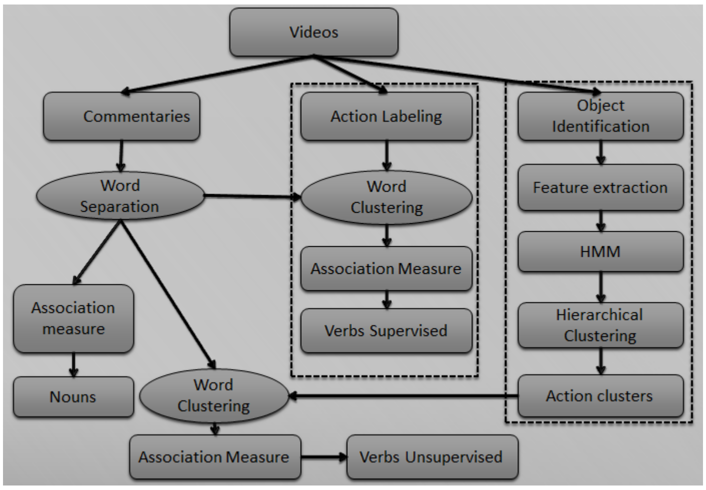


Figure 1.3: The schematic of the whole process

Chapter 2

Vision and Language Dataset

In order to recognize the actions and corresponding language labels for the language, we need proper datasets of visual sequences as well as language. The visual dataset should be something which contains various type of actions clearly represented. And similarly language database for the visual sequences should be correctly aligned to the visual sequences.

2.1 Visual Database :

The video used in this project are psychological videos. These videos were made by Dr. Uta Frith. She is a professor at Institute of Cognitive neuroscience, University of London. These videos were initially created for the purpose of identification of Autism in children and their further evaluation. There are three classes in the dataset, each containing 4 videos :

1- GD : This class contains videos with goal directed actions.The videos in this class are chase,dance, fight, lead.

2- TOM : This class contains videos testing Theory of Mind. The videos in this category are coax, mock, seduce, surprise.

3- Random : In this class the videos represent random actions likely drift, billiard, star, tennis.

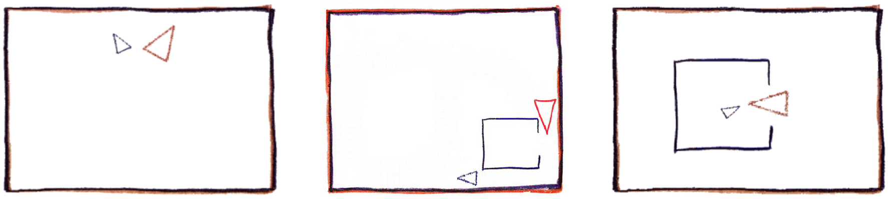


Figure 2.1: Stills from animations (a) Dancing Video, (b) Chase Video and (c) Coaxing Video

We recognised action clusters for chase and coax videos. The language modelling is done for the coaxing video due to its rich content of actions.

2.1.1 Motion Data of objects in the video :

In the videos we need to get the motion data of the objects in it. Motion data involves the frame intervals in which a particular object is moving and when it is not moving. Since in our videos only the triangles move, we need the motion data of these two triangles. This data is further used in recognition of nouns and verbs subsequently.

2.2 Language Database :

Language database is created by manually taking the commentaries for the videos and then transcribing them into text. We took commentaries in two languages primarily for one video 'Coaxing'. 22 commentaries in Hindi and 8 commentaries in English.

2.2.1 Collecting Commentaries :

Different people were asked to give commentary on the video. One subject could give at most on commentary in each of the language. Each subject was shown the video multiple times before giving the commentaries. The purpose of showing the video before commentary was to make the subject familiar with the environment. This would help him to speak simultaneously with the actions in the video at the time of giving the commentary.

Each subject was given certain instructions about the commentary. The instructions are as follows :-

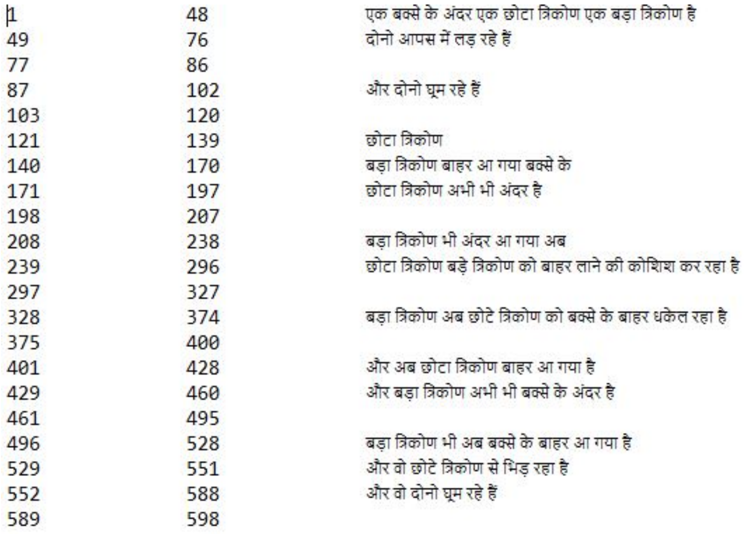
1 - Describe the objects in the video and their interactions.
2 - Do not involve yourself or any other external object as agent in the video.
3 - Do not metaphorize the sentences, explain them as they appear.
4 - Try to speak simultaneously

The subjects showed huge variety in their description of the video. Majority of the subjects considered the two triangles and the one side opened rectangular box. But some of subjects considered the rectangular shape enveloping all other objects as another object. Therefore their narration consisted of 4 agents. Most of the subjects were stuck on their initial notion of the objects in the video, e.g. If a subject refers the larger triangle(also the red one) as बडा त्रिभुज् then he continues to refer it as this word only in the whole commentary and if he refers it as लाल त्रिभुज् then he does it in whole video. Very a few subjects mixed these words.

After recording the commentaries, the audio files were embedded the the video. For each of the subject and language, a separate video was created with equal frame rate with the original animation video.

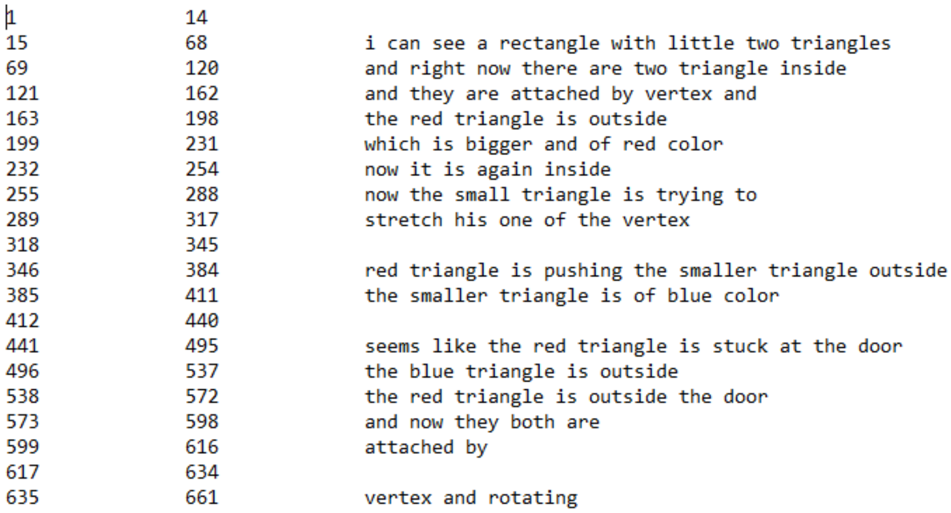
2.2.2 Transcribing to text :

We did manual transcription of the commentaries. We followed a common procedure for transcribing commentaries of both the languages. The minimal unit of transcription was words. The words were separated from each other by spaces. The morphological variations of the words were kept as they were spoken. The small grammatical errors were corrected. Now the utterances were divided into sentences or phrases. The basis of division was the length of sentences or phrases and the pauses made by the subject. If a sentence was larger than some predefined length and could be broken into two parts, we wrote it in two parts. Or if the subjects pauses for more than 5 frames before uttering next words, we consider it in next line. The pauses of less than 5 frames were bound to accuracy of measurement.



1	48	एक बक्से के अंदर एक छोटा त्रिकोण एक बड़ा त्रिकोण है
49	76	दोनो आपस में लड़ रहे हैं
77	86	
87	102	और दोनो घूम रहे हैं
103	120	
121	139	छोटा त्रिकोण
140	170	बड़ा त्रिकोण बाहर आ गया बक्से के
171	197	छोटा त्रिकोण अभी भी अंदर है
198	207	
208	238	बड़ा त्रिकोण भी अंदर आ गया अब
239	296	छोटा त्रिकोण बड़े त्रिकोण को बाहर लाने की कोशिश कर रहा है
297	327	
328	374	बड़ा त्रिकोण अब छोटे त्रिकोण को बक्से के बाहर धकेल रहा है
375	400	
401	428	और अब छोटा त्रिकोण बाहर आ गया है
429	460	और बड़ा त्रिकोण अभी भी बक्से के अंदर है
461	495	
496	528	बड़ा त्रिकोण भी अब बक्से के बाहर आ गया है
529	551	और वो छोटे त्रिकोण से भिड़ रहा है
552	588	और वो दोनो घूम रहे हैं
589	598	

Figure 2.2: Transcribed text for Hindi commentary



1	14	
15	68	i can see a rectangle with little two triangles
69	120	and right now there are two triangle inside
121	162	and they are attached by vertex and
163	198	the red triangle is outside
199	231	which is bigger and of red color
232	254	now it is again inside
255	288	now the small triangle is trying to
289	317	stretch his one of the vertex
318	345	
346	384	red triangle is pushing the smaller triangle outside
385	411	the smaller triangle is of blue color
412	440	
441	495	seems like the red triangle is stuck at the door
496	537	the blue triangle is outside
538	572	the red triangle is outside the door
573	598	and now they both are
599	616	attached by
617	634	
635	661	vertex and rotating

Figure 2.3: Transcribed text for one English commentary

Chapter 3

Prelinguistic Image Schemata

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

3.1 Agents in the psychological videos

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 image schema. The image schema can be considered as combination of two things. First is the objects participating in the image schema. Second is the feature vectors of the set of frames constructing the image schema. The function associating the feature vectors of the frames is the characteristic of the given image schema. Here we consider mainly two type of such schema i.e. for nouns and verbs. The nouns are used for the shapes occurring in the video. The prepositions are used for the relation objects . The verbs are used for interaction of the objects in the video and their motions.

To get such image schema, we needed data which has identifiable objects and their actions. So we use Frith-Happe [7] animations. In these videos the objects are red triangle([**RT**], also the bigger triangle), blue triangle([**BT**] also the smaller triangle) and a rectangular box [**RB**]. The rectangle is static at specific position in all the frames. The two triangles move inside a frame. Through their motion, they represent various actions.

Subject	Interval	Commentary
Subject1	51-89	एक छोटा त्रिभुज और एक बड़ा त्रिभुज आपस में खेल रहे हैं
Subject2	60-83	एक दुसरे को पकड़ कर घूम रहे हैं
Subject3	58-128	लाल त्रिभुज और एक नीला त्रिभुज दोनो एक दुसरे का सिरा पकड़ कर घूम रहे हैं

Table 3.1: The commentaries from the different subject on the at the time of the frames show above. The time interval of these commentaries completely contain the interval of the frames shown

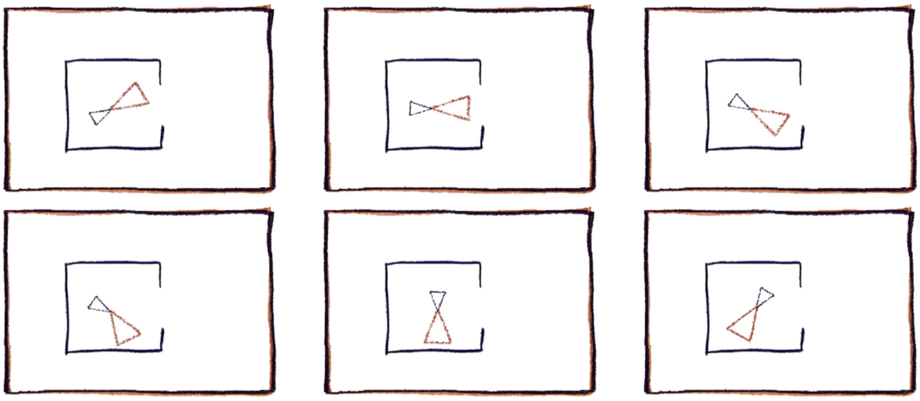


Figure 3.1: The images representing the states of the triangles from Frame 62 to Frame 78 in the coaxing video. This image schema correspond to circling as spoken by most of subjects.

3.2 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 fr_0 and for each of frames fr_i, we calculate the difference of frames d_i :

$$d_i = fr_i - fr_0$$

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.

3.2.1 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.

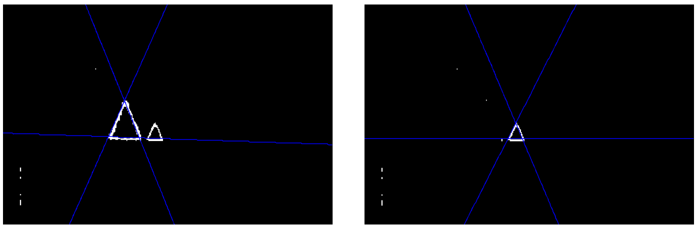


Figure 3.2: Two triangles as detected by Hough Transform

3.2.2 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.

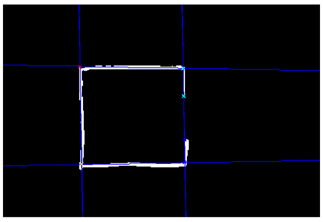


Figure 3.3: The detected rectangle and its vertices in Coaxing Video

3.3 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 :-

3.3.1 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.

3.3.2 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.

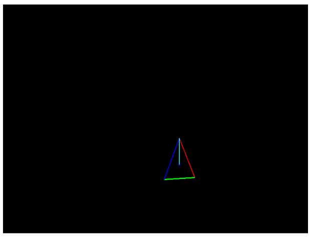


Figure 3.4: The detected rectangle and its vertices in Coaxing Video

3.3.3 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 :-

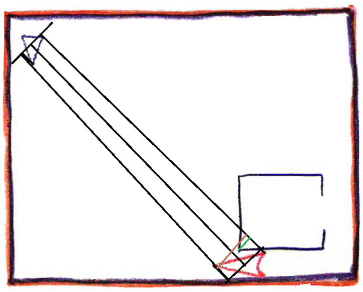


Figure 3.5: Visibility of triangles : The brown line between the two supporting tangents is the distance between the lines and the green line is measure of occlusion and the ratio of lengths of green and brown lines is the measure of visibility

3.3.4 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

$[\mathbf{C_{t_1}\ C_{t_2}\ \Theta_1\ \Theta_2}]$

2- Include the visibility factor with other features

$[\mathbf{C_{t_1}\ C_{t_2}\ \Theta_1\ \Theta_2\ v}]$

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.

$[\mathbf{C_{t_1}\ C_{t_2}\ sin\Theta_1\ cos\Theta_1\ sin\Theta_2\ cos\Theta_2}]$

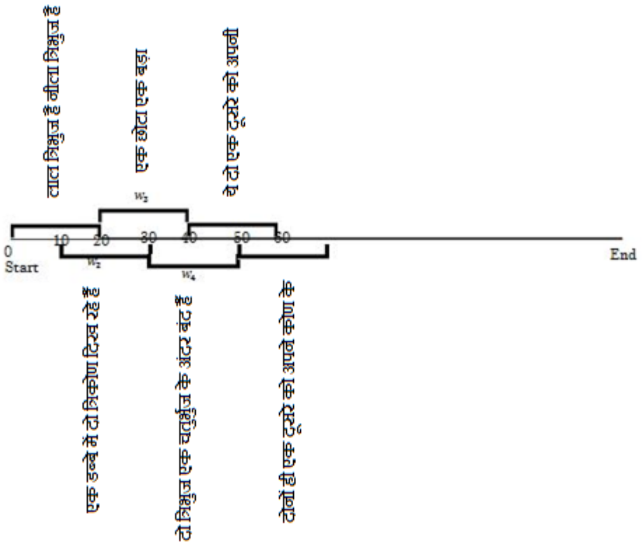
3.4 Actions in the Visual Sequence :

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 labeled 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

3.4.1 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.


लाल त्रिभुज है नीला त्रिभुज है
एक छोटा एक बड़ा
वे दो एक दूसरे को अपनी
w_3
0
Start
10
20
30
40
50
60
End
w_2
w_4
एक डब्बे में दो त्रिकोण दिख रहे हैं
दो त्रिभुज एक चतुर्भुज के अंदर बंद हैं
दोनों ही एक दूसरे को अपने कोण के

3.4.2 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.

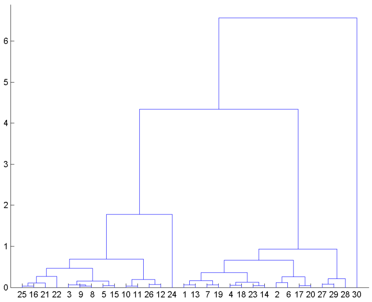


Figure 3.6: The hierarchical clustering created for Coaxing video, top 30 nodes are shown

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.

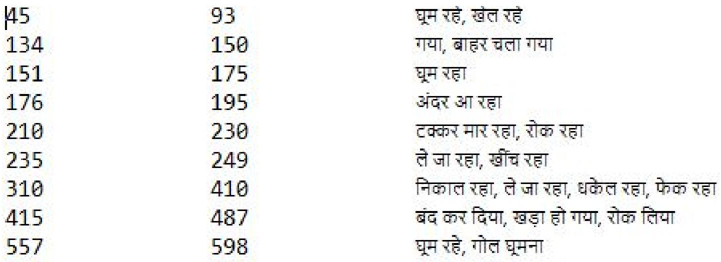
B. User labeled Unsupervised Method

3.4.3 Labelling the actions :

In the Coaxing video, there are 4-5 types of actions namely घूम , टक्कर मारना, खेल, लड. Here we label all the actions and their occurrence interval in the video manually. Multiple intervals can have same actions.

3.4.4 Learning HMM on labeled actions and Merging labels:

In the next step we learn HMM on each of the labels. These HMMs are characteristics of the action described by that label. Now based on Mutual acceptance distance we merge



45	93	घूम रहे, खेल रहे
134	150	गया, बाहर चला गया
151	175	घूम रहा
176	195	अंदर आ रहा
210	230	टक्कर मार रहा, रोक रहा
235	249	ले जा रहा, खींच रहा
310	410	निकाल रहा, ले जा रहा, धकेल रहा, फेक रहा
415	487	बंद कर दिया, खड़ा हो गया, रोक लिया
557	598	घूम रहे, गोल घूमना

Figure 3.7: User labeled actions in the Coaxing Video

the nearest HMMs. After merging we will be left with set of intervals each representing a single type of action.

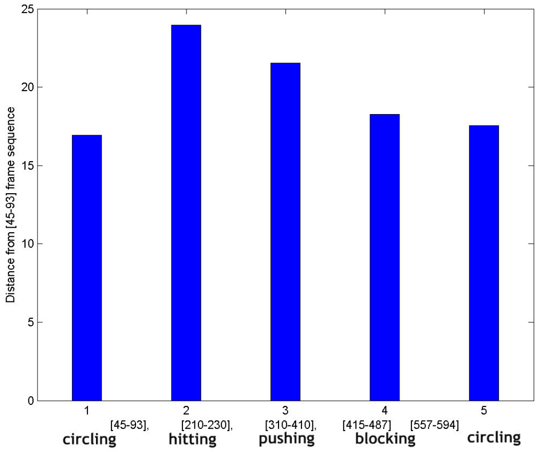


Figure 3.8: HMM distances of first labeled action sequence with all other major action sequences

Chapter 4

Language Association

Till now we are able to separate different objects in the video. Now we need to associate these objects with words. In the similar way by clustering we merge similar action and separate actions with different attributes. But we don't know how are these actions associated with the language.

4.1 Properties of Language :

A language has four properties :-

lexicons : Lexicons of a language are all the words consisting of the language. This includes all words in language and corresponding lexemes. Any sentence in a language is a subset of lexicons. To recognize the lexemes one must be able to understand the word boundaries. A child knows the words boundary very late in it's learning process

Phonology : Phonology of a language corresponds to the pronunciations in the language. Syllables of a language represent the phonological property of the language. In early learning process of child, it doesn't know the word boundaries. So it tries to map syllables to the objects and actions.

Morphology : The lexemes can be broken as morphemes. Morphemes are the smallest fragmentation of words which have meaning. To able to recognise the lexemes in a language one need to know the meanings of words and be able to distinguish between meaningful and meaningless syllables

Syntax : Syntax is structure of the sentences in a language. Syntax can be learned only with prior knowledge of lexemes, phonology and morphology of the language.

Here we study only lexical and phonological properties of the language i.e syllables and words. And we don't assume any prior information about any aspect of the language. So our model will be true for any language.

4.2 Words Corpus :

In the commentary, most common words of the language don't refer to the object being described. Therefore we need to remove these types of words from our commentaries. Most common words are identified by a corpus. Below is the description of the corpus of the both language and corresponding filtering process :-

4.2.1 Hindi Corpus :

We use CFILT, IIT Bombay[?] corpus. We calculate the absolute frequency as well as fractional frequency of each word in the corpus. Now we use two methods of filtering words from our commentaries -

1- We calculate the fractional frequency of each word in our commentaries. Then we remove all words in the commentary whose fractional frequency matches with that in the corpus.

2- We take top 1000 words in the corpus and remove them from the commentary.

4.2.2 English Corpus :

For English, we took most common words from DuBois Learning Center[?]. There are 100 most common words available here. Then we removed these words from our commentaries.

4.3 Noun Learning :

Before discovering the verbs, we need to filter out the nouns in the visual sequence. Therefore we need to discover the nouns in the commentary. These nouns will primarily correspond to the moving objects and the objects more involved into the interactions. To identify the nouns, we need to build attention model :-

4.3.1 Attention Model for objects :

When any word in the commentary is said about some object in the video, we say that the object is attended.

The attention model says that the subject will attend to the objects which are moving. The subject will unlikely utter about the temporarily or permanently static object. he utters about the objects where his gaze is focused. And the gaze of human is more likely to follow the motions in a video[8]. In in a single image, the salient parts are mostly the more contrast parts and the rare or different object. But in image sequences, these factors affect the gaze much less than motion. Therefore the words in the commentary will be highly co-related to the objects which are moving in the video. Based on this assumption we identify the nouns.

4.3.2 Determining concepts :

Concepts play similar role in image schema as the words or syllables play in the language. The concepts are instances of image schema. Upon hearing the utterances, it is mapped to the words which is further mapped to the concepts. Similarly on seeing the image schema, the object is being recognized and words start becoming available in symbolic unit.

In this project we have considered following types of concepts for noun learning :

(i)Concept-1 : This concept occurs when the red triangle **RT** is moving and is attended by the speaker.

(ii) Concept-2 : This concept occurs when the blue triangle **BT** is moving and is attended by the speaker.

(iii) Concept-1 Not Concept-2 : This concepts happens when the red triangle is moving and is attended, but the blue triangle is not attended by the speaker. This favors **RT** strongly.

(i) Not Concept-1 Concept-2 : This concept happens when the blue triangle is moving and is attended while the red triangle is not attended by the speaker. This favors **BT** strongly.

(i)Concept-1 and Concept-2 : This happens when both triangles are moving and both are attended by the speaker. This doesn't favor any object particularly.

(i) Not Concept-1 Not Concept-2 : This concept happens when none of the objects are attended in the speaker. This also doesn't favor any specific object or other aspect of the image schema. We use first 4 concepts for extracting nouns and verbs.

4.3.3 Association Measures :

Association measures are computations on the labels and visual sequences which provide the measure of co-occurrence of given label and visual sequence.
Given a label l, a concept c at time t, following probabilities are defined :-
Probability that speaker s has attended the concept c at time t is :-

$$\begin{aligned} P(c|s,t) &= 1 \text{ if } c \text{ is attended by speaker } s \text{ at time } t \\ &= 0 \text{ otherwise} \end{aligned}$$

$$\begin{aligned} P(l|s,t) &= 1 \text{ if } l \text{ is attended by speaker } s \text{ at time } t \\ &= 0 \text{ otherwise} \end{aligned}$$

Let's assume all the speakers are S and the concepts are represented by C anl the utterances are L.
Now the joint probability of the concept c and utterance l for all the speakers is represented as :-

$$J(l,c) = \frac{1}{T * \|S\|} * \sum_{t=1}^{T} \sum_{s \in S} P(c|s,t) * P(l\|s,t)$$

Where : T is total duration of the occurrence of concept
In the same way, the concept probability is defined :

$$P(c) = \frac{1}{T * \|S\|} * \sum_{t=1}^{T} \sum_{s \in S} P(c|s,t)$$

and label probability is :

$$P(l) = \frac{f(l)}{\sum_l f(l)}$$

Where $f(l)$ is the frequency of label l in the commentary
A good association measure is one which gives very high values for the labels and visual categories which co-occur frequently. It also penalize the labels which occur frequently with different visual categories. Here are some association measures used in in this project

4.3.4 Joint Probability Results for Concepts :

Following are the results for joint probability :

Hindi Language :

Concept-1(Red Triangle)					
Monograms		Bigrams		Trigrams	
त्रिभुज	0.245	लाल त्रिभुज	0.100	वो छोटे त्रिभुज	0.018
बाहर	0.237	नीला त्रिभुज	0.054	गोल गोल घूम	0.016
लाल	0.143	छोटे त्रिभुज	0.045	बड़ा त्रिभुज बाहर	0.014
बड़ा	0.124	बाहर धकेल	0.0410	लाल त्रिभुज नीले	0.011

Table 4.1: Joint Probability for Concept-1(Red Triangle)

Concept-1 Not Concept-2(Red Triangle)					
Monograms		Bigrams		Trigrams	
त्रिभुज	0.034	बड़ा त्रिभुज	0.010	बड़ा त्रिभुज छोटे	0.004
बड़ा	0.030	लाल त्रिभुज	0.010	त्रिभुज छोटे त्रिभुज	0.004
त्रिकोण	0.029	छोटा त्रिभुज	0.009	त्रिकोण छोटे त्रिकोण	0.004
छोटा	0.124	नीला त्रिभुज	0.008	बड़ा लाल त्रिभुज	0.003

Table 4.2: Joint Probability for Concept-1 Not Concept-2(Red Triangle)

Concept-2(Blue Triangle)					
Monograms		Bigrams		Trigrams	
बाहर	0.255	लाल त्रिभुज	0.103	लाल त्रिभुज नीले	0.022
त्रिभुज	0.237	नीला त्रिभुज	0.054	त्रिभुज नीले त्रिभुज	0.022
लाल	0.143	नीले त्रिभुज	0.045	वो छोटे त्रिभुज	0.018
त्रिकोण	0.124	त्रिभुज बाहर	0.0410	लाल त्रिभुज बाहर्	0.012

Table 4.3: Joint Probability for Concept-2(Blue Triangle)

Not Concept-1 Concept-2(Blue Triangle)					
Monograms		Bigrams		Trigrams	
त्रिभुज	0.033	नीला त्रिभुज	0.024	नील त्रिभुज बाहर	0.008
बाहर	0.030	लाल त्रिभुज	0.014	नीला त्रिभुज लाल	0.006
नीला	0.024	त्रिभुज बाहर	0.011	त्रिभुज लाल त्रिभुज	0.006
लाल	0.017	त्रिभुज लाल	0.006	नीला त्रिभुज अन्दर	0.004

Table 4.4: Joint Probability for Not Concept-1 Concept-2(Blue Triangle)

English Language :

Concept-1(Red Triangle)					
Monograms		Bigrams		Trigrams	
triangle	0.165	smaller triangle	0.064	smaller triangle outside	0.029
smaller	0.127	**bigger triangle**	0.060	started playing outside	0.012
outside	0.112	small triangle	0.032	both started playing	0.012
box	0.104	triangle outside	0.029	small object follows	0.011

Table 4.5: Joint Probability for Concept-1(Red Triangle)

Concept-1 Not Concept-2(Red Triangle)					
Monograms		Bigrams		Trigrams	
triangles	0.022	basically both	0.008	blue smaller object	0.006
box	0.014	smaller object	0.006	triangle over here	0.003
triangle	0.010	**red object**	0.006	work tries move	0.00
object	0.010	blue smaller	0.006	within space bigger	0.00

Table 4.6: Joint Probability for Concept-1 Not Concept-2(Red Triangle)

Concept-2(Blue Triangle)					
Monograms		Bigrams		Trigrams	
triangle	0.204	**smaller triangle**	0.096	**smaller triangle outside**	0.028
ouside	0.156	bigger triangle	0.064	started playing outside	0.012
smaller	0.146	triangle outside	0.028	both started playing	0.012
box	0.112	red triangle	0.023	bigger triangle seems	0.012

Table 4.7: Joint Probability for Concept-2(Blue Triangle)

Not Concept-1 Concept-2(Blue Triangle)					
Monograms		Bigrams		Trigrams	
triangle	0.036	**smaller triangle**	0.014	box ring basically	0.008
outside	0.031	red triangle	0.012	big box ring	0.008
smaller	0.027	gets success	0.009	**smaller triangle tries**	0.005
box	0.016	ring basically	0.007	roam around outside	0.005

Table 4.8: Joint Probability for Not Concept-1 Concept-2(Blue Triangle)

4.3.5 Analysis of JP Results in Both Languages :

In all of the association measures, the single concept is weaker than the corresponding multiple concept. For example Concept-1 is weaker than Concept-1 Not Concept-2 due to obvious reasons of exclusion of the simultaneous occurrences of concepts. In the videos, most of the time both the triangles move simultaneously, therefore there is a lot of overlapping in their motion. Since we are considering words, the association of words with concept is quite discretized. So the accuracy of results varies with different measures. And sometimes, the results are unexplainable.

Joint probability captures the co-occurrence of the concepts and words. Following is the analysis of the result :
1- For Concept-1(**[RT]**) and Concept-1 Not Concept-2(**Strong [RT]**) we get लाल त्रिभुज

as dominating.[**0.10 against 0.054**] and [**0.02 against 0.017**]
2- For Concept-2 ([**BT**]) we get लाल त्रिभुज as dominating.[**0.103 against 0.099**]. The explanation can be given as described earlier in beginning of analysis i.e. the motions of the objects is highly overlapping and the mapping is discretized.
3- For Not Concept-1 Concept-2 **Strong** [**BT**] we get नीला त्रिभुज dominating.[**0.024 against 0.014**]
Similar analysis can be done for other English language results. For [**RT**] the association in English are not very strong. But for the [**BT**] the association measure are quite strong.[**0.096**] **against** [**0.064**]

4.3.6 Relative Frequency :

This measure is used in both the noun discovery as well as in verb discovery. For nouns this does not give good results because the nouns are uniformly distributed in the commentary. Relative frequency is calculated as follows :

$$RF(l,c) = \frac{\text{Frequncy of } l \text{ when } c \text{ is in focus}}{(\text{freq of } l) * (\text{freq of } l \text{ when } c \text{ is not in focus})}$$

Relative frequency gives high measures of words occurring in relatively high frequency. But it also gives high values for those words which occur seldom in the commentary. These words spoil the noun results with relative frequency measure. The solution for this is mutual information measure.

4.3.7 Mutual Information :

Mutual information gives weight to the occurrence of a label with respect to the all words occurring for the visual sequence. Therefore it doesn't give weight to the labels seldom occurring in the commentary. Apart from that it favors, the rare concepts which co-occur with specific label frequently. And it also penalizes the words which occur most frequently with many image schema. Mutual information is defined as :-

$$MI(l,c) = J(l,c) * log\left(\frac{J(l,c)}{P(c) * P(l)}\right)$$

Hindi Language Results :

Concept-1(Red Triangle)					
Monograms		Bigrams		Trigrams	
त्रिकोण	0.510	लाल त्रिभुज	0.120	वो छोटे त्रिभुज	0.042
बाहर	0.404	बाहर धकेल	0.117	गोल गोल घूम	0.030
बड़ा	0.335	छोटे त्रिकोण	0.101	दोनो त्रिभुज घूम	0.022
त्रिभुज	0.206	छोटे त्रिभुज	0.089	बड़ा त्रिभुज बाहर	0.020

Table 4.9: Mutual Information results for Concept-1 (Red Triangle)

Concept-1 Not Concept-2 (Red Triangle)					
Monograms		Bigrams		Trigrams	
त्रिकोण	0.158	छोटा त्रिभुज	0.025	त्रिकोण छोटे त्रिकोण	0.021
बड़ा	0.094	बड़ा त्रिभुज	0.023	बड़ा त्रिभुज छोटे	0.017
छोटा	0.065	त्रिकोण छोटे	0.021	बड़ा लाल त्रिभुज	0.014
तकरा	0.023	छोटे त्रिकोण	0.0206	छोटा नीला त्रिभुज	0.012

Table 4.10: Mutual Information results for Concept-1 Not Concpet-2(Red Triangle)

Concept-2 (Blue Triangle)					
Monograms		Bigrams		Trigrams	
बाहर	0.448	लाल त्रिभुज	0.151	वो छोटे त्रिभुज	0.041
त्रिकोण	0.430	बाहर निकल	0.099	त्रिभुज नीले त्रिभुज	0.041
छोटे	0.216	छोटे त्रिभुज	0.090	लाल त्रिभुज नीले	0.036
त्रिभुज	0.188	नीले त्रिभुज	0.0871	दोनो त्रिभिज घूम	0.021

Table 4.11: Mutual Information results for Concept-2 (Blue Triangle)

Not Concept-1 Concept-2 (Blue Triangle)					
Monograms		Bigrams		Trigrams	
नील	0.050	नीला त्रिभुज	0.055	नीला त्रिभुज बाहर	0.034
बाहर	0.037	अन्दर ना	0.018	नीला त्रिभुज ऊपर	0.012
कोस्हिस्ह	0.032	लाल त्रिभुज	0.013	त्रिभुज अन्दर ना	0.012
अन्दर	0.028	वो बाहर	0.013	नीला त्रिभुज लाल	0.012

Table 4.12: Mutual Information results for Not Concept-1 Concept-2 (Blue Triangle)

English Language Results :

Concept-1(Red Triangle)					
Monograms		Bigrams		Trigrams	
triangle	0.211	**bigger triangle**	0.094	smaller triangle outside	0.051
trying	0.205	smaller triagle	0.087	started playing outside	0.041
smaller	0.190	inner square	0.072	both started playing	0.041
both	0.180	small triangle	0.089	small object follows	0.020

Table 4.13: Mutual Information results for Concept-1 (Red Triangle)

Concept-1 Not Concept-2 (Red Triangle)					
Monograms		Bigrams		Trigrams	
triangles	0.071	basically both	0.047	blue smaller object	0.032
basically	0.039	smaller object	0.032	triangles over here	0.016
here	0.036	blue smaller	0.032	work tries move	0.001
squares	0.030	squares here	0.030	within space bigger	0.000

Table 4.14: Mutual Information results for Concept-1 Not Concpet-2(Red Triangle)

Concept-2 (Blue Triangle)					
Monograms		Bigrams		Trigrams	
triangle	0.324	**smaller triangle**	0.186	**smaller triangle outside**	0.047
outside	0.291	bigger triangle	0.107	started playing outside	0.041
smaller	0.245	red triangle	0.050	both started playing	0.041
box	0.185	**blue object**	0.0416	triangle outside again	0.028

Table 4.15: Mutual Information results for Concept-2 (Blue Triangle)

Not Concept-1 Concept-2 (Blue Triangle)					
Monograms		Bigrams		Trigrams	
outside	0.070	gets success	0.051	box ring basically	0.040
triangle	0.066	red triangle	0.051	big box ring	0.040
smaller	0.054	ring basically	0.040	**smaller triangle tries**	0.029
success	0.051	cannot enter	0.040	roam around outside	0.029

Table 4.16: Mutual Information results for Not Concept-1 Concept-2 (Blue Triangle)

4.3.8 Analysis of MI Result in Both languages :

Mutual Information captures the concepts and words occurring simultaneously and it penalizes the words which occur with many concepts. Following is the analysis of the results:

1- For Concept-1 [**RT**] we get लाल त्रिभुज as strong association.[**0.12 against 0.11**]

2- For Not Concept-1 Concept-2**Strong** [**BT**], we get नीला त्रिभुज as strong association.**0.055 against 0.013**.

Similarly in English language :

1- For concept-1 [**RT**], we get bigger triangle as stong association.[**0.094 against 0.087**]

The deviation of results from actual values is also due to morphological variations of words in Hindi.

4.4 Verb Learning

4.4.1 Completely Unsupervised Method

Here we get labels for the action clusters generated by hierarchical clustering. Since these clusters have larger distance between them as compared to the distances between their subsets, they are assumed to represent separate action. Therefore each cluster is treated as a separate concept. Now for labels in each concept we calculate the relative frequency measure. After ranking the values in decreasing order, we get that top-most values have labels mostly verbs, rare words and some of nouns. Here we don't get nouns because nouns are uniformly distributed in all the clusters.

We filter out top labels based on some threshold. Now we do more filtering by Hindi corpus and previously determined nouns. After this process, we majorly get the verbs for each cluster. In a cluster, verbs with highest measure and with synonyms of high measures are the action labels of the the cluster.

4.4.2 Verb Learning Results :

Hindi Language :

Table 4.17: Hindi language results

Intervals	Exptected	Detected Verbs
190-255	आना ,टकराना	भिड़ा, हिचक, निकालना,चिढ, स्हरमा, खींचने, **टक्करमारने**
150-195	घूमना, आना	डरता, खेलना, घूमे,**घूमते**
520-595	घूमना	वापस
370-405	धक्का	
400-455, 460-515	बंद करना	युध

English Language :

Table 4.18: English language results

Intervals	Exptected	Detected Verbs
190-255	Come, Hit	**Coming**, Rotating, Refusing, **Came**, Corner, Back, Ways, Funny, Pull, Turn, Follows
150-195	Rotate, Come	Cut,**Dancing**, Separated, Vertex, Chasing, Attached, Game, Side, Gone:0.67
520-595	Circle, Rotate	Exploring, Interacting, Fighting, Surroundings, Also, Touched, Enjoying, Over, Happy, acquainted, Forcing, **Circling, Playing:0.67**
370-405	Push	Win, Drag, Here, Ahead, **Pushing:0.75**
400-455, 460-515	Block	Force, Forcefully, Enter, Move, **Blocks**, Ring, Size, Slowly, Taunt, Explore, Roam, Push:0.67

4.4.3 User labeled Unsupervised Method

We use similar method for learning the verbs in user labeled action clusters as used for hierarchical cluster based action cluster. Each of the clustered intervals are treated as separate concepts. Now for labels each concept, we calculate the relative frequency measure. Here again we get most verbs, rare words and some nouns as top words. Then we filter words based on association measure value, corpus and previously calculated nouns. After that we get the verbs as labels for the cluster. We compare this with the user labeled values for accuracy of results.

Hindi Language:

Table 4.19: Verb results on user labeled clusters in Hindi

Intervals	Ground Truth	Detected Verbs
45-93, 151-175, 557-198	घूम रहे, खेल रहे,गोल घूम	खुस्ह, पकड़, गोल, घूम , कोण दोनो, दोस्त
134-150	गया, बहर चला गय	घूमे
176-195	अन्दर आ रहा	वापस
210-230	टक्कर मार रहा, रोक रहा	
235-410	ले जा रहा, खींच रहा	
310-410	निकाल रहा, ले जा रहा, धकेल रहा, फ़ेक रहा	निकाल्, धकेल, फ़ेक, तरीको, धक्का, सिरे, जुड, धक्के, मारकर, छोता, अधूरे, धकेलने, चतुर्भुज, नीले
415-487	बन्द कर लिया, खड़ा हो गया, रोक लिया	रास्ता, बदे, खदा, युध, दरवाजा, ना, भगा, विरुध, पाये, पड़, वो, रोक

English Language:

Table 4.20: Verb results of user labeled clusters in English

Intervals	Ground Truth	Detected Verbs
45-93, 151-175, 557-198	Circle, Rotate, Play	Fighting, Playing, Interacting, **Circling**, Touching,Connected, Tips, Enjoying, Align, Enclosed
134-150	Go Away	Chasing, Dancing
176-195	Coming	Still
210-230	Hit, Stop	Funny, Ways, Corner
235-410	Pull	
310-410	Throw, Push	**Pushing, Push**, Win, Wait, Moving, Ahead
415-487	Block, Stand, Roam	**Blocks, Completely, Explore, Roam,** Reason

4.4.4 Comparison of Results in both Languages with Both Methods :

Hindi :

For completely unsupervised method, we get 2 correct mapping for words in 5 clusters out of 12 detected words.
For user labeled cluster method, we get 11 correct mapping for 3 clusters out of 7 clusters and out of 20 detected words.
In the user labeled learning we were able to get the verbs for all the major clusters. We could not get words for the clusters which were so small that very less commentaries could be assigned for that.
For completely unsupervised learning due to impurity of clusters, we were getting less accuracy.

English :

Similarly in English , for completely unsupervised method, we get 5 correct mapping for 5 clusters and 7 correct word mapping out of 40 detected words. In English we get large variety of words, so we get large number of words even after filtering.

Chapter 5

Conclusion

In this project, we proposed a generalized model for identification of nouns. By this model, we can detect nouns in commentary and corresponding image schema. We further tried to discover more complex schema e.g. actions. The action discovery was also generalized for multiple languages. We followed two methods for action discovery. First is by hierarchical clustering of HMMs on different segments of image schema. We create clusters from the cluster tree resulting from hierarchical clustering. And on these clusters, we identify the verbs associating the words with these clusters and then filtering out un-necessary words. The second method involved user input in form for classifying the video into segments of importance. We then merge these clusters using HMMs to get smaller number of clusters. And use similar method applied to the previously described clusters.

5.1 Further Work :

The HMMs were not able to cluster the video into very meaningful frame sequences, therefore the verbs learnt by completely unsupervised method were not as good as that in user labeled method. We can use eye gaze data for the 'Coaxing Video' to get better accuracy in noun and verb discovery. The gaze information tells the attention point of the speaker in the video. With this information, we can know, which object is being talked about. So the notion of concepts is not only based on the attention model, but is based on real data. This increases the accuracy of the results.

At these stage , we have nouns for the shapes in the video and verbs for the action in the video. Since we are considering mostly transitive verbs, with this much of information, we can get the predicate of actions i.e. agents involved in the actions. In that way, the syntax of a language is learnt.

Based on the information of noun and verbs, we can discover the anaphora[6] in the commentary of a language.

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

[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] Frederic Kaplan, Pierre-Yves Oudeyer, and Benjamin Bergen. Computational models in the debate over language learnability. *Infant and Child Development*, 17(1):55–80, 2008.

[4] 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.

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

[6] K. Neema and A. Mukerjee. Discovering the concept of anaphora from grounded verb models. In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 305–310. IEEE.

[7] 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 149-154, Instute of Cognitive Neuroscience, 2011.

[8] 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.

[9] Mutsumi Imai Etsuko Haryu Hiroyuki Okada Li Lianjing Jun Shigematsu. Revisiting the noun-verb debate: A cross-linguistic comparison of novel noun and verb learning in english-, japanese-, and chinese-speaking children.