

Grounded Label Learning in Telugu  
from Multimodal Input

CS499 Project Report

Teegala Sudhamsh Goutham  
Y6501  
Email: *sudhamsh@iitk.ac.in*

Advisor: Prof. Amitabha Mukerjee

April 25, 2010

### **Abstract**

This work discusses a method of learning grounded labels in Telugu language using multimodal input. The focus is on learning labels for actions. Semantic information obtained from a video is used to learn schemas for binary actions (involving two objects) like *chase*, *come closer* and *move away*. A Computational Model of Dynamic Visual Attention is used to find the most salient object in a frame. The feature vectors are computed using relative position and relative velocity of salient objects for a set of frames. The frames are then clustered using an unsupervised clustering method called the *Merge Neural gas Algorithm*. The clusters are then identified as action schemas for different verbs. Audio commentaries collected from people, while showing them the video, are typed in a word separated form. By associating the words to the frames during which they were uttered, the labels for these clusters are learnt. The labels are grounded as we associate them to action schemas derived from a video.

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Problem Structure</b>	<b>2</b>
<b>3</b>	<b>Methodology</b>	<b>2</b>
3.1	Input: Heider and Simmel Video . . . . .	3
3.2	Salient Objects . . . . .	3
3.2.1	Perceptual Theory of Mind . . . . .	3
3.2.2	Computational Model of Dynamic Visual Attention . . . . .	3
3.3	Merge Neural Gas Algorithm . . . . .	4
<b>4</b>	<b>Work Done</b>	<b>4</b>
4.1	Feature Vector Extraction . . . . .	4
4.2	Clustering . . . . .	4
4.3	Commentaries . . . . .	5
<b>5</b>	<b>Association Measure</b>	<b>5</b>
<b>6</b>	<b>Results</b>	<b>6</b>
<b>7</b>	<b>Conclusion</b>	<b>7</b>

# List of Figures

1	Frames from Heider and Simmel Video . . . . .	3
2	Frame with salient object identified . . . . .	3
3	Cluster Accuracy . . . . .	5
4	Sample Telugu Commentaries . . . . .	6
5	Sample Telugu Commentaries after Editing . . . . .	6
6	Results . . . . .	7

# 1 Introduction

Language is more than a mere set of well formed syntax on a lexicon of words. Language has meaning. It is due to its meaning that we can use language to perceive, describe and interact with the world. These activities result in grounding language. By grounding, we mean an association between a symbol in language to a pattern, a sensorimotor perception or other extra-linguistic elements.

Consider the following example. Imagine that we are in a foreign country whose language we do not know. We have a dictionary of that language with us. We observe a sign post printed in the foreign language and try to read it. We look up the first word on the sign post in the dictionary and find its definition. Since the definition is also in foreign language, we look up the first word in the definition and so on. So, no matter how many words we have in the foreign language, we will never be able to tell what is on the sign post. This is called the Symbol Grounding Problem.<sup>1</sup>

From the above example we see that grounding the symbols is important for a language. In the above example, at least some of the symbols need to be grounded. Hence, we propose a grounded method to learn labels in Telugu language. Noun grounding is relatively easy and much work has been done in the past[2][3]. We focus on learning labels for actions by a model used by [2]. We acquire concepts from a video and learn labels for them using commentaries by different people. Here we focus only on binary object interactions i.e., actions involving two objects only.

## 2 Problem Structure

We identify the following major parts of the problem.

1. Identifying the most salient object(s) in the frame. We use a computational model of dynamic visual attention to identify the most salient object(s) in a frame.
2. Computing Feature Vectors. We compute feature vectors using the relative positions and relative velocities of salient objects over a period of time.
3. Clustering the feature vectors using Merge-Neural Gas Algorithm. We obtain the clusters that represent an action schema. We then correlate these clusters to action schemas of chase, move-away or come-closer using human labels.
4. Collecting audio commentaries for the video. The commentaries are then typed in and properly separated into sentences and words and the corresponding frames are noted.
5. Language Label Learning by matching the words of a frame or frame sets to the clusters.

## 3 Methodology

Since activity learning is hard on 3D videos[3], we attended only to 2D videos. We use a 2D video of a chase sequence as our input.<sup>2</sup> We then use a bottom-up computational model of dynamic visual attention called Gaze Predictor Model[4]

---

<sup>1</sup>Adapted from Harnad, 1990 by [1]

<sup>2</sup>The particular video was developed by Bridgitte Martin Hard of the Space, Time, and Action Research group at Stanford University.

to find the salient objects in the video. The algorithms and techniques used are described below.

### 3.1 Input: Heider and Simmel Video

We use a 2D video derived from a psychology work by (Heider and Simmel,1944), modified by (Martin and Twersky,2003). In this video (consisting of 2532 frames), there are three agents, a big square, a small square and a circle which interact in and around a square room or box with a door. The interactions include chasing each other, moving in and out of the box, hitting each other and coming close to each other. The binary actions we focus on include chase, come-closer and move-away.

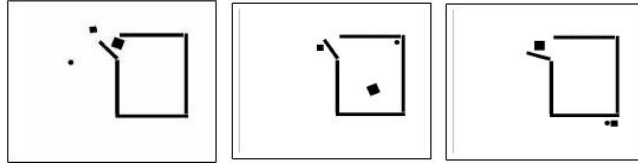


Figure 1: These are a few frames from the input video used. Three objects viz., a big square, a small square and a circle interact with each other.

## 3.2 Salient Objects

### 3.2.1 Perceptual Theory of Mind

In our experiments, we neither have information about the object or event the speaker is attending to nor do we instruct the speaker to attend to any particular object. We also do not follow the cues from the speaker’s gaze direction. We focus on the model of *Perceptual Theory of Mind*[5]. In this model, we assume that the speaker attends to the same object or event the learner finds salient. We assume from this model that the objects or events attended by the agent were also salient for the speaker when he narrates the scene.

### 3.2.2 Computational Model of Dynamic Visual Attention

We use a computational model of visual attention to find the most salient object in a frame. These are of two types. The top-down processes are dependent on task and the bottom up process are task independent. We use a bottom-up process. For static images, colour, intensity etc., are used to find the most salient object. This was extended for dynamic scenes by incorporating motion features in saliency computation[4]. We use this model for dynamic visual attention and find the most salient object or objects in a frame.

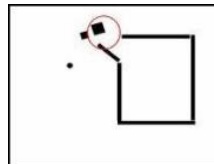


Figure 2: A frame from the Video after the dynamic visual attention model is applied. The most salient object is shown in the red circle.

### 3.3 Merge Neural Gas Algorithm

A neural gas algorithm learns topological structures in given input signals in an unsupervised manner. The algorithm is related to the Kohonen Maps or Self-Organising maps. It takes a distribution of high-dimensional data and returns a network in lower dimensions with similar topology. The algorithm involves a set of neurons with feature vectors of the same dimension as the input feature data. In each iteration of the algorithm, an input signal is selected and its distance from all neurons is computed. The closet neuron is declared as winner and feature vectors of other neurons are adapted to the feature vector of winner neuron. After some iterations, clusters begin to emerge which represent the topology of input data.

For input feature vectors containing temporal data, the time history information is not properly used by usual neural gas algorithms. Merge Neural Gas algorithm[6] adds an explicit context vector to the neural gas algorithm which captures the temporal context of the input data. A vector, *Context*, is defined for both signals and neurons. In each iteration  $i$  of the algorithm on the input signal data, Context ( $c_{s_i}$ ) of the current signal ( $s_i$ ) is defined as a combination of the feature vector ( $f_{w_{i-1}}$ ) and the context vector( $c_{w_{i-1}}$ ) of the previous winner neuron.

$$\begin{aligned} c_{s_i} &= (1 - \beta).f_{w_{i-1}} + \beta.c_{w_{i-1}}. \\ d_{n_j} &= (1 - \alpha).dist(f_{s_i}, f_{n_j}) + \alpha.dist(c_{s_i}, c_{n_j}) \end{aligned}$$

The parameters  $\alpha$  and  $\beta$  are the influence of context distance on distance function and context update rate respectively.  $\alpha$  determines the weightage given to temporal data over present data. We used  $\beta = 0.55$  and varied  $\alpha$  between 0.02 to 0.6. Clustering is done in another iteration of the algorithm where the winning neurons are identified for each pattern and the signals having winner neurons of same cluster are grouped into one cluster.

## 4 Work Done

### 4.1 Feature Vector Extraction

Every Object in the input video is identified and their positions in each frame are calculated. We define velocity of an object to be the difference of its positions between current frame and previous frame. We consider a period of 20 frames around each frame. All objects salient in that period of 20 frames are considered as being attended to at this frame. We compute the feature vectors for each pair of objects in focus. Our feature vector consists of two attributes,

1. *pos.veldiff* : the inner product of position difference and velocity difference.
2. *pos.velsum* : the inner product of position difference and velocity sum.

In the data, due to some errors in evaluating positions of objects in some frames, information about some frames (around 400 frames) is lost and so they were not considered for clustering.

### 4.2 Clustering

We now use the Merge-Neural Gas Algorithm to cluster the feature vectors. For every pair of objects, feature vectors from frames in which these two are in focus are considered for clustering. It is important to choose different parameters correctly as they affect the influence of Context and the number of clusters. Clusters with less than 15 frames on whole are ignored. We observed that there are four prominent

clusters in each run of the algorithm.

Having obtained the clusters, we need to correlate them to the human labels. Three people were shown the video once and in the second run, they were asked to utter one of the labels, *come-closer(CC)*, *move-away(MA)* and *chase(CH)*. Now given the time when these were uttered, we can identify the label for each frame. We then related the clusters to these labels frame by frame. For the four clusters obtained Fig.3 shows the results of the validation. We observe that Cluster C1 corresponds to CC and C2 corresponds to MA. We have two clusters for CH in C3 and C4. We later observed that C3 represents chase when the focus object is *chasing*, whereas C4 represents chase when the focus object is *being chased*.

	C1	C2	C3	C4	Total	Accuracy	TCA
CC	272	1	17	8	298	91	
MA	50	159	41	1	251	63	80
CH	18	11	120	38	187	84	

Figure 3: The rows represent the number of frames identified as CC(Come Closer) or MA(Move Away) or CH(Chase). Accuracy is in %. The columns represent the frames from clusters C1, C2, C3, C4. TCA is Total Clustering Accuracy(%)

### 4.3 Commentaries

Commentaries are collected in Telugu language from 8 different people of age group 18 to 22 with enough knowledge of the language. They were told the following(in Telugu or English):

*You are going to watch a video of length around 2 minutes. You need to describe what is going on in the video. The Video contains some shapes like squares and circles. You will be shown the video twice and you can practise before describing it the third time. You can use the words Big square, Small Square, Circle, Room and Door for the shapes or any other nouns you think appropriate, but be consistent with them*

Observe that we have given away the nouns we do not focus on them. After recording the commentaries, they are typed in text form. The corresponding frame sets were also noted using the time when they spoke a particular sentence. Later on, nouns and pronouns were identified and tagged. A standard form of Telugu is used to remove the variations in dialects and other usage variations. This include conversion of verbs into transitive form, using the same kind of separation of words in all commentaries, etc. A sample Telugu commentary, its English transliteration and translation are given in Fig.4. A sample commentary with the changes made is shown in Fig.5.

## 5 Association Measure

We use a conditional probability measure to find the association strength between a word and a cluster. But we need to decide whether it should be  $P(w_j|C_i)$  or  $P(C_i|w_j)$ . Both these conditional probabilities are related by the Bayesian formula

$$P(C_i|w_j) = \frac{P(w_j|C_i)P(C_i)}{P(w_j)}$$

చిన్న చదరం, chinna chadaraM,	వృత్తం vRittaM	వెనకాతల venakAtala	పెద్ద చదరం pedda chadaraM	పరిగెడుతుంది parigeDutuMdi
Big Square is running after small square and circle				
	వృత్తం vRittaM	చిన్న చతురస్రంలని chinna chaturasraMlani	వెంబడిస్తున్నది veMbaDistunnadi	
Following Circle and small square (Who?)				

Figure 4: The sample Telugu commentary for the frameset 2093-2178 in two different commentaries and their transliterations and translations. These were the sentences before doing any corrections in them.

చిన్న చదరం, chinna chadaraM,	వృత్తం vRittaM	వెనకాతల venakAtala	పెద్ద చదరం pedda chadaraM	పరిగెడుతుంది parigeDutuMdi
Big Square is running after small square and circle				
చదరం chadaraM	వృత్తం vRittaM	వెనకాల venakAla	డబ్బా DabbA	పరిగెడుతుంది parigeDutuMdi
	వృత్తం vRittaM	చిన్న చతురస్రంలని chinna chaturasraMlani	వెంబడిస్తున్నది veMbaDistunnadi	
Following Circle and small square (Who?)				
డబ్బా DabbA	వృత్తం vRittaM	చదరంలని chadaraMlani	వెంబడిస్తున్నది veMbaDistunnadi	

Figure 5: The sample Telugu commentary in Fig.4. is edited and the result is shown in Telugu and transliterated. In the first sentence, the word venakAtala is changed to venakAla and in second sentence, anaphora resolution is done by identifying the noun. In both sentences noun tagging is done by using common words for the three shapes.

Now,  $P(w_j|C_i)$  can be taken as  $\frac{k_{ji}}{n_i}$  where  $k_{ji}$  is the frequency of word  $w_j$  in the cluster  $C_i$  and the  $n_i$  is the total frequency of all words in cluster  $C_i$ . By using this, we find the conditional probability  $P(C_i|w_j)$  and use it as an association measure. This measures the probability of a given word  $w_j$  being a label for the cluster  $C_i$ .

## 6 Results

We followed some rules in associating the words with clusters. In each commentary, a sentence is considered to belong to a frame set depending on the time when it is spoken. In each frame, only those sentences which contain the noun for the object in focus in that frame are considered. Else, the sentence may be describing something else that is not in focus and so not in cluster. The words are associated to clusters using the measure described above. Fig.6 shows the results for the clusters C1, C2, C3 and C4 when monograms (single word labels) are used.

We observe that the results for chase when the focus object is *chasing* are pretty good. It is observed that the action *hit* has been classified as frequent alternate actions of *come-close* and *move-away*. Hence the clusters C1 and C2 have high

Cluster 1(Come Closer)			
గుడ్డుకుంటున్నాయి	guddukuMTunnAyi	hitting(plural)	0.053
వెంబడిస్తుంది	veMbaDistuMdi	chasing	0.04
వెనకాల	venakAla	back/after	0.038
వచ్చాయి	vachchAyi	came(plural)	0.035
Cluster 2(Move Away)			
గుడ్డుకుంటున్నాయి	guddukuMTunnAyi	hitting(plural)	0.036
కదిలింది	kadiliMdi	moved	0.032
గట్టగా	gaTTigA	strongly	0.031
తీసింది	tOsiMdi	pushed	0.031
Cluster 3(Chase A)			
వెంబడిస్తుంది	veMbaDistuMdi	chasing	0.037
వెనకాల	venakAla	back/after	0.036
పరిగెడుతుంది	parigeDutuMdi	running	0.036
వచ్చాయి	vachchAyi	came(plural)	0.034
Cluster 4(Chase B)			
వెంబడిస్తుంది	veMbaDistuMdi	chasing	0.012
వెనకాల	venakAla	back/after	0.007
పరిగెడుతుంది	parigeDutuMdi	running	0.006
పడింది	paDiMdi	fallen(as in chase)	0.006

Figure 6: Results for each cluster, with the Telugu word, transliteration, translation and probability measure.

measure for the word *hit*. Also, the use of bigrams can make the labelling more consistent as can be seen in case of C3 and C4 where the words occurring together have high probability measure.

## 7 Conclusion

We have shown a method of learning grounded labels for transitive verbs in Telugu language. By using bigrams and higher n-grams, the accuracy of the labelling can be increased. Another method widely used in linguistics is to use Stemming. Languages like Telugu change the base word by adding cases or prepositions to describe actions, gender, tense etc. In our method we did not do stemming. So there will be words which are counted differently for each variation of base form. This can be countered by stemming. In stemming we take only the base forms of the words. So, stemming can lead to an increase in the accuracy of associations. Also it can be used to learn grounded translations between languages by repeating this on other languages. The work in English for actions had been done by [2]. By using the cluster labels in both the languages, the translations between English and Telugu languages can be learnt using grounded sources.

## References

- [1] David A. Robertson Arthur M. Glenberg. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43:379–401, 2000.
- [2] G. Satish. Unsupervised concept and language acquisition using perceptual attention on multimodal input, 2008.
- [3] Priwathijit Guha. *Unsupervised Concept Aquisition From Surveillance Videos*. PhD thesis, 2009.
- [4] Subhransu Maji Vivek Kumar Singh and Amitabha Mukerjee. Confidence-based updation of motion conspicuity in dynamic scenes. In *Third Canadian Conference on Computer and Robot Vision*, 2006.
- [5] Amitabha Mukherjee and Mausoom Sarkar. Grounded acquisition of containment prepositions. In *International Conference on Natural Language Processing*, 2007.
- [6] Marc Strickert and Barbara Hammer. Merge som for temporal data. *Neurocomputing*, 64:39–71, 2005.
- [7] Clayton T. Morrison Paul R. Cohen and Erin Cannon. Maps for verbs: The relation between interaction dynamics and verb use. In *International Joint Conferences on Artificial Intelligence*, 2005.
- [8] G.Satish and Amitabha Mukherjee. Acquiring linguistic argument structure from multimodal input using attentive focus. In *International Conference on Development and Learning*, 2008.
- [9] Deb K. Roy and Alex P. Pentland. Learning words from sights and sounds. *Cognitive Science*, 26:113–146, 2002.
- [10] Kobus Barnard and Pinar Duygulu et al. Matching words and pictures. *Journal of Machine Learning Research, Special Issue on Machine Learning Methods for Text and Images*, 3:1107–1135, 2003.