

Multilingual Associations by Language Label Learning on a Video

CS498 Project Report

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

Advisor: Prof. Amitabha Mukherjee

November 6, 2009

Abstract

This report presents a method of associating words from different languages by using them to describe a video. Semantic information from video can be used to find out the various actions and by using commentaries from various sources, labels for those actions can be learnt in a language. A Computational Model of Visual Attention is used to find the most salient object or action in the current set of frames and by matching this with the commentary, the corresponding word or phrase is learnt. By doing this for three languages namely, English, Hindi and Telugu, we propose to see if we can learn associations between words of these languages by choosing the best match word in each language. Thus, we wish to construct a multilingual lexicon for such actions.

This report presents my work in this project and also the future work that should be done in the next semester.

Contents

1	Introduction	2
1.1	Past Work	2
2	Problem Structure	2
3	Methodology	3
3.1	Input: Heider and Simmel Video	3
3.2	Gaze Model	3
3.3	Concept Learning	4
3.4	Language Label Learning	4
4	Work Done	4
4.1	Commentaries	5
4.2	Analysis of the Commentaries	5
5	Work to be done Next Semester	6

List of Figures

1	Commentary for a sample set of frames from PETS2000 dataset	2
2	A frame from Heider and Simmel Video	3
3	A frame from the Video after Gaze Predictor is applied	4
4	Multilingual Lexicon	5
5	Example Commentaries	6

1 Introduction

Obtaining semantic information from a video is of great importance today. Much work is being done in the area of visual concept acquisition in complex multi-object videos. It is natural for human beings to associate things they see with some *word*. A similar thing can be done for such actions or objects identified in a video i.e., associate the concepts with a language label. This can be used in linguistics for associating words from different languages by using the labels from different languages for various objects or actions acquired from a video. This B.Tech Project deals with this Multilingual association, specifically for the languages English, Telugu and Hindi.

1.1 Past Work

Some amount of work has been going on in this project for the last couple of years in IIT Kanpur and elsewhere. Language Label Learning from 3D videos by Guha[1] is a recent work from which this project draws its inspiration. Figure 1 shows an example commentary for a sequence of frames from the PETS2000 dataset used by Guha. Also, significant work has been done in acquiring containment prepositions and linguistic argument structure from videos. Other important works in similar fields are done by Cohen and Morrison[2], Duygulu and Barnard[3], Roy and Pentland[4].



Figure 1: These are the frames numbered 125, 150, 175 from the PETS2000 dataset. A commentary in English for the frameset 125-175 could be 1. *A car is going from right to left* or 2. *A red car is moving..* Similar commentaries can be collected in other languages.

2 Problem Structure

The Problem can be divided into the following steps.

1. Acquiring the visual concepts from a video. It involves the problem of object tracking.
2. Identifying the most salient visual concept in a frame. This is done by a computational model of visual attention. This gives us the object or action most salient in a set of frames.
3. Collecting commentaries for the video in the three languages. The commentaries are then typed in and separated according to sentences.

4. Language Label Learning by matching the sentences to the most salient object or action in the corresponding set of frames and finding the best match among them.
5. Associating and then analysing the best matches from the three languages.

3 Methodology

Various literature were studied to understand the problem and the past work done in this area. To understand the basics of video processing and object tracking, a simple background subtraction algorithm is implemented on a sample dataset (PETS2000). Object extraction and Language label learning on 3D videos has already been done by Guha[1]. But it focuses on learning nouns. Our project focuses on activity learning, which includes verbs and verbal phrases. Since activity learning is hard on 3D videos, we shifted our attention to 2D videos.

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, there are three agents, a big square, a small square and a circle which interact in and around a square box with an opening. The interactions include chasing each other, moving in and out of the box, hitting each other and coming close to each other.

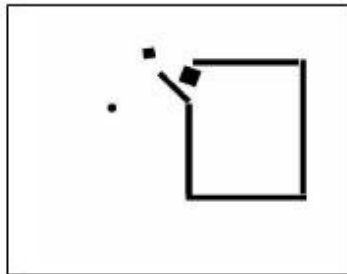


Figure 2: This is a frame from the input video we use. Three objects viz., a big square, a small square and a circle interact with each other.

3.2 Gaze Model

To find out the most salient visual concept in a set of frames, the gaze predictor model proposed by Singh, Maji and Mukherjee[5] is used. The original computational model of visual attention is based on the work by Koch and Ullman. This gaze predictor includes extensions to dynamic image streams to incorporate motion features in saliency computation.

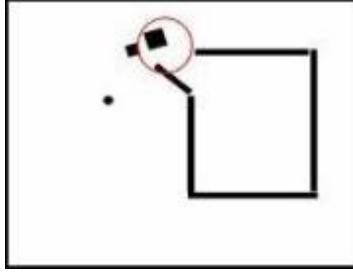


Figure 3: A frame from the Video after Gaze Predictor is applied
The most salient object is shown in the red circle.

3.3 Concept Learning

Our goal is to learn activity from a video in different languages and associate them. Learning transitive verbs from a 2D video is discussed in (Satish and Mukherjee, 2008)[6]. This work deals with learning action schemas from a video using *Merge Neural Gas Algorithm*. Using this algorithm, visual events are organised into action classes. The computational model of visual attention is used to find the most salient object or action. It is assumed that the commentator focuses on the action that the model finds salient (*Perceptual Theory of Mind*). Now by using noun-referrant mappings and considering only those linguistic expressions referring to objects or actions in perceptual focus, action schemas for these actions are learnt. These schemas are then validated by using them for language production to describe a 3D video. The input video used there is same as ours, the Heider and Simmel Video.

3.4 Language Label Learning

We wish to learn the labels for the objects and actions from the video in the following categories.

- **Objects:** Big Square, Small Square, Circle, Box
- **Relations:** Near, In, Out etc.
- **Activity:** Chase, Move closer, Hit, Go to etc.

For this atleast ten commentaries are collected for each language and respective labels are learnt. Then we construct a multilingual lexicon based on semantics by associating these labels. Figure 4 describes the structure of our lexicon. Observe that our lexicon is based on the semantics of the languages but not their syntax.

4 Work Done

Many papers were studied for this project and they are mentioned in the references. Apart from that, a single gaussian based background subtraction algorithm was im-

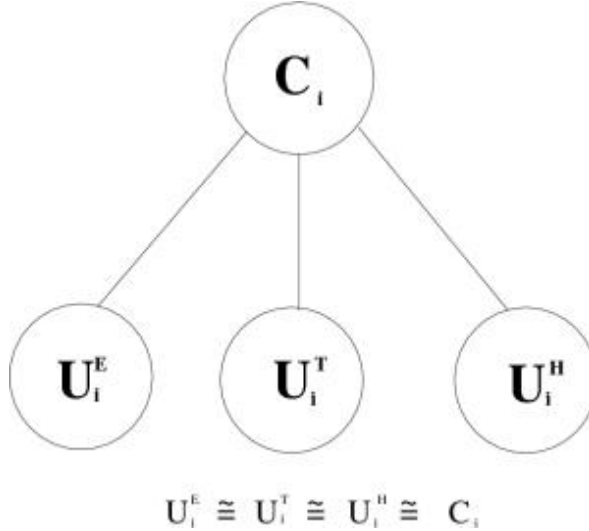


Figure 4: C_i is a concept learnt from the video. Let U_i^E , U_i^T and U_i^H be the labels associated with the concept C_i in the languages English, Telugu and Hindi. Then we wish to construct a multilingual lexicon $U_i^E \cong U_i^T \cong U_i^H \cong C_i$ based on semantics.

plemented initially, to understand a few basics in video processing. Later on, commentaries were collected on the Heider and Simmel video.

4.1 Commentaries

The Video is shown to three individuals of age group 20-22, who have enough knowledge in one of the languages, English, Hindi and Telugu. They are asked to comment on the video as it is happening. The commentaries are stored in wav files. Each commentary is then typed according to the sentences and the corresponding *start-time* and *end-time* are noted. The video frame rate is *30fps* and so the corresponding frames are given by

$$\begin{aligned} \text{start-frame} &= 30 * \text{start-time} \\ \text{end-frame} &= 30 * \text{end-time} \end{aligned}$$

The audio trails the video by a few frames but this can be neglected as the length of the set of frames is very large compared to this lag i.e.,

$$\text{lag} \ll (\text{end-frame}) - (\text{start-frame})$$

4.2 Analysis of the Commentaries

We considered only one commentary from each language, for now. Figure 5 shows the commentaries in different languages for a few set of frames.

Even in such a small data, it can be observed that for the framesets 1 and 5, the sentences are translations of each other. But in the frameset 4, the commentary in English and Telugu refers to *coming out of door* whereas the commentary in Hindi

S.No	Frame set	English
1	180 - 210	A circle and a small square have entered the frame
2	395-436	The small square seems to be moving around the circle
3	977-1004	Circle enters the door
4	1767-1807	Circle goes outside the door
5	2469-2492	The big square breaks the door

Telugu	Hindi
చిన్న చదరం, వృత్తం వచ్చాయి	छोटा वर्ग और वृत्त बक्से के बाहर हैं।
పెద్ద చదరం గదిలో బయటకు రావడనికి ప్రయత్నిస్తూ ఉంది	बड़ा वर्ग बक्से में घूम रहा है।
వృత్తం గది లోపలికి వెళుతుంది	वृत्त बक्से के अंदर जा रहा है।
వృత్తం తలుపు నుండి బయటకు వచ్చేసింది	वृत्त बक्से से बाहर आ गया।
పెద్ద చదరం తలుపు పగలగొట్టింది	बड़े वर्ग ने दरवाजा को तोड़ दिया।

Figure 5: Example Commentaries

refers to *coming out of box*. This sort of confusion between *box* and *door* can be eliminated with more commentaries. A similar confusion is seen in frameset 3 between *box* and *door*.

In frameset 2, the commentaries differ greatly in the action they focus on. The English commentary refers to the *small square moving around circle* whereas in Hindi it is *big square roaming around in box* and in Telugu it is *big square trying to come out of box*. In this frameset, the gaze model shows a change in attention from the small square to the big square.

5 Work to be done Next Semester

Work has to be done further to achieve desired results in this project. As yet, only one commentary has been taken from each language. More commentaries (at least ten in each language) should be collected and analysed to get better results in each language. Only then, we can associate the words from each language.

First, the video needs to be segmented into framesets using the gaze model. Then, the object in focus in each frameset is found. This is used to find the associations

between the actions or objects to the word-separated commentaries. As we are working with a video which has a small set of objects, mapping nouns is easy. Then learning verbs and the predicate structure is necessary to achieve our goal of activity learning. Learning prepositions like *in* and *out* has been successfully achieved by Mukherjee and Sarkar[7]. Acquiring predicate structure has been achieved by Satish and Mukherjee[6]. These works will be used in learning the verbs and predicate structures.

Then we map the best match word for an action or object in each language to construct our multilingual lexicon. This will be the course of this project in the next semester.

References

- [1] Priwathijit Guha. *Unsupervised Concept Aquisition From Surveillance Videos*. PhD thesis, 2009.
- [2] 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.
- [3] 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.
- [4] Deb K. Roy and Alex P. Pentland. Learning words from sights and sounds. *Cognitive Science*, 26:113–146, 2002.
- [5] 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.
- [6] G.Satish and Amitabha Mukherjee. Acquiring linguistic argument structure from multimodal input using attentive focus. In *International Conference on Development and Learning*, 2008.
- [7] Amitabha Mukherjee and Mausoom Sarkar. Grounded acquisition of containment prepositions. In *International Conference on Natural Language Processing*, 2007.