Word and Phrase Learning based on Prior Semantics

Amitabha Mukerjee and **Nikhil Joshi**
Department of Computer Science and Engineering
Indian Institute of Technology,
Kanpur (UP, India) - 208016
{amit,joshins}@cse.iitk.ac.in

Abstract

Based on the evidences of preverbal conceptual development in infants, we adopt semantics-first approach for word-learning. We first cluster several perceptual categories from a complex visual interaction. Using a surveillance traffic video, we a) identify the moving objects by separating these from a static background, and b) group the similar appearances into clusters. The resulting models are found to be noisy approximations of traffic object categories and motion actions. Next, we consider these models along with parallel commentaries that describe the scene in free, unconstrained language. A bottom-up model of dynamic attention is applied to identify objects in perceptual focus, which are mapped to words in co-temporaneous utterances. Using no language-specific knowledge such as syntax, we show the ability to learn words for the object classes and also for the motion actions.

1 Introduction

The problem of word-learning primarily focuses on mapping the linguistic representation of a word to its semantics. In most of the attempts to learn the language to semantics mappings, semantic representations were often limited to the logical representations (Zettlemoyer and Collins, 2005; H Alshawi, 2011). However, the need of much richer semantic representations such as percptual schema (Barsalou, 1999), image schema (Mandler, 1992) is argued for grounding the meanings of words(Harnad, 1990). A number of approaches have tried to construct such term-meaning associations from sensorimotor data (Steels and Kaplan, 2002; Gorniak and Roy, 2004; Roy and Pentland, 2002; Oates et al., 2000). However, these approaches used scenes with simple objects and constrained linguistic descriptions. Also, learning was guided by considerable feedback.

In this work, we consider learning objects and interactions from a complex 3D-scene and mapping them to words and phrases from free, unconstrained language with full sentences describing the scene. The key to handle referential uncertainty (Siskind, 1996) is the visual saliency predicted using a bottom-up attention model. The salient objects are then associated with the co-occurring utterances in the narratives to learn the labels for the visual concepts. Owing to evidences of preverbal conceptual development (Mandler, 1992), we adopt semantics-first approach (Yu and Ballard, 2004) where we learn visual semantics first and then discover appropriate word associated with it

For learning objects and interactions, image sequences from a fixed camera, as typically used in surveillance scenarios, are considered. The stable patterns of background are first learned, and used to extract foreground blobs corresponding to the objects of interest. The object blobs are tracked across the frames and regions of occlusion are identified. Unoccluded object appearances are then projected to a feature space based on the "Pyramidal Histogram of visual Words" (PHOW) approach (Bosch et al., 2007). The resulting PHOW descriptor for the blobs are then classified in an unsupervised manner, resulting in a number of object classes. For every object tracked, a trajectory is modeled using the position and velocity of object blobs in successive frames. These trajectories are then clustered

Proceedings of Recent Advances in Natural Language Processing, pages 616–621,
Hissar, Bulgaria, 12-14 September 2011.

to obtain a number of motion classes. The object and motion classes obtained are evaluated based on the user labels (the *ground-truth*). We note that the resulting models are similar to the *conceptualizer* of (Steels and Kaplan, 2002), but unlike in that work, the model here is learned and not programmed beforehand.

For the word association task, we first compile a set of narratives by asking nine adults to describe objects and activities in free unconstrained language. The transcribed narratives (text) are then aligned with the objects and activities in visual focus, as identified by the bottom-up attention model. We are able to discover the appropriate nouns for four object classes with high visual purity viz. BICYCLE, MOTORCYCLE, TRUCK and CAR. Phrases like "*bAe-N se dAe-N*" and "*geT kI taraf*" are also discovered for the trajectory LEFT-TO-RIGHT and TURN. During association, we remove units that are very frequent in general discourse, assuming these to be non-relevant to this context. However, no linguistic knowledge (pos, syntax or morphology) except the knowledge of word segementation is assumed.

Our unsupervised approach to word learning implies two important scalability advantages. Since we use no knowledge of the camera placement or the types of objects in the scene, the visual analysis is potentially applicable to a wide range of scenes. Also, since we use no knowledge of the syntax of the target language, it is possible to use the approach to other languages as well. Since the terms learned are grounded in the visual domain, it can be flexibly related to new input situations. This is demonstrated in this work via successful queries on novel traffic video.

2 Unsupervised object classification

In recent years, supervised learning for visual object categories has been able to distinguish hundreds of classes of objects with high accuracies (Bosch et al., 2007; Mutch and Lowe, 2006). The critical step in these approaches is to project the images onto a set of patterns, called "words", so that each image is characterized as a distribution on the words. This class of approaches, known as "bag of words" after similar approaches in document analysis, classify novel images based on their similarity to the trained models. In this work, we extend these ideas to unsupervised object classification. Here the object images (foreground blobs from surveillance video) have the advantage that these are relatively tightly cropped around the region of interest. We track salient patches in each blob to identify the same agent across contiguous frames - sample views of some agents are shown in Figure 1. As can be seen, the results are very noisy owing to occlusions, shadows, tracking errors, agent appearance changes etc.

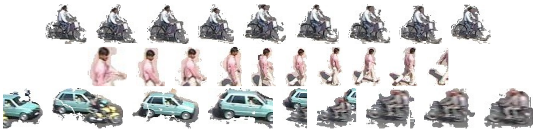

Figure 1: *Agents as sequences of isolated foreground blobs.* Bottom row (agent 130): the sequence is initially tracking a car - but after it exits, it is erroneously mapped to a motorcycle.

The tracking step considers substantially overlapping sequences of blobs. Only where an agent is isolated is the blob considered for modeling its appearance. We use the pyramidal histogram of words (PHOW) approach (Bosch et al., 2007), based on computing the SIFT operator (Lowe, 1999) on a very large number of points (100K) on these blobs. These are clustered to obtain a code-book of 300 "words". Next, each foreground blob in a tracked agent is projected onto these words, and the agent is modeled as a probability distribution on the space of words (estimated by the histogram).

Using a Bhattacharya distance metric, the histograms are clustered using k-means (results reported for $k = 30$). This results in an oversegmentation of the category space, and to evaluate the effectiveness of the clusters, we manually categorize the agents into seven *groundtruth* classes: TEMPO (T), BICYCLE (B), MOTORCYCLE (M), TRUCK (L), HUMAN (H), CAR (C), and also a small category NOISE (N) with object fragments and lighting effects etc. The purity of each cluster is defined as the percentage of its dominant class. We assign the dominant ground-truth category in a cluster as ground-truth of that cluster. The average

Class: # agents	Clusters	Purity
H:52	C1,C2,C4,C10, C11,C12,C14,C21	51/63 (81%)
M:36	C3,C8,C9,C22, C23,C24,C26	35/48 (73%)
B:32	C5,C6,C7,C15, C20,C28	22/25 (88%)
T:21	C0,C16,C17, C18,C25	15/27 (56%)
L:12	C12,C29	11/13 (83%)
C:16	C19	9/10 (90%)
N:8	C27	2/4 (50%)

Table 1: *Clusters from k- means (k = 30).* Clusters are assigned to one of six ground-truth categories. Purity of a cluster = degree to which it is dominated by a single object category.)

purity of the clusters obtained by this process is 76.5%. By training the model with a $N-M$ of agents and testing with the remaining M, we obtain a cross-validation accuracy of 70.8% (for $M=5$). Table 1 shows the ground-truth distribution for 30 clusters obtained using k-means. Figure 2 shows blobs of agents from some of the clusters formed for k=30.

Some clusters appear to have fine-graded semantic significance - e.g. the class C16 (“passengers getting off from tempo”) and C21 (“humans either on some vehicle”)in Figure 2.While such classes were not marked in the ground-truth, this finer discrimination may actually help in detecting activities. Some other clusters are less meaningful; e.g. cluster C27 , is mostly noise.

For every agent tracked across the frames, we define its trajectory based on position and velocity of object blobs in ten frames at regular intervals. Positions of an agent are taken relative to its position in the starting frame to avoid locational bias. Based on these features and euclidean distance measure, trajectories are clustered into seven clusters using *k-means* algorithm. For evaluation purpose, we marked the ground-truth of these trajectories as one of the five categories: LEFT-TO-RIGHT (LR), RIGHT-TO-LEFT (RL), TURN (T), CROSS (C) and NOISE (N) with not so meanigful trajectories. The purity of each cluster is calculated in the same way as it is calculated for object clusters. Table 2 shows the distribution of ground-truth categories for each of the seven trajectory clusters discovered. Similarly, many vehicles crossing the road come from left, move towards right and then cross the road resulting in low purity of C5. The very low purity of cluster C6 is mostly because of noisy trajectories of human blobs which move arbitrarily in the scene. Errors in tracking agents also result in noisy trajectories and lead to inaccuracies in the clustering.

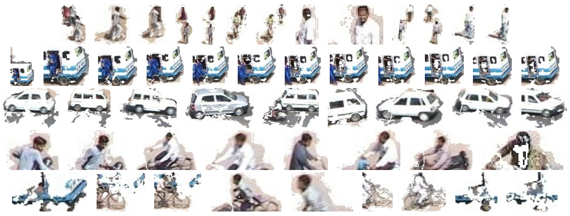

Figure 2: *k-means (k = 30) clusters* Clusters C10, C16, C19, C21, C27. Representative views from all agents in each class are shown. The membership of these clusters can be seen in Table 1. Whereas C10 and C19 are relatively clean classes, C27 has several noise agents

Cluster/GT	LR	RL	T	C	N	Purity
C1 (RL)	0	20	0	0	1	20/21
C2 (LR)	15	0	1	0	1	15/17
C3 (LR)	20	0	2	0	1	20/23
C4 (RL)	0	26	8	1	3	26/38
C5 (LR)	21	2	4	8	4	21/39
C6 (LR)	13	8	4	2	7	13/34
C7 (T)	0	3	14	3	0	14/20

Table 2: *Ground-Truth distribution*: Distribution of ground-truth categories for each of seven trajectory clusters

3 Attention Model

We use attention model to find the most salient part of the scene that humans are likely attend to. The words used in the description are more likely to refer to objects that are in perceptual focus. This resolves the referential uncertainty.

In general, attention combines bottom-up mechanisms (independent of task) with top-down mechanisms (task dependent). While a number of models are available for bottom-up attention, on both still (Itti and Koch, 2001) and dynamic (Singh et al., 2006) images, top-down attention is far more difficult to model owing to complexities in modeling the task. Also, in our context, commentaries were collected without providing any specific task, so we use a dynamic bottom-up model.

In our work, we have an advantage over traditional dynamic attention models since the objects of attention are already segmented and available as tracked sequences of segmented foreground blobs. These are the scene regions that are competing for attention. Unlike many computational models that consider saliency of pixels in the data, we are in a position to evaluate the saliency of the segmented foreground region directly. Our attention model is based on the findings that a) Objects with higher speed are likely to be more salient, and b) Objects with a larger image size are more likely to be attended (Itti and Koch, 2001). We ignore some other factors such as colour and texture, which are more relevant in still images; for image sequences, motion and size are more significant. In addition to the saliency map based on the above factors, we also need to construct a confidence map, based on how recently was the object attended. Objects which have not been attended for some time tend to decay in their confidence, and thus become more likely to be attended to. We combine all these aspects to define saliency of object blob j as

$$S_j = (1 - e^{-k\Delta t})(w_1 A_j + w_2 v_j)$$

where A_j is the image area (in pixels) and v_j is image speed (in pixels per frame) of the object j. Δt is the time elapsed since the object was last updated. Parameters w_1, w_2 and k capturing relative importance of object size, velocity and confidence are all set to 1.

4 Learning language labels

For the purpose of learning language labels for concepts learned from video, we use human narratives describing the same visual scene. We asked 9 native speakers (college students: all male) to watch the video once, and give their commentary on it the next time around. In the instructions, they were asked to focus on people, vehicles in the scene and their activities. The narratives were broken into segments at sentences boundaries as well as at pauses longer than 1.5s, and transcribed without correcting grammar errors. Also, initial 40 seconds and final 20 seconds of data were discarded since people appeared to be talking more generally at the beginning of the video, and events in the end could not be commented upon. Around 600 sentences with 3398 words were used in the analysis.

Since the subjects were not constrained in their descriptions in any way, the lexical choice and linguistic constructions varied widely. Thus the same event may be described as "gADI dAe.N se bAe.N or gayI" (car went from right to left), "ek sa.NTro gayI"(one Santro went) etc. As perspectives varied tremendously, for the same time interval in the video, different subjects said: "ek kAr aAyI" (One car came), "vah saD.ak krOs kar rahA hai" (He is crossing the road) etc. Even after asking the narrators to focus on the people, vehicles and their activities during instructions, the commentaries collected include considerable peripheral descriptions like "bIch meIn koI DivAiDar nahIn hai" (There is no divider in the middle).

In order to identify the relevant linguistic units, we align segments of the commentary with the most salient objects in the video as identified by the attention model described above. For computational purposes, we assume linguistic units to be contiguous at word-level and associate k-grams (for $k = 1$ to 4) with co-occuring salient concepts in the video. We seek to identify the unit having maximal *conditional probability* given a concept.

4.1 Object-Label associations

Table 3 reports the top two 1-gram at the word level for six ground-truth object classes. The conditional probabilities shown are multiplied by 100. Dominating associations are discovered for four object categories: BICYCLE, MOTORCYCLE, TRUCK and CAR (*sAikal*, *bAik*, *Trak* and *kAr* respectively). Units like *lefT* ("left"), *dAe.N* ("right") indicating the directions of movement are also appearing among top2 *1-word* associations.

Part of the reason for difficulty in learn-

Concept	Word	Cond. Prob
TEMPO	bAik	6.70
	dAe-N	6.46
BICYCLE	sAikal	3.17
	moTarsAikal	1.59
MOTOR-CYCLE	bAik	8.37
	Tempo	8.29
TRUCK	Trak	19.54
	lefT	6.13
HUMAN	dAe-N	10.86
	pe	10.29
CAR	kAr	7.50
	pe	5.00

Table 3: *Association results:* top2 1-word associations for each of object categories

Trajectory	3-gram	Prob
C1	purI KalI hai	1.71
	saD.ak pUrI KalI	1.71
C2	bae-N se dAe-N	3.16
	lAl SirT me-m	2.73
C3	bae-N se dAe-N	4.44
	puch rahA hai	3.96
C4	roD krOs kar	4.62
	krOs kar rahA	4.47
C5	krOs kar rahA	4.67
	roD krOs kar	4.20
C6	kuch log roD	2.20
	dae-N kI taraf	2.18
C7	geT kI taraf	3.57
	Ai Ai TI	3.57

Table 4: *Association results:* top2 3-word associations for each of trajectory clusters

ing labels for other categories can be seen in Table 1, where we see that the average purity for CAR, BICYCLE and TRUCK is quite high whereas that for TEMPO is very poor. Though the purity of HUMAN is moderate. we find that there are many relevant labels in the narratives; e.g. a person with bicycle is described as *sAikalwAlA* (bicyclist) or as *aAdmI* (man). Also, attentional salience is more often on the larger, faster-moving vehicles and not on smaller human blobs. Possibly for these reasons, label for humans is not learned.

4.2 Trajectory-Label association

Table 4 shows top2 3-grams according to conditional probability measures for seven clusters of trajectories. As can be seen, for clusters C2 and C3 representing Left-to-Right (LR), *bAe-N se dAe-N* ("left to right") appears as the strongest 3-gram. Similarly, for clusters C5 and C6 *dAe-N kI taraf* ("towards right") appears third (not reported here). For the cluster C7 representing TURN (T), *geT kI taraf* ("towards the gate") appears as the strongest label as the agents in the cluster C7 are generally turning towards the gate of an institute. For other two clusters, C1 and C4, however, appropriate labels could not be learnt. Perhaps, the event of RIGHT-TO-LEFT may not have been commented as profoundly as the events of LEFT-TO-RIGHT or TURN.

4.3 Testing on Novel scenes

In order to test our semantic model, we used two different videos of the similar scene, and attempted to recognize the three classes of objects with high viusal purity.

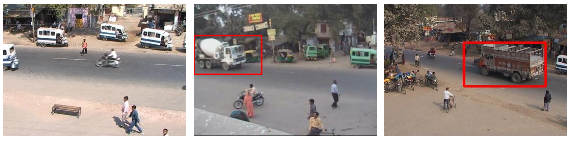

Figure 3: *Test videos.* Training video at left. Samples from two test videos, from novel camera positions, at middle and right.

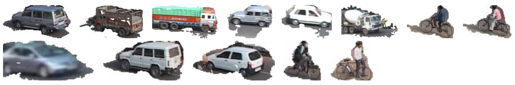

Figure 4: *Test agents from novel videos. Sample blobs from thirteen test agents. Agents on bottom row were not correctly labeled.*

These videos were shot on different days, from different vantage points, and varied considerably in the imaging (Figure 3). Our video query consisted of identifying objects of a given class. For evaluation, we manually identified TRUCK (3), BICYCLE (5), and CAR (5) agents. Sample blobs for each agent shown in Figure 4. The truck query responded with all three agents of TRUCK. The CAR agents did not fare that well, only two out of five were correctly identified; two being labeled as TEMPO (possibly because the class TEMPO was very noisy and had several CARs in it), and one as MOTORCYCLE. One of the cars seen in the novel video is a sedan (leftmost in bottom row, Figure 4), which was not present in the training data. Three of the BICYCLE agents were correctly identified; other two were HUMAN (man standing besides his bicycle) and TEMPO but were misinterpreted as BICYCLE.

As we scale up and include more videos and different vantage points for training, more refined models of object classes are expected to be learned, so that such production or recognition errors would go down.

5 Conclusion & Future Work

In this work we have attempted to learn visual concepts for some object classes and motion trajectories, and map these to Hindi words or phrases, based on a) an unsupervised model that discovers object categories from a fixed-camera video; b) a model of synthetic blob-based attention that identifies the most salient agent among many moving objects; and c) an association between the concepts learnt from the video and the *k*-grams in the user commentaries. The model has been demonstrated in a video querying task.

Our unsupervised object clustering is able to distinguish among several object categories and also some motion trajectories even from a very short video of around 4.5 minutes. With greater exposure, the models may be refined further. Further, there were only 600 sentences of narrative with which to work. To put it in context, a typical child is exposed to a much larger corpus of co-occurrent text and visual context every *hour*. As NLP searches for richer models of semantics, such multimodal data mining will become more widely used. To help bootstrap this process, both the multimodal corpora and and textual database has been made available.

Given the unsupervised nature and particularly the minimal dependence on linguistic knowledge, we are currently expanding this approach to learn several languages. A larger goal is to integrate models of motion-trajectories with the knowledge of nominals, and begin to attempt to build the kind of defeasible knowledge structures.

References

[Barsalou1999] L Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609.

[Bosch et al.2007] A. Bosch, A. Zisserman, and X. Munoz. 2007. Image classification using random forests and ferns. In *ICCV'07*, pages 1–8.

[Gorniak and Roy2004] P. Gorniak and D. Roy. 2004. Grounded semantic composition for visual scenes. *JAIR*, 21(1):429–470.

[H Alshawi2011] M Ringgaard H Alshawi, P Chang. 2011. Deterministic statistical mapping of sentences to underspecified semantics. In *Proceedings of the Ninth IWCS*, pages 15–24.

[Harnad1990] S Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335 – 346.

[Itti and Koch2001] L Itti and C Koch. 2001. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

[Lowe1999] D Lowe. 1999. Object recognition from local scale-invariant features. In *ICCV, 1999*, volume 2, pages 1150 –1157.

[Mandler1992] J M Mandler. 1992. How to Build a Baby: II. Conceptual Primitives. *Psychological Review*, 99(4):587–604.

[Mutch and Lowe2006] J. Mutch and D Lowe. 2006. Multiclass object recognition with sparse, localized features. In *IEEE CVPR*, volume 1, pages 11–18.

[Oates et al.2000] T Oates, Z Eyler-walker, and P. Cohen. 2000. Toward natural language interfaces for robotic agents: Grounding linguistic meaning in sensors. In *Proceedings of the 4th ICAA*, pages 227–228.

[Roy and Pentland2002] D Roy and A Pentland. 2002. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146.

[Singh et al.2006] V K Singh, S Maji, and A Mukerjee. 2006. Confidence based updation of motion conspicuity in dynamic scenes. In *Third Canadian CRV 2006*, page 13.

[Siskind1996] J M Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):1–38.

[Steels and Kaplan2002] L. Steels and F. Kaplan. 2002. Bootstrapping grounded word semantics. In *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, chapter 3, pages 53–74. Cambridge University Press.

[Yu and Ballard2004] C. Yu and D.H. Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 1(1):57–80.

[Zettlemoyer and Collins2005] L S. Zettlemoyer and M Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI*, pages 658–666.