Language Label Learning for Visual Concepts Discovered from Video Sequences

Prithwijit Guha[1] and Amitabha Mukerjee[2]

[1] Department of Electrical Engineering,
Indian Institute of Technology, Kanpur,
Kanpur - 208016, Uttar Pradesh,
pguha@iitk.ac.in
[2] Department of Computer Science & Engineering,
Indian Institute of Technology, Kanpur,
Kanpur - 208016, Uttar Pradesh,
amit@cse.iitk.ac.in

Abstract. Computational models of grounded language learning have been based on the premise that words and concepts are learned simultaneously. Given the mounting cognitive evidence for concept formation in infants, we argue that the availability of pre-lexical concepts (learned from image sequences) leads to considerable computational efficiency in word acquisition. Key to the process is a model of bottom-up visual attention in dynamic scenes. Background learning and foreground segmentation is used to generate robust tracking and detect occlusion events. Trajectories are clustered to obtain motion event concepts. The object concepts (image schemas) are abstracted from the combined appearance and motion data. The set of acquired concepts under visual attentive focus are then correlated with contemporaneous commentary to learn the grounded semantics of words and multi-word phrasal concatenations from the narrative. We demonstrate that even based on a mere half hour of video (of a scene involving many objects and actvities), a number of rudimentary concepts can be discovered. When these concepts are associated with unedited English commentary, we find that several words emerge - approximately half the identified concepts from the video are associated with the correct concepts. Thus, the computational model reflects the beginning of language comprehension, based on attentional parsing of the visual data. Finally, the emergence of multi-word phrasal concatenations, a precursor to syntax, is observed where they are more salient referents than single words.

1 Conceptual Spaces and Linguistic Labels

A traditional view of cognition holds that the concepts are declarative, amodal and conscious - perceptual abstractions are procedural schemas that reflect important cognitive skills, but do not qualify as concepts [1]. In this *late-conceptualization* view, concepts underlying language do not arise until the end of the sensorimotor stage (about one and a half years), roughly the same time as language itself.

However, mounting evidence for infant skills in categorization and event structuring has challenged this position leading to what may be called the *Perceptual-conceptualization* view: that processes of perceptual abstraction lead directly to symbolic structures. (see debates following the lead articles [2–4]).

Computationally, a fallout of the late-conceptualization position is that concepts and linguistic tokens must be learned simultaneously. Here the computational task involves simultaneously learning the concepts and their associations [5, 6]. This ignores any abstractions that may have formed over months of perceptual interaction and concepts are learned ab initio the moment linguistic tokens begin to appear. On a naive view, the perceptual-conceptualization position, where some degree of language-independent concept formation occurs in the pre-lexical stage, should be easier since these concepts are already available and they only have to be associated with the linguistic tokens. This approach also ties in with cognitive linguistics, where language is viewed as part of an embodied cognitive process, a mechanism for expressing (and transferring) categories acquired from sensory experience [7] rather than a purely formal symbol manipulation system.

In this work, we consider this debate in a computational perspective by simulating pre-lexical concept learning from complex natural images, followed by a very rudimentary model for associating these concepts with words from a word-separated language commentary. First, we seek to determine if a cognitively motivated model of visual cognition is competent to form concepts from complex real-life image data in the pre-lexical stage. Second, we explore if the availability of such concepts make it any easier to acquire language based on contemporaneous image sequences and word-segmented-textual descriptions.

The main difficulty in this process - which is also one of the traditional objections to perceptual symbols - is how to identify which part of a scene is relevant to the concept [8] - e.g. in the action of pouring milk from a jug, is it the colour of the jug that is relevant? We posit bottom-up visual attention as a mechanism for determining visual saliency, and show how this results in significant pruning of the possible concepts that can be associated with language labels. We use a computational model of dynamic visual attention [9, 10] to compute the saliency distribution over the image space.

1.1 Developmental Models of Perception

Consider the traffic scene of figure 1, say, with the complex interactions between vehicles, pedestrians, animals, bicycles, etc. How is the system to make sense of this complex domain? We feel that a developmentally motivated approach, focusing on the capabilities that an infant brings to bear on such a task, may be relevant.

Around the age of six months [11], infants are seen to observe the background for some time before beginning to pay attention to figure objects (foreground). This corresponds to well-known techniques in visual surveillance for learning a background model in order to identify and track the foreground objects. This

(a) (b) (c)

Fig. 1. *Traffic Scene* input. Multiple moving objects with uncalibrated camera: (a) Frame 50: White car moving from right to left (8 objects); (b) Frame 70: white car overlapping rickshaw; (c) Frame 539 : truck moving from right to left. Note that the occluded objects are also being tracked. Object shapes and trajectories are analyzed to abstract agent concepts, which are then associated with an unaltered textual narrative.

together with the occlusion behavior which has been widely studied in developmental literature, provides some evidence for the initial capabilities that infants may be bringing to the task of constructing structures in the perceptual space.

A key component of this process is a model of visual attention. For this we use an extension to dynamic images of the Itti-Koch model for static scenes [9]. This model is key to identifying the objects and actions in a scene, and eventually, in associating them with linguistic labels [12].

Another aspect of our work is the role of occlusion. In computer vision, occlusion is often viewed as an obstacle to be overcome. Increasingly, developmental models of perception seem to suggest that occlusion is one of the most salient aspects of a scene that an infant pays attention to from very early on. In our work, we have had some reasonable success in modeling interaction events between objects by using occlusion sequences as part of the visual signature for these events. An overview of the system can be seen in Figure 2.

Our approach differs from earlier computational work on grounded language acquisition. We are not limited to learning static features such as object shape [6, 13]. Other computational models either use simplified line drawing animations [5], or assume pre-defined force dynamics primitives based on which higher constructs are learned through observation [14]. The emphasis in social models of grounded learning is on evolving a lexicon via interaction games [15]. The attention-based approach of Yu and Ballard [12] is very close in spirit to our model, except that where as they identify objects in focus by actually tracking the speaker's gaze, we use a synthetic attention model.

Although the learner is observing scenes of considerable visual complexity - more than ten objects are often active simultaneously, we do not use camera calibration or obtain any $3D$ motions - all imagery abstractions are computed on the image plane alone. We also use no visual priors for the scene, nor any linguistic priors for the language elements, and show that attentional focus may be sufficient to associate actions and objects with words in textual narratives by

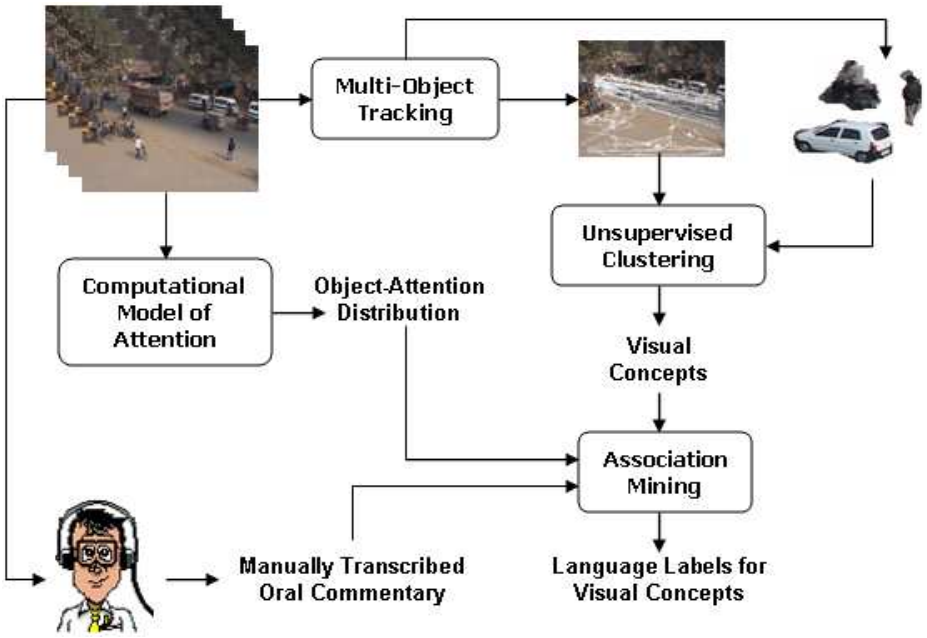


Fig. 2. *System overview.* Multiple targets are tracked in the input image sequence, and object shape templates, trajectories, and occlusions are mined to obtain appearance models and actions. These constitute the class of visual concepts. Oral commentaries acquired synchronously with the image sequence are now associated with the images. The association of a visual concept (concerning a certain object) to a language label (single or multi-word phrase) is computed as a function of the probability that the object is in attentive focus.

adult observers. Some samples of the commentary are shown in table 1); these are used as is without any simplification.

2 Object Detection, Tracking and Modeling

In the first phase of the work, concepts are built up from the image sequences. Each object instance is a space-time manifold characterized by the time indexed set of appearances - a collection of position (XY) and corresponding color (RGB) vectors along with the centroid-trajectory $\{\mathbf{c}(t)\}_{t=t_s^{(i)}}^{t_e^{(i)}}$ over its scene presence characterizes the object $A_i(t_s^{(i)}, t_e^{(i)})$. This object model encodes both its appearance and behavior and constitutes part of the cognitive percept or perceptual schema for the object.

Object models are acquired based on perceptual units mediated by attentive processes. Connected pixels moving in coherent motion are assumed to be objects, and occlusions between objects are handled. The visual input consists

Table 1. *Sample Commentaries* showing frame numbers spanned by each sentence. Note the diversity in the focus.

Frame Interval	Narrator #1
1 - 67	car left to right
68 - 111	white car right to left
112 - 159	one jeep right to left
160 - 209	bike going right to left
210 - 312	cycle left to right
313 - 508	person coming very slowly from the right
509 - 559	truck coming right to left
Frame Interval	**Narrator #2**
1 - 67	rickshaw moving down-wards
68 - 142	car moving and another tata Sumo seen
143 - 205	motor bike seen
206 - 247	person seen crossing the road
248 - 362	motorcycle right to left
363 - 432	person taking his cycle and walking
433 - 590	motorcycle, lorry and auto moving
Frame Interval	**Narrator #3**
1 - 47	car from left to right
48 - 94	cycle from left to right
95 - 138	car from left to right
139 - 163	person is crossing the road
164 - 217	bike is moving from left to right
218 - 310	cycle from left to right
311 - 364	bike is moving from the right side to the left
365 - 421	cycle entering from the left of the screen
422 - 524	bike from right to left
525 - 550	truck moving from right to left

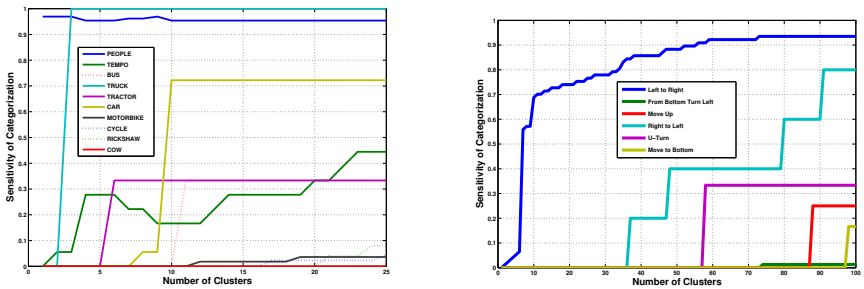
of traffic scenes with cars, people, bicycles, and vehicles - a total of 367 objects in 10 categories, captured with a static camera. In constructing models for each object, we use only image data; no $3D$ motions based on calibration data are used.

Objects are identified as foreground regions based on one of two kinds of evidence: first, as regions of change with respect to a learned background model [16]; and second, as regions exhibiting motion [17]. The background model is learned as a pixel-wise mixture of Gaussians only for those pixels which exhibit no image motion. Foreground blobs are associated with an object based on its motion-predicted support region. The objects are further localized by iterative centroid updates [18]. After all objects are localized, object-blob associations are re-computed and object models are updated only for those objects which are unoccluded by others.

2.1 Object Categorization

We perform unsupervised object categorization using the appearance (shape) and trajectory features, constituting a 3-manifold in image-space × time for each object. The shape features (dispersedness, area and aspect ratio) and the trajectory data of each discovered object are clustered by agglomerative hierarchical clustering [18]. We discover a total of 376 objects categorized into 19 different classes, of which several are infrequent outliers and a few appear due to misidentification of merged blobs and other tracking errors. Owing to the relative infrequency in 9 such classes, we remove them from the present study. The remaining ten concepts are then taken and the perceptual schema corresponding to these are used for associative language learning: " MAN " (130 out of 376 or 34.57%), " TEMPO " (4.78%), " BUS " (0.80%), " TRUCK " (0.27%, one instance), " TRACTOR " (0.80%), " CAR " (4.79%), "MOTORBIKE " (14.63%), " CYCLE " (11.70%), " RICKSHAW (6.65%)" and " COW " (4.52%).

The simplest model of agent behavior constitutes a clustering in the space of the trajectories, modeled in terms of the temporally ordered centroids. Trajectories are scaled onto time intervals of equal length on which we learn a mixture of Gaussians. The four major trajectory categories obtained by unsupervised trajectory clustering are LEFT TO RIGHT (77 out of 376 or 20.48%), RIGHT TO LEFT (20.21%), FROM-BOTTOM-TURN-LEFT (1.33%) and U-TURN (3.99%). The infrequent categories as well as the outliers are removed from the analysis. The final set of concepts then include ten categories of objects, and four categories of behaviors - thus the set of concepts $\Gamma = \{\gamma_r\}_{r=1}^{n_\Gamma}$ is the set of these learned categories assumed to be available to the language learner. The sensitivities of unsupervised categorization of object appearances (shape features) and actions (trajectories) with respect to the number of clusters is shown in figure 3.



(a) (b)

Fig. 3. The sensitivities of unsupervised categorization w.r.t. the number of clusters for (a) object shape templates (appearances) and (b) object trajectories (actions)

3 Visual Attention and the Perceptual Theory of Mind

Language Learning is largely a social activity, reflected in the *Theory of Mind* hypothesis [19] - that the learner has a model for aspects of the speaker's mind, including a sensitivity to the object being attending to, intentions, belief structures, etc. When the learner is presented with only the visual stream and is not in the presence of the speaker, attention is mediated by visual saliency alone, and not by cues received from the speaker's gaze. In many learning situations where both speaker and viewer are looking at the same scene, this appears to be the case, and we call this the *Perceptual Theory of Mind* – i.e., we assume that the speaker would have attended to those parts of the scene that the learner also finds salient.

Models of Visual Attention involve both bottom-up and top-down processes [10, 20]. While top-down processes are task-dependent, bottom-up processes capture those features of the scene that have the highest payoff in terms of generating conceptual abstractions in most relevant domains. Top-down processes require a conceptual sophistication which is still not available to our pre-lexical learner, and even bottom-up visual attention processes are in the formational process. Nonetheless, we assume a degree of perceptual saliency measures are available to our language learner.

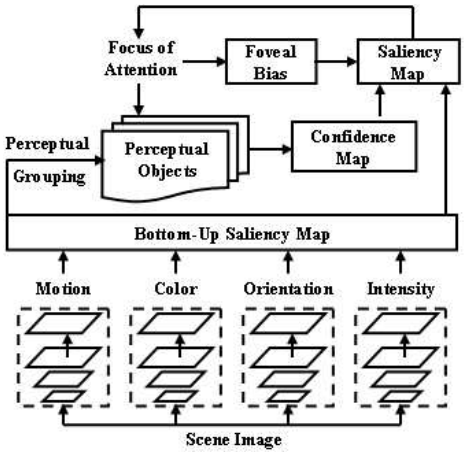
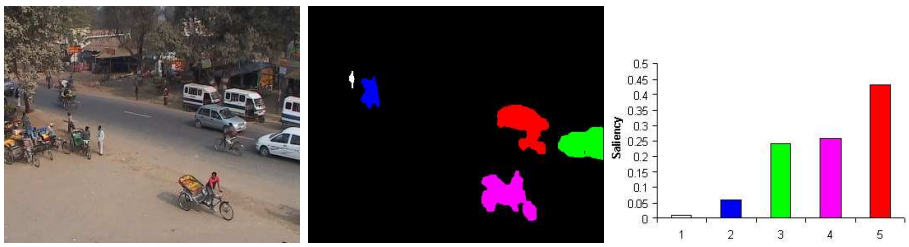


Fig. 4. *Bottom-Up Dynamic Visual Attention Model.* Feature maps for static images (color, intensity and orientation) are extended for motion saliency, computed from an optical flow pyramid. Persistent connected blobs constitute perceptual objects, characterized by shape, appearance, and motion features. Saliency is computed and the focal object identified via winner-take-all. Finally, this fixated object is associated with words from the co-occurring linguistic utterance.

Models for bottom-up attention in static images have been encoded based on multi-scale extraction of intensity, color and orientation contrast feature maps

[10]. This static model has been extended to dynamic scenes [9] by incorporating features for motion saliency (computed from optical flow), and an inhibition of return based on a confidence map reflecting the uncertainty accumulating at image points not visited for some time. A small foveal bias is introduced to favor proximal fixations over large saccades when saliencies are comparable. The saliency map is thus the sum of the feature maps and confidence maps, mediated by the foveal bias, and a Winner-Take-All (WTA) isolates the most conspicuous location for the next fixation. In this work, we use this model of visual attention (Figure 4) to compute the saliency distribution (Figure 5) which indicates the probability of an object being attended.



(a) (b) (c)

Fig. 5. *Saliency distribution of the tracked objects.* (a) Frame 20 : 6 objects tracked in a traffic scene. The blobs associated to the tacked objects are colored in (b) according to their saliency distribution as shown in (c)

4 Learning from Textual Narratives

Data. A group of 18 student volunteers (Indian English speakers, ages $18-25$, 16 male, 2 female) were shown the video and instructed to "describe the scene as it happens" without any further cues about the experimental objectives. Each sentence in the resulting oral narrative was synchronized with the images, and each word in the sentence correlated with the objects under attentive focus in that time span.

The learning task then becomes one of associating conceptual image-schemas γ_r from the set of acquired concepts Γ, with words from the narrative constituting the lexicon Σ. In order to retain generality, we consider k-word concatenations $\sigma_k(l)$ appearing in the narrative; so that $sigma_1$ consists of single words. Thus, from a sentence such as "*Bus moves from left to right*", we would have the set of σ_2 phrases: { "*Bus moves*" , "*moves from*" , "*from left*" , "*left to*" , "*to right*" }.

We now search the set of k-word concatenations $\Sigma_k = \{\sigma_k(l)\}$ for the best match to a co-attentive pre-linguistic concept γ_r. We measure the degree of association between the concept (γ_r) and the l^{th} k-length concatenation $\sigma_k(l)$ using

extremely elementary probability measures: the joint probability $P(\sigma_k(l), \gamma_r)$ and the conditional probability $P(\gamma_r|\sigma_k(l))$. In the absence of sufficient data (most combinations appear too infrequently to compute joint probabilities), we find it productive to use the *conditionally weighted joint probability measure* $\mathcal{J}(\gamma_r, \sigma_k(l))$ given by,

$$\mathcal{J}(\gamma_r, \sigma_k(l)) = P(\gamma_r|\sigma_k(l))P(\sigma_k(l), \gamma_r) \tag{1}$$

Also, the probability of longer concatenations needs to be normalized by the probability of k-length sequences, but given the very small sample of text, this cannot be computed reliably and we make the weak assumption that this likelihood is inversely proportional to the segment-length $(1/k)$, so that k-word strings have their probability multiplied by k. The association measure is a very small fraction and results reported in tables 2 and 3 are multiplied by 1000.

While our results are limited to this particular scene, we assume that the learning agent is also exposed to other contexts. Thus, it is likely that the more common words (the, of, etc.) have been encountered in many other contexts - thus their conditional probabilities are low. For single word matches ($k = 1$), we discount the hundred most common words (based on the Gutenberg corpus).

4.1 Association Results

The set of concepts available include ten object categories and four trajectory categories. For all concepts, utterances co-temporaneous with attentive focus result in correlations with all words in the utterance. Concepts that have very strong (frequent) associations are likely to be learned earlier.

Our narrative shows a preponderance of motion / trajectory words - most frequent is the word *left* (447 instances) followed by *right* (387). Next, generic motion verbs such as *moving* (128) and *going* (126) overwhelm the first nouns - *bike* (111), *car* (81), etc.

In the following, we start to learn word associations (using phrase-length $k = 1 \ldots 4$) for both the trajectory and the agent concepts. Immediately we discover that motion concepts are learned adequately at the $k = 3$ level (Table 2), whereas object labels are overwhelmed by trajectory descriptors like "left" or "going". Based on the mutual exclusivity principle [21], the early learner assumes that different labels apply to different concepts - and therefore, having learned the motion words, we drop the learned tags from the lexicon before proceeding to learn the object labels. Inverting this order, attempting to learn the objects first, results in a weaker correlation, e.g. the term "cycle" fails to get a high association. This is of course atypical, since infants learn the first object labels somewhat earlier than the first motion labels; it is no doubt an idiosyncracy of the traffic scene where motions are preponderant.

For trajectory labels, single word tags such as "left" or "right" have weaker associations, and multi-word concatenations, "left to right" and "right to left", emerge with the strongest association for the concepts LEFT TO RIGHT and RIGHT TO LEFT. The categories FROM-BOTTOM-TURN-LEFT and U-TURN have very few

Table 2. Associating Language Labels to object Behavior (Trajectory)

LEFT-TO-RIGHT		RIGHT-TO-LEFT		FROM-BOTTOM TURN-LEFT		U-TURN	
ONE WORD LONG LINGUISTIC LABELS ($k = 1$)							
left	1.609	left	10.61	left	0.021	bus	0.66
to	1.471	to	9.441	person	0.018	left	0.36
right	1.334	right	8.715	the	0.017	to	0.30
moving	0.836	the	6.841	cycle	0.016	right	0.27
the	0.807	moving	4.991	to	0.013	from	0.24
TWO WORD LONG LINGUISTIC LABELS ($k = 2$)							
to left	1.598	to left	13.01	gate and	0.056	bus coming	0.80
to right	1.422	right to	11.17	man in	0.040	bus comes	0.65
left to	1.368	to right	6.124	walking with	0.038	the Vikram	0.50
right to	1.312	from right	6.078	and turns	0.038	from left	0.44
from left	0.858	left to	5.894	IIT and	0.038	entering from	0.40
THREE WORD LONG LINGUISTIC LABELS ($k = 3$)							
left to right	2.751	right to left	23.33	gate and going	0.105	entering the Vikram	0.96
right to left	2.607	left to right	11.45	person moving left	0.078	entering from the	0.77
moving right to	1.194	from right to	8.421	and turns left	0.072	left to right	0.75
from left to	0.960	moving right to	5.910	IIT and turns	0.072	tempo moves towards	0.67
white car moving	0.921	to the left	5.166	of IIT and	0.072	in tempo moves	0.67
FOUR WORD LONG LINGUISTIC LABELS ($k = 4$)							
moving right to left	2.364	from right to left	15.28	gate and going to	0.184	lady entering the Vikram	1.71
from left to right	1.744	moving right to left	12.07	person moving left to	0.140	bus coming from left	1.32
from right to left	1.540	going right to left	9.996	IIT and turns left	0.124	tempo moves towards the	1.22
moving left to right	1.384	going from right to	7.368	of IIT and turns	0.124	in tempo moves towards	1.22
white car moving right	1.240	from left to right	7.256	out of IIT and	0.124	comes in tempo moves	1.22

instances and may require more observations before they can be learned. After removing the trajectory labels from the set of words for object (noun) learning, we find single-word results outweigh multi-word text, and only single word results are reported (Table 3).

Table 3. Labels for object Concepts (word set $\Sigma - \{left, to, right\}$)

MAN		TEMPO		BUS		TRUCK		TRACTOR	
moving	6.613	going	4.985	bus	0.081	lorry	0.141	tractor	0.046
going	6.280	moving	4.814	state	0.017	truck	0.046	loaded	0.019
motorbike	5.817	tempo	4.571	govt.	0.015	going	0.008	green	0.013
cycle	3.284	motorbike	3.057	big	0.010	motorbike	0.008	stuff	0.011
two	2.992	bus	3.018	exits	0.009	moving	0.007	fully	0.010
CAR		MOTORBIKE		CYCLE		RICKSHAW		COW	
moving	1.488	moving	0.809	cycle	1.509	going	1.144	two	0.045
going	1.287	going	0.528	moving	1.429	moving	1.063	motorbike	0.041
motorbike	1.125	car	0.499	going	1.180	rickshaw	0.736	moving	0.041
car	1.054	motorbike	0.475	two	0.752	motorbike	0.680	tempo	0.037
coming	0.760	coming	0.373	tempo	0.669	car	0.669	going	0.030

4.2 Discussion

Some labels are easier to learn compared to others for several reasons. First, there are instances of *Synonymy*, e.g. a concept like MAN can have labels *people, sardarji, person, guys, guy* etc., diluting the effect of any particular label (we do not remove plurals or do any kind of morphological processing on the text). This is true also for CAR and for TEMPO. Secondly, our computational *Visual Saliency* model may not have selected the objects mentioned in the narrative. This is particularly true of people, who are preponderant in the scene but are not selected either in the narrative nor by the visual focus. When they do appear in the narrative, they are sometimes not in attentive focus, and we see that for the category MAN, no relevant label appears in the top five. On the contrary, MOTORBIKEs are mentioned quite frequently, but are not as frequently in attentive focus, and given the preponderance of objects (varying between five and twenty at any time), MOTORBIKE emerges as one of the high contenders for several concept categories. On the other hand, large objects like truck, which appeared only once, despite two equal synonyms (*truck* (11), *lorry* (9)), have both these labels at the top of the list. This is due to the high visual saliency of this large moving region; the same may also hold for BUS. Finally, there are issues related to the *Categorization Level*, i.e., the narratives may refer to objects at a subordinate (or superordinate) level. Thus, the concept CAR is referred to by model names such as *maruti, Sumo, Zen* as well as *taxi, van, car, cars* etc. There are also eight instances of the superordinate "vehicle" being used. Clearly,

a much richer characterization of objects and their subcategories would need to be learned before these distinctions can be mastered.

To reiterate the main results - this work represents a completely unsupervised process relying on visual attention to parse the visual input. Place the camera at the scene, and observe the goings on for about half an hour. At some point, have some adults comment on what is happening, and even with very primitive statistical association measures, our infant learner is able to build mappings for six new words/phrases. We feel that given the enormous prior knowledge deployed in many computational learners, this is not bad going at all for our infant learner.

5 Conclusion

In this work, we have presented a model that acquires concepts of object shape and appearance, as well as actions, from complex multiagent videos. Despite the complexity of the input, we demonstrate that some of these concepts can then be successfully associated with word labels. The same task, if performed with simultaneous concept and language acquisition, is considerably more difficult (e.g. see [5] on prepositions). More importantly, such a procedure ignores any possible perceptual abstractions that may have formed in the first year and a half of life. While this does not rule out any other alternatives, it provides some computational weight for the perceptual-conceptualization position.

To our knowledge, this is the first work that takes a complex visual scene, identifies a number of concepts in a completely unsupervised manner, and then associates these with unedited text inputs, to obtain a few phonetic to perceptual schema mappings. The main burden of computation in this task is in the visual processing - i.e. the visual concepts may be harder to learn than (at least some) of the linguistic mappings.

Another key outcome is that some insight has been gained into the phrase "image schema", which has been used in a wide variety of meanings e.g. [7] presents a linguistic perspective and [1] a perceptual view. Our approach provides a plausible computational approach to constructing image-schemas from real perceptual data. These are internalized as probability distributions ranging over spatio-temporal manifolds. In our model, we find that certain image schemas have correlations that may already be viewed as symbolic arguments - e.g. we discover that action concepts such as *left-to-right* or *right-to-left* involve a single moving object in an agentive role. Thus, their valency (a grammatical notion related to the number of arguments a verb takes in a sentence), is determined from these semantic considerations, and in the long run this may provide a semantic basis for many considerations in syntax.

Such models clearly have immediate application value in visual surveillance; the user has only to describe a few scenes, and it would be possible to then identify salient aspects of the scene and code these in future encounters with these objects.

As for syntax, it is tempting to claim that the approach is oblivious to syntactic (and morphological) niceties, but it is important to remember that we are learning primarily motion descriptors and nominals in a weakly inflected language. In case-rich languages, the learning rate would surely be slower, and some prior morphology learning may be needed before learning most of the grounded nouns. This is true even for child learning, as attested in Turkish vs English learners [22].

While our approach is rich in terms of perception, the learner is not an active participant in the scene. Thus crucial aspects such as intentionality, purposive action, and social interaction have been ignored in the present study. While some amount of language learning may involve passive inputs, contingent interaction is undoubtedly a powerful force that would be important to explore in following work.

While the specific appearance models are indexed upon the specific view, the object classes per se, as well as the occlusion-based interaction primitives, are more general and can be applied to novel situations. It would be important to consider the correlations between multiple views in constructing the appearance models, so that all canonical views can be covered.

Finally, while we have used attentive focus to associate visual concepts with words, we have not used attention at all for the task of forming conceptual clusters. The use of attention for learning concepts is significant since the learned concepts can then act as top-down mediators and bring in elements of intentionality into the system. On the whole, such associative maps for word meanings are clearly just the first step - the vast majority of adult vocabularies are acquired by extrapolation from a few grounded words, primarily by reading[19]. However, these first grounded words constitute the foundation on which these other meanings can be anchored.

References

1. Mandler, J.M.: Foundations of Mind. Oxford University Press, New York (2004)
2. Jones, S.S., Smith, L.B.: The place of perception in children's concepts. Cognitive Development **8** (1993) 113–139
3. Mandler, J.M.: A synopsis of the foundations of mind: Origins of conceptual thought. Developmental Science **7** (2004) 499–505
4. Barsalou, L.W.: Perceptual symbol systems. Behavioral and Brain Sciences **22** (1999) 577–609
5. Regier, T.: The Human Semantic Potential: Spatial Language and Constrained Connectionism. Bradford Books (1996)
6. Roy, D.K., Pentland, A.P.: Learning words from sights and sounds: a computational model. Cognitive Science **26** (2002) 113–146
7. Langacker, R.: Foundations of Cognitive Grammar, vol. 2: Descriptive Application. Stanford University Press, Stanford, CA (1991)
8. Quine, W.V.O.: Word and Object. John Wiley and Sons, New York (1960)
9. Singh, V.K., Maji, S., Mukerjee, A.: Confidence based updation of motion conspicuity in dynamic scenes. In: Third Canadian Conference on Computer and Robot Vision (CRV' 06). (2006)

10. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Reviews Neuroscience **2** (2001) 194–203
11. Coldren, J.T., Haaf, R.A.: Priority of processing components of visual stimuli by 6-month-old infants. Infant Behavior and Development **22** (1999) 131–135
12. Yu, C., Ballard, D.H.: A multimodal learning interface for grounding spoken language in sensory perceptions. ACM Transactions on Applied Perception (2004)
13. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135
14. Siskind, J.M.: Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. J. of Artificial Intelligence Res. **15** (2001) 31–90
15. Steels, L.: Evolving grounded communication for robots. Trends in Cognitive Sciences **7** (2003) 308–312
16. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition. Volume 2. (2004) 28–31
17. Proesmans, M., Gool, L.V., Pauwels, E., Osterlinck, A.: Determination of optical flow and its discontinuities using non-linear diffusion. In: The 3rd Eurpoean Conference on Computer Vision. Volume 2. (1994) 295–304
18. Guha, P., Mukerjee, A., Venkatesh, K.: Spatio-temporal discovery: Appearance + behavior = agent. In: Fifth Indian Conference on Computer Vision, Graphics and Image Processing. Volume LNCS 4338. (2006) 516–527
19. Bloom, P.: How Children Learn the Meanings of Words. MIT Press, Cambridge, MA (2000)
20. Rothenstein, A.L., Tsotsos, J.K.: Attention links sensing to recognition. Image and Vision Computing (2006) 1–13
21. Regier, T.: Emergent constraints on word-learning: A computational review. Trends in Cognitive Sciences **7** (2003) 263–268
22. Stromswold, K.: The cognitive neuroscience of language acquisition. In (ed.), G., ed.: The new cognitive neurosciences. MIT Press, Cambridge, MA (1999) 909–932