Perceptual Theory of Mind: An intermediary between visual salience and Noun / Verb Acquisition

Amitabha Mukerjee and Mausoom Sarkar

Dept of Computer Science and Engineering, IIT Kanpur, India
{amit,mausoom}@cse.iitk.ac.in

***Abstract*— We present computational models based on visual attention that learn object-name mappings and action semantics from simple 2D multi-agent visual streams co-occurring with word-separated utterance streams. We use no perceptual priors and the nominals are acquired by an early learner who has no syntactic knowledge. A late learner then uses the knowledge of nominals to identify actions referring to these arguments, and acquires the semantics of motion verbs like *run* or *chase*. Both early and late learners use visual attention to determine which parts of the scene are salient at the time of the utterance; we use a synthetic model of dynamic visual attention. Simple statistical measures based on joint probability are found sufficient to identify nominal participants from word separated input text, and a simple recurrent network is used to learn the verb semantics that encodes fine-grained image schemas as well as the argument structure as part of the semantic model.**

***Index Terms*— Multimodal Learning, Grounding, Gaze Prediction , Focus of Attention**

I. Introduction

This paper presents computational models based on visual attention that support two claims of developmental learning:

- that nominals can be acquired from word-separated language utterances without any knowledge of syntax.
- that motion verbs can be acquired directly from perceptual sequences of motion features in temporally referenced semantic schemas. This process also encodes their argument structures as part of this semantics.

We consider a language learner acquiring grounded meanings for words at two stages in the acquisition of language. The object name learner (early) acquires nominals solely from word-object correlations in word-separated utterance streams, without regard for syntax. The verb learner (late) is aware of the difference between actions involving single vs multiple participants (intransitive / transitive verbs) - a sensitivity that may be present in children as early as 11 months, which is also the time when they become aware of word boundaries [11].

Here, the learner is not in the presence of the speaker, and cannot follow cues from the gaze of the speaker to determine attentive focus. Instead, it is assumed that the learner realizes that her attentive mechanisms are similar to that of the speaker, a hypothesis we call the *Perceptual Theory of Mind*. Computational mechanisms of visual attention [18]. are then used to constrain the region of visual computation, and identify the constituents participating in an action.

There is some agreement that language learners can acquire some nominals from language usage based on correlation alone, e.g. Bloom ([2], p. 198): “Syntax is not necessary for at least some nominal learning.”. But how does this work? In computational models, models for noun learning tend to look at single objects [15], or using single words [17] which are then easily correlated with perceptual precepts. But most words are learned from multi-object scenes and not accompanied by one-word labels but by words appearing in usage contexts. In this direction, [1] use a head mounted camera on the narrator and incorporate gaze, head and hand movements in their model for grounded word acquisition. But here also, a single object is the focus in the utterance.

For our work, we take as input a video made famous in social psychology by Heider and Simmel [7]. The co-occurring text were collected as part of an experiment on how users segment events into hierarchical subtasks [10]. In this task, users were asked to segment the actions in the scene and also to describe the action in an unconstrained narrative. Consequently, the linguistic input has the wide variety expected in multiple articulations for the same scene (see Table I below). We use the word-separated text directly, and also the event boundaries that were attested in their experiments. We then correlate words co-occurring with perceptually salient objects to learn the object names, and despite this variation in the input, we find that nominals are easily associated with their object tokens.

While we are able to learn nominals without syntax, the extent to which syntax informs the learning process in verb learning has been the subject of much debate ([6], [14]). Without entering substantially in this debate, we assume that the learner is able to learn the syntactic distinction between intransitive and transitive verbs based on the number of participants (arguments) in the corresponding actions. These correspond to different feature-sets used in the learning task - *monadic* (only one agent) or *dyadic* (two-agent interactions). Using simple recurrent networks (SRNs), we are able to capture the fine-grained temporal semantics for actions such as *chase* or *run* in terms of perceptual image schemas. Unlike other work on linking video to language that assumes a set of action priors such as those involving contact etc, which are characterizable a priori, - ([5], [3], we only assume that the feature set is available to the learner, possibly as a part of

earlier perceptual categorization processes.

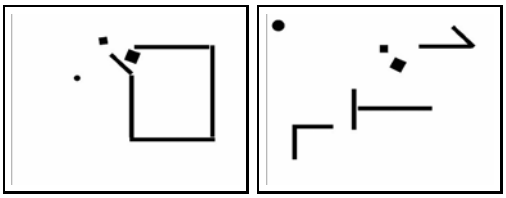

Fig. 1. *Input Videos.* Chase sequence (derived from [7] and Hide and Seek sequence both created by Bridgette Martin [10]). The same three agents, "big square", "small square" and "circle" participate in two activities in two different spaces.

Start Frame	End Frame	Subject One	Subject Two
617	635	the little square hit the big square	they're hitting each other
805	848	the big square hit the little square	and they keep hitting each other
852	1100	the big square hit the little square again; the little circle moves to the door; the big square threatens the little circle	now the circle is blocking the entrance for the big square; now the circle is inside the square
1145	1202	the big square goes inside the box; (and) the door closes	another square went inside the big square

TABLE I

Description of the Events by Subjects. DIFFERING STATEMENTS BY TWO SUBJECTS IN THE [*Chase* VIDEO]

II. SYNTHETIC MODELS OF VISUAL ATTENTION

Computational models of Visual Attention involve bottom-up and top-down processes. While top-down processes vary depending on task requirements, bottom-up aspects are more stable and have been encoded for static images [8] based on parallel extraction of intensity, colour and orientation contrast feature maps. Colour and intensity contrast maps are obtained as feature pyramids (maps at different scales), along with center-surround maps (multi-scale difference of feature maps). The center-surround feature processing is similar to the difference of gaussian convolved images (DOGs). For orientation specific processing, gabor filters are used with different frequencies and at different scales to generate the orientation specific feature map.

The static model, which replicates saliency map structures likely to be present in the LGN or V1 regions of the mammalian cortex, is extended here to model dynamic scenes based on motion saliency. Motion saliency is computed from the optical flow, and a confidence map is introduced to record the uncertainty accumulating at scene locations not visited for some time. A small foveal bias is introduced to mediate in favour of proximal fixations as opposed to large saccadic motions. The saliency map is 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. The overall architecture is highlighted in "Fig.2".

Perceptual Theory of Mind

The Theory of Mind hypothesis [2] holds that the learner has a model for several aspect of the speaker's mind, at various levels from a sensitivity to the object being attending to, to belief structures (e.g. children under three are found to be incapable of entertaining false beliefs). In this work, we focus at the lowest end of this spectrum, and focus on what we call the *Perceptual Theory of Mind*. While much of the Theory of Mind work has focused on gaze following based on cues from the speaker's eyes or her gaze direction, the Perceptual Theory of Mind makes a much weaker claim: in the absence of direct cues from a speaker, it assumes that the speaker would have attended to those parts of the scene that the learner also finds salient. This is probably a valid assumption for children from the age of six months onwards [4], although the mechanisms for perceptual salience are themselves being developed at this stage. In our work, we do not specify a particular development status for our learning agent, but assume this model to infer that the scene objects being attended to by the agent were also salient for the speaker at the moment of utterance.

Language Acquisition experiments tend to cast doubts on the efficacy of a purely associationist model of learning words, and it is true that a large percentage of our vocabulary is not learned using multimodal inputs but from reading. Nonetheless, this work presents some evidence that for the beginning learner, multimodal associations mediated by attentional processes provide strong and reliable cues for learning nominals and their properties, verbs and their argument and event structures.

III. EARLY LEARNER STAGE : NAME LEARNING

Data. Two video sequences were used in the experiments. Both show two squares and a circle moving in a 2D space "Fig.1". The first sequence (*Chase*) shows a room with a door, and a large square that chases the other two. The second (*Hide & Seek*) shows a game of hide & seek.

Grounded semantics for nominals appearing in the text are associated with the corresponding image object token based on the following steps:

1) *Tracking and Recognition*. Object recognition is based on shape.
2) *Synthetic Gaze Prediction* Estimates the entities that are being attended to during a particular utterance, under the perceptual theory of mind assumption.
3) *Associative Learning*. The words in the commentary are associated with their perceptual correlates.

A. Tracking and Shape-based recognition

Spelke [16], among others has demonstrated that infants perceive objects as connected blobs with coherent motion. In our videos, these blobs are rather simple, and these are extracted and tracked for the duration of the video. A global list of all the objects and their pose and orientation is used for computation of motion primitives.

Shape recognition is required to obtain shape universals for different object tokens with the same perceptual signature.

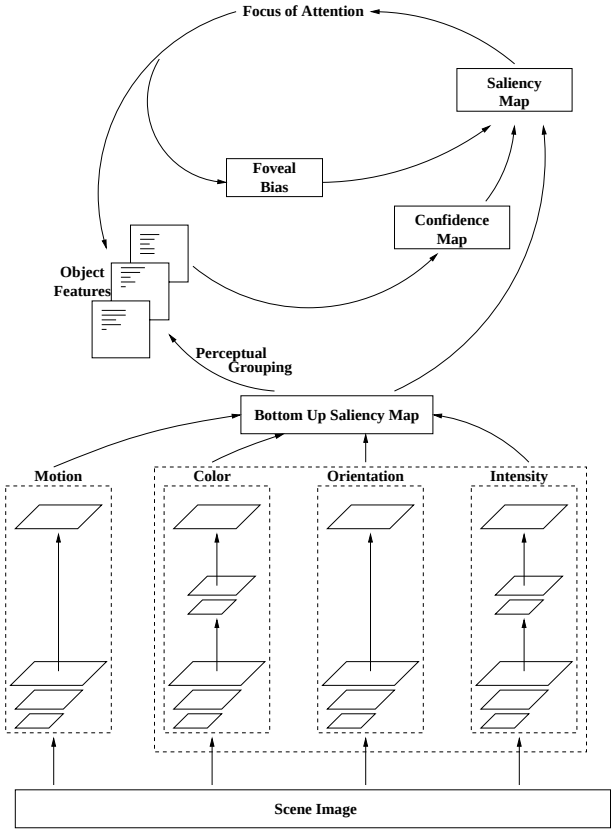

Fig. 2. *Bottom-Up Dynamic Visual Attention Model*. The feature maps for static images (colour, intensity and orientation) are extended with a motion saliency map (based on optical flow). In addition a confidence map records which sites have not been visited for a longer time. Winner-Take-All determines the next fixation.

This is important in both correlating objects in the same scene (squares) and also across different perceptual input situations, as in combining word associations across the two separate videos. Shape matching for 2D objects is implemented based on a histogram of the tangent direction at each point along the boundary, a simple scale, rotation and transformation invariant metric that is cognitively plausible. The normalized (scale invariant) histograms were circularly shifted and compared using statistical divergence function as in Roy [15] to determine the closest transformation between two shapes, which serves as our model of shape similarity.

$$d_v(X,Y) = \sum_i \frac{(x_i - y_i)^2}{(x_i + y_i)}$$

where $X = \bigcup_i x_i$ and $Y = \bigcup_i y_i$ are two histograms and x_i , y_i are the values of a histogram.

B. Gaze prediction

The synthetic model of visual attention for dynamic scenes was used to predict the gaze for the two videos. The predicted gaze for the two videos are shown in "Fig.3".

C. Association of meanings with words

The attended objects are now associated with the temporally correlated words using one of two probability measures. At this point, we also assume that the learner has been exposed to other linguistic fragments before this, so that words like "the" and "is" are known to be more general than this discourse context, and are not applied to this situation. (In the BNC, "the" occurs 1500 times more frequently than "square", say). Using perceptual equivalence relations based on shape, we associate the objects with words from the multiple narratives using the probability measures outlined below.

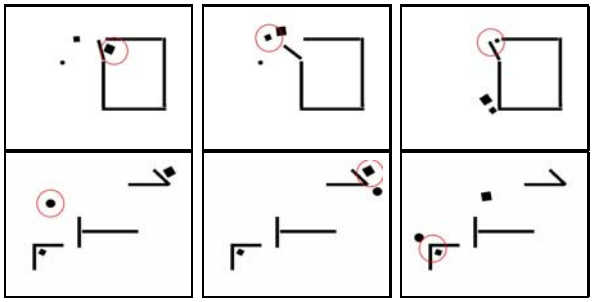

Fig. 3. *Focus of attention on the video*. Red circles represent the focus of attention; Oscillating attention between multiple objects makes multiple participants more likely.

Two measures for associating the words with the objects are used.

1) Mutual Information Measure.
2) Joint Probability Measure.

1) Mutual information Measure: After temporally correlating words with objects the association is defined as the product of mutual information of word w_i and object o_j with their joint probability.

$$A = \Pr(w_i, o_j) \log \frac{\Pr(w_i, o_j)}{\Pr(w_i)\Pr(o_j)}$$

If W and O ($W = \bigcup_i w_i$ and $O = \bigcup_i o_i$) are two random variables then their Mutual Information $I(W,O)$ would be

$$I(W,O) = \sum_i \sum_j \Pr(w, o_j) \log \frac{\Pr(w_i, o_j)}{\Pr(w_i)\Pr(o_j)}$$

where $\Pr(w_i, o_j) \log \frac{\Pr(w_i, o_j)}{\Pr(w_i)\Pr(o_j)}$ is the contribution of each word object pair.

2) JointProbability Measure: This measure is the product of the conditional probability of object o_j given word w_i with their joint probability.

$$A = \Pr(w_i, o_j) \Pr(o_j | w_i)$$

The conditional probability will give high values for the words that occurred specifically for a given object, while joint probability will highlight the number of word object pairs. Hence this measure should bring out the stronger labels associations.

Both the measures showed object names being highly associated with the corresponding object. Results are shown for the chase video and both videos together "Fig.4 and Fig.5" respectively.

IV. Late Learner Stage : Action Labels

In the next stage, we posit a late learner with limited syntactic knowledge, who then uses the learned entity names as

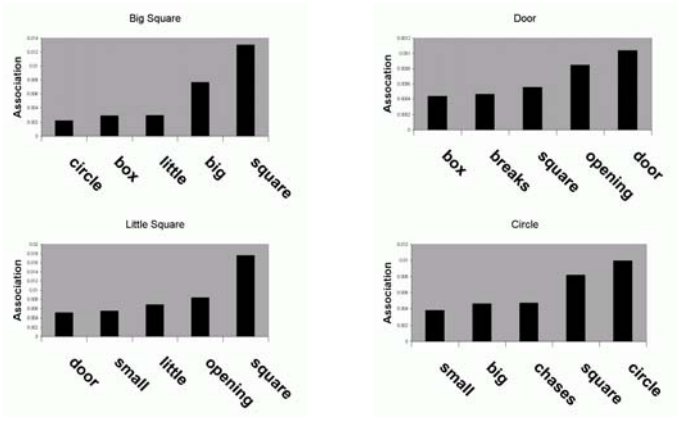

Fig. 4. *Early Learner: Noun Learning using First Encounter*. Association using Joint Probability of words with Big Square, Door, Little Square and Circle in Chase video. Most objects except "little square" are well characterized.

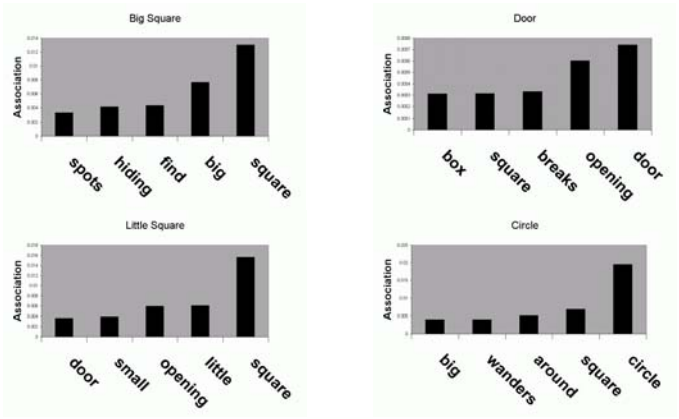

Fig. 5. *Noun Learning using Multiple Encounters*. Associating words using Joint Probability - remembering shapes across two encounters (both videos).

arguments in a k-ary predicate structure, based on the valences of the verbs appearing in the text. Actions are pre-linguistically categorized using different k-ary feature sets, depending on the number of participants in the action, using a recurrent neural network (SRN) [12]. This also results in discovering the valence of the verb.

A. Features for Spatio-Temporal Analysis

The feature set determines the dimensionality of the space in which the word meaning is grounded. The valence of the predicate is a crucial input for this information, it is seen that a *monadic* feature set is sufficient for actions like run (intransitive), whereas a *dyadic* feature set is needed in actions like chase (transitive).

There is considerable evidence that infants have pre-linguistic perceptual notions for Path concepts such as source, trajectory etc, and also other notions such as Up-Down, Containment, Force, Part-Whole and Link [9]. Some of these aspects have also been implemented in computational systems [13]. In this work, we assume that the pre-linguistic visual system has the capability to abstract the following:

(a) Shape classes (square vs circle, on a high contrast image)
(b) Motion characteristics for individual agents (monadic features)
(c) Motion characteristics for pairs agents (dyadic features)

Specifically, we define the following abstract features:

- *Monadic Features:*
  1) Velocity- vx and vy in respective direction
  2) Acceleration- ax and ay in respective direction
  3) θ- Angular displacement of the object
  4) $d\theta$- Change in displacement of the object.
  5) ω - Angular velocity of the object.
  6) α - Angular acceleration of the object.
- *Dyadic Features*
  1) Proximity- It is inverse of the boundary-to-boundary distance between two objects.
  2) Relative Velocity between the two objects
  3) Relative Acceleration between the two objects
  4) Measure of Parallelism between the direction of motion of the two objects i.e. $\cos(\hat{v_a} - \hat{v_b})$
  5) Leader A - Measure of leadership (in motion) of A w.r.t B i.e $\cos(\hat{v_a} - \theta_{ba})$
  6) Leader B - Measure of leadership (in motion) of B w.r.t A i.e. $\cos(\hat{v_b} - \theta_{ab})$
  7) Chamfer- Measure of chamfering between the two objects i.e. $\frac{\sin(\hat{v_a} - \theta_{ba}) + \sin(\hat{v_b} - \theta_{ab})}{2}$

For example, an action such as "X chases Y" may associate *Proximity* (indicative of spatial clustering between X and Y), *Parallelism*, high *Leader X* and low *Relative Velocity*. The dyadic parameter *Chamfer* feature reflects if two objects are moving together or are moving one ahead of the other this feature, when low, may indicate a follow or chase action; when close to 90, it is indicative of a move-together action.

B. Similarity Clustering of Verbs

Action intervals in the video are surprisingly consistent across viewers [10], and we use these boundaries for denoting temporal intervals for actions that are to be labelled. The diverse statements of the experiment participants are clustered. Lexical units occurring extremely infrequently (just one instance) are removed from consideration, and pairs of head-verbs that are used consistently in the same interval, with the same set of agents as arguments, by different subjects, are considered as synonymous, with suitable changes in the agent order as necessary. The cluster of such head-verbs was labeled using the most frequent lexical unit, resulting in the clusters. The objective was to acquire the semantics for these actions.

C. Verb Learning Results

Assuming that each video sequence is of l duration during which a verb is reported for τ time units and not reported for $l - \tau$ units. The Detection Accuracy, or true positives, is computed as a percentage of positive classifications out of τ. The False Negatives, again, are computed as a percentage of

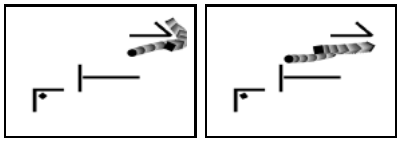

Fig. 6. *Big Square* ***Chases*** *circle in the Hide and Seek video.* 1 Second intervals

.

τ. However, the false positives as well as focus mismatches are computed as a percentage of $l - \tau$. Results are presented for three learning scenarios:

1) *One-Verb-One-Network:Human Subject Tags*. Here Different SRNs are trained for each action cluster; Table II presents the results with unsupervised text and image correlations based on the statements of human subjects. "Fig.9" shows some of the video fragments which are classified by the learning system as chase.

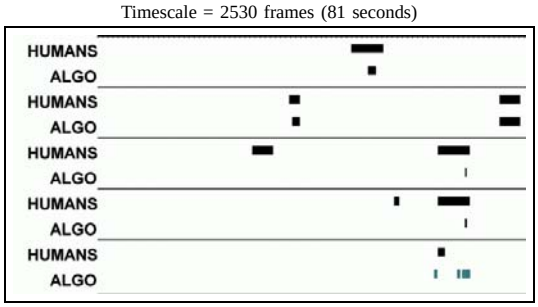

Fig. 7. *Comparing machine and human classifications,* for verb:*chase*. Each row is a different pairwise combination of agents.

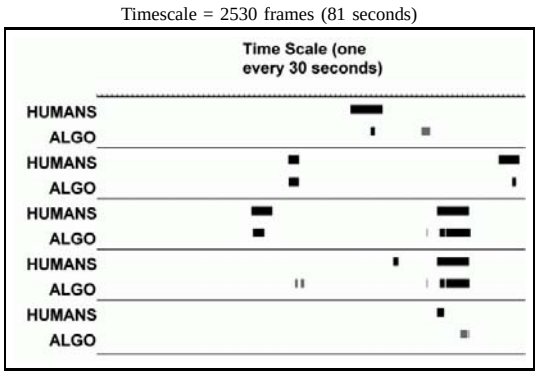

Fig. 8. *Comparing outputs - trained on synthetic data.* One-Verb-One-Network for *chase* video on Synthetic Data.

"Fig.7 ,Fig.8 and Fig.10" present results along a time line, each row reflects a different combination of agents (small square, big square, circle). Dark gray color indicates false positive classification, while light gray color indicates focus mismatches.

2) One-Verb-One-Network: Synthetic Data: To overcome the very severely limited data in the videos, a synthetic video with only two agents was created, executing canonical versions of the actions over 2660 frames. These may be thought of as pre-linguistic categorization

Verb	TP	FP	FN	FM
hit	3	3	1	1
chase	6	0	3	4
come Closer	6	20	7	24
move Away	8	3	0	14
spins	22	0	1	9
moves	5	1	2	7

where TP=True Positive,FP=False Positive,FN=False Negative and FM=Focus Mismatch

TABLE II

Interval results. ONE-VERB-ONE-NETWORK, UNSUPERVISED

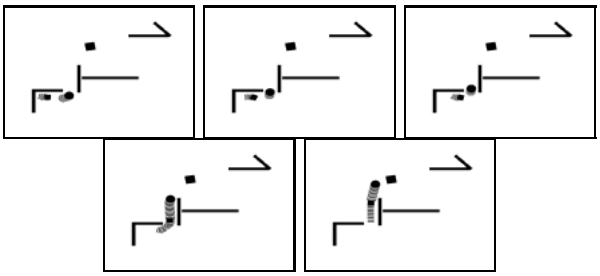

Fig. 9. Example of classification of **chase** event, one for every 1 Second (30 frames), between the circle and the small square

of the actions. The system is trained on the synthetic data, and tested on the original human-annotated video; Table III shows the results.

Verb	TP	FP	FN	FM
chase	52.59%	0.21%	47.09%	1.24%
come Closer	53.61%	11.29%	45.52%	20.41%
move Away	65.30%	12.07%	33.37%	17.15%

where TP=True Positive,FP=False Positive,FN=False Negative and FM=Focus Mismatch

TABLE III

PERCENTAGE RESULTS FOR ONE-VERB-ONE-NETWORK SYNTHETIC DATA

Between the two scenarios - learning from one pair of agents in a human-tagged data vs learning from elaborately created synthetic data, it is clear that having more data improves the classification accuracy. However, since the training set has only two agents and has no focus mismatch, the extent of this problem is actually higher when trained on synthetic data.

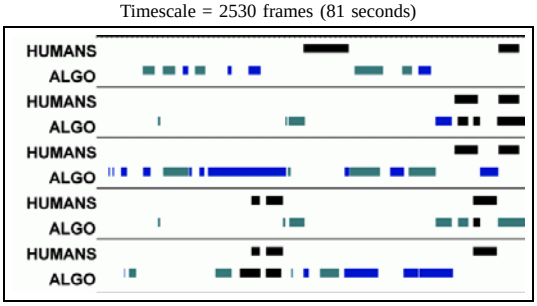

Fig. 10. A comparison of system and human descriptions of *come closer* verb over a time line. Note the large extent of focus mismatch (light grey).

3) All-Verbs-One-Network: Synthetic Data
Here we train a single recurrent network to distinguish among all the verbs, and use the synthetic video for training. Only dyadic features are used. There is one input neuron for each feature, one output neuron for each verb and the hidden layer has the same number of nodes as the input layer. An additional output neuron indicates no output.
The accuracy of events detected improves considerably if we constrain the sequence based on the participating agents as well as the verb. The Interval accuracy is computed as a ratio of correctly classified intervals to the total intervals for each pair of objects. Table IV lists the performance of the system in this mode.

Video	Total Frames	Objects involved	Frames Classified	Correct Classification	Interval Accuracy %
Hide & Seek	2530	SS, C	2492	2380	94.07
		SS, BS	2329	2156	85.22
		BS, C	2530	2028	80.16
Chase	2530	SS, C	2334	2285	90.30
		SS, BS	2329	2156	85.22
		BS, C	2384	2207	87.23

SS = Small Square, BS = Big Square, C = Circle. Percentage accuracy computed over total number of frames and not on number of classified frames.

TABLE IV
FRAME-BY-FRAME INTERVAL ACCURACY RESULTS FOR ALL-VERBS-ONE-NETWORK.

V. CONCLUSION

In this work we make the assumption that the learner observing a visual scene identifies the objects she is attending as those also being attended to by the speaker. This makes it possible for us to use a computational model of dynamic vision that identifies the participants in the co-occurring text.

The present system makes a number of assumptions on the learning process. The set of features in verb learning are presumed to be available to the learner, and are not learned per se. While one may argue that some of these notions (such as size or motion trajectory information) may be preferred as a part of the innate perceptual apparatus, we strongly suspect that some of the relative motion parameters may actually be learnable given enough training data, and we hope to explore this with the additional data.

In this work, no attempt was made here to learn the syntactic structure of the statements per se, and effects such as tense, case, gender etc are not modeled at all. An important extension would be to learn more grammatical structures as in the work of [5]. In our case, once the verbal heads of various phrases are known, and with some knowledge of closed-class words, it would be possible to identify some of the roles played by grammatical elements in different constructions appearing in the narrative.

A particularly attractive line of inquiry following this work is that of delineating the spatial primitives involved as adjunct modifiers to the nominals and verbal phrases in this very text. These modifiers are important cues to the action, and are also strongly related to the gaze fixation model adopted here, and we are actively looking into modeling the semantics of these spatial correlates in the coming months.

VI. ACKNOWLEDGMENTS

We thank Bridgette Martin & Barbara Tversky for initial discussion regarding the idea behind this paper. The videos and the text were provided by Bridgette Martin. We also thank Vivek Singh for his visual attention model, and Dana for his help in developing it. V. Shreenivas and Achla Raina helped in the verb acquisition work.

REFERENCES

[1] Dana H. Ballard and Chen Yu. A multimodal learning interface for word acquisition. In *International Conference on Acoustics,Speech and Signal Processing(ICASSP03)*, volume 5, pages 784–7, april 2003.
[2] Paul Bloom. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA, 2000.
[3] Alan P Fern, Robert L Givan, and Jeffrey Siskind. Learning temporal, relational, force-dynamic event definitions from video. In *Proceedings of AAAI*, pages 159–166, 2002.
[4] J.H Flavell. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly*, 50:274–29, 2004.
[5] Dominey Peter Ford. Learning grammatical constructions in a miniature language from narrated video events. In *cognitive science*, 2003.
[6] Lila R. Gleitman and Jane Gillette. The role of syntax in verb learning. In Paul Fletcher and Brian MacWhinney, editors, *Handbook of Child Development*, chapter 15, pages p.413–427. Blackwell, 1995.
[7] F. Heider and M. Simmel. An experimental study of apparent behavior. In *American Journal of Psychology*, volume 57, pages 243–59, 1944.
[8] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, Pasadena, California, Jan 2000.
[9] Jean M Mandler. How to build a baby ii. conceptual primitives. *Psychological Review*, 99:587–604, October 1992.
[10] Bridgette Martin and Barbara Tversky. Segmenting ambiguous events. In *Proceedings of the 25th annual meeting of the Cognitive Science Society*, 2003. Crucial for our Data-Collection chapter.
[11] Roberta Michnick-Golinkoff and Kathy Hirsh-Pasek. Reinterpreting children's sentence comprehension: Toward a new framework. In Paul Fletcher and Brian MacWhinney, editors, *Handbook of Child Development*, chapter 16, pages 430–461. Blackwell, 1995.
[12] Amitabha Mukerjee, Pradeepsinh B Vaghela, and V Shreeniwas. Prelinguistic verb acquisition from repeated language exposure for visual events. In *International Conference on Natural Language Processing*, Hyderabad, India, 2004.
[13] Tim Oates, Paul R Cohen, Marc S Atkin, and Carole R Beal. Building a baby. In *Proceedings of the 18th annual conference of the Cognitive Science Society*, pages 518–522, 1996.
[14] Steven Pinker. How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377–410, April 1994.
[15] Deb Roy. Integration of speech and vision using mutual information. In *International Conference on Acoustics,Speech and Signal Processing(ICASSP00)*, 2000.
[16] Elizabeth S. Spelke. Principles of object perception. *Cognitive Science*, 14:29–56, 1990.
[17] Luc Steels. Language learning and language contact. In *Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks, ECML-97*, pages 11 – 24, 1997.
[18] 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 CRV 2006*, June 7th-9th 2006.