Acquiring Linguistic Argument Structure from Multimodal Input using Attentive Focus

G Satish and Amitabha Mukerjee
Computer Science and Engineering,
Indian Institute of Technology Kanpur, Uttar Pradesh, India
{*satish,amit*}@*cse.iitk.ac.in*

***Abstract*—This work is premised on three assumptions: that the semantics of certain actions may be learned prior to language, that objects in attentive focus are likely to indicate the arguments participating in that action, and that knowing such arguments helps align linguistic attention on the relevant predicate (*verb*). Using a computational model of dynamic attention, we present an algorithm that clusters visual events into action classes in an unsupervised manner using the Merge Neural Gas algorithm. With few clusters, the model correlates to coarse concepts such as *come-closer*, but with a finer granularity, it reveals hierarchical substructure such as *come-closer-one-object-static* and *come-closer-both-moving*. That the argument ordering is non-commutative is discovered for actions such as *chase* or *come-closer-one-object-static*. Knowing the arguments, and given that noun-referent mappings that are easily learned, language learning can now be constrained by considering only linguistic expressions and actions that refer to the objects in perceptual focus. We learn action schemas for linguistic units like "moving towards" or "chase", and validate our results by producing output commentaries for 3D video.**

I. Introduction

Computationally, learning noun-referent mappings from unsupervised multimodal input is quite well understood [19], [16], [15], but verbs present a more complex challenge [5], [21]. It has been hypothesized that this is because of unavailability of the grounded instance, as is available for nouns. Yet verbs constitute the central structure that define how an utterance is to be interpreted. Computational models that attempt to learn verbs often involve manually encoding some parts of an interaction space [14], [4], [12], [21].

In this work we take the view that the semantics of the verb guides its syntax, sometimes known as the *semantic selection hypothesis* [22]. Cognitively, this may imply that early language learners may be acquiring some basic action schemas along with their argument structure, in a pre-linguistic stage [9]. Later, language labels get associated with these available concepts. Computationally, the availability of the argument structure makes it easy to constrain the search to linguistic fragments involving the corresponding nominals.

We postulate that a key aspect of this process is the role of perceptual attention [15], [1]. Attention constrains visual search, but also helps limit the set of agents participating in the action, which eventually generalizes to the argument structure. However, there have been very few attempts that use attention for learning language predicates or motion categories. Direct human gaze was tracked in [1], who use the narrator's gaze, head and hand movements for grounded word acquisition, and verbs such as "picking up" and "stapling" are associated with the actions. However, the verbal concepts learned are difficult to generalize into action schemas, applicable to new scenes or situations. Top down attention guided by linguistic inputs is used to identify objects in [17]. More recently, in [6] attentive focus is used to learn labels for simple motion trajectories, but this is also restricted to a particular visual domain.

A. From Concept to Label to Language

Here we wish to explore the possibility of learning transitive verbs, along with their argument structure (involving two agents), in an unsupervised manner, based on initial perceptual input, and later multimodal input. We proceed in three stages: concept learning, language association, and finally, validation through language production in a novel context.

In the *concept learning phase*, the system induces a model or schema for the action observed, from the perceptual input alone. Such models, often called *Image Schema* in Cognitive Linguistics [8] or *Perceptual Schema* in Experimental Psychology [9], involve abstractions on low-level features extracted from sensorimotor modalities (positions and velocities), as well as the argument structure. These schemas are implicit, and instances can only be verified against the schema, the structure itself is not explicitly available. Constructing such action templates has a long history in Computer vision [11], but the emphasis has been on recognizing single-agent activities (e.g. gestures), and less so on the interactions between agents. Most of the work has used visual priors for recognition, and only recently have unsupervised approaches become prevalent [13], [6]. We restrict ourselves to two-object interactions, using no priors, and our feature vectors are combinations of relative position and velocity vectors of the objects (we use a simple inner product). We use the Merge Neural Gas algorithm [20], which maintains a temporal context against which event similarity is measured for clustering purposes. By considering different levels of cluster granularity in the unsupervised learning process, we also learn subsets of coarse concepts as finer action concepts, resulting in an action hierarchy which may be thought of as a rudimentary ontology. However, since the concepts are grounded, the model is considerably flexible, unlike a predicate-inheritance based structure.

In the *Label learning* phase, the action semantics learned are associated with a co-occurrent textual input. For this, we

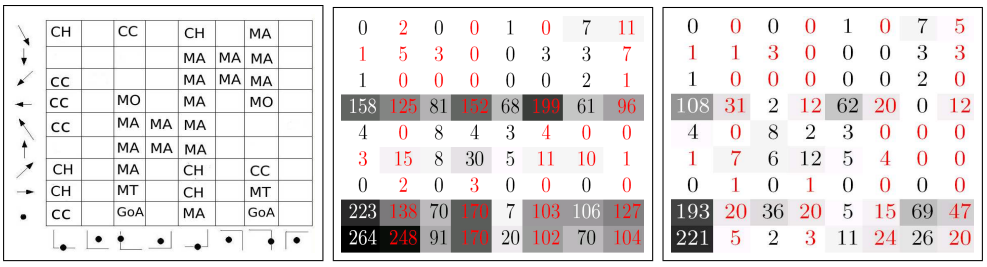

Fig. 1. *Attention-based selection of object pairs more likely to select for salience.* (*a*) Qualitative analysis of two object interaction: relative positions and relative velocity of second object along X & Y respectively (reference object is at origin and moving along x-axis). Cases when motion does not have a simple English label are blank. Others labels are: Come Closer (CC), Move {Away,Opposite, Together} (MA,MO,MT), Chase (CH) and Go Around (GoA). (*b*). Distribution observed when object pairs are chosen without using attention: More than half (58%) the cases are un-labelled motions (in red). (*c*) Distribution with attentive focus: 76% of frames have labels.

use noun labels which are already associated with the scene objects, and we identify nominals present in any sentence, and consider only actions that involve these agents in the scene. This constrains the association search, and we show that simply maximizing the conditional probability is sufficient to obtain relatively good fits, unlike more complex association measures such as those used in Machine translation [2].

In the final *Language production* phase, we attempt to generate linguistic predicates instantiated with arguments, to describe novel 2-body motions. We take a 3D surveillance video, in which the depth of a foreground object is indicated by it's image y-coordinate. We show that the motion features of humans can be labelled using the action schemas learned. However, since we have not learned the morphology or syntax, we simply produce the verb head along with the arguments.

II. Role of Attentive Focus

One of the key questions we ask in this paper is about the relevance of attentive focus. It turns out that restricting computation to attended events somehow results in a better correlation with motions that are named in language (Fig.1). Like other models that use attention to associate agents or actions to language [1], [6], we use attentive focus to constrain the region of visual salience, and thereby the constituents participating in an action. We use a computational model of dynamic visual attention [18] to identify agents possibly in focus.

While the process we adopt for learning two-agent interactions is completely unsupervised, in order to simplify the visual processing, we use a 2D video [1] of blocks and circles moving around, well known in psychology [7]. It is assumed that the perceptual apparatus can segment coherently moving objects as "wholes", and the motions of individual objects can be tracked.

Fig.1 shows a simple qualitative analysis of 2-body motions from the video (Fig.2), distinguishing only the signs of the relative position and velocities. We observe that certain motion-position combinations have ready labels in English, whereas others are more difficult to name. When computing

[1]The particular video, as well as the commentaries used, were developed by Bridgitte Martin Hard of the Space, Time, and Action Research group at Stanford University.

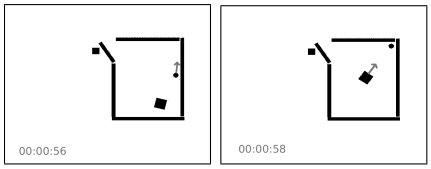

Fig. 2. *Scenes from "Chase" Video*: Three agents, "big square", 'small square" and "circle" play and chase each other. Circle *moves away* from big square (left) but big square *comes close* to the circle (right). Velocities are shown with gray rays.

the relative motions for a pair of objects, we find that the motions are much more likely to be labels when the object-pair is in attentive focus (Fig.1(c), 76%), than not (Fig.1(b), 42%). In some sense, this is not very surprising, because we give names to phenomena that are salient, those we attend to.

After learning the action schemas, we consider commentaries from thirteen users which are correlated with the actions learned. Four "action"s are found to dominate the descriptions of two agent interactions: *Come Close(CC), Move Away(MA), Chase1(A,B) Chase2(B,A)*. Chase1 and Chase2 differ in the argument order, the others are commutative. One of our results is that the set of clusters obtained by the learning system has a high correlation with this set of actions, but using a clustering model with a finer granularity, it reveals hierarchical substructure such as *come-closer-one-object-static* and *come-closer-both-moving*.

III. Unsupervised Clustering

In order to learn the actions, we use only two features - abstractions defined on the raw motion data - as input. Both features are dyadic (involving two bodies) and are based on relative velocity and positions of the bodies - *pos·velDiff*: $(\vec{x}_B - \vec{x}_A) \cdot (\vec{v}_B - \vec{v}_A)$ and *pos·velSum*: $(\vec{x}_B - \vec{x}_A) \cdot (\vec{v}_B + \vec{v}_A)$. Here '·' is the inner product and $\vec{x}_A$ and $\vec{v}_A$ refer to the unit position and velocity vectors of object A, which is taken as the reference object during feature vector computation. The first feature captures the combination of relative position and velocity, the second the relative position and magnitude.

These feature vectors are then clustered into categories in an unsupervised manner based on a notion of distance between individuals. We use the Merge Neural Gas(MNG) algorithm[20]

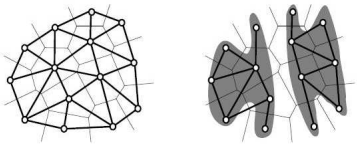

Fig. 3. (*Left*) Delaunay triangulation of vectors in $\Re^n$. (*Right*) Induced Delaunay triangulation(dark edges) in high signal activity region(dark region)

for unsupervised learning which has been shown to be well-suited for processing complex dynamic sequences as compared to the other existing models for temporal data processing like Temporal Kohonen map, Recursive SOM etc. This class of temporal learning algorithms are more flexible with respect to the state specifications and time history compared to HMMs or VLMMs.

A. Merge Neural Gas algorithm

Neural gas algorithm [10] learns important topological relations in a given set of input vectors(signals) in an unsupervised manner by means of a simple Hebb-like learning rule. It takes a distribution of high-dimensional data, P(ξ) and returns a densely connected network resembling the topology of the input. The main steps of the algorithm are explained below.

A fixed number of random neurons are taken in $\Re^n$, *n* being the dimension of the signal space. For every signal ξ_i an edge is introduced between the two closest neurons. The resulting network would be a sub-graph of Delaunay triangulation of the set of neurons (Fig. 3) with edges present in the regions of high signal activity. The neurons that do not participate in this *edge growing* process are called *dead units*. To make use of all the neurons, adaptation should take place towards the signal area. This is achieved in [10] by a Vector Quantization procedure called Neural Gas. For every signal the neurons are adapted towards the signal; the adaptation falls off exponentially as the distance of neuron from the signal increases. This step makes the *dead units* move towards the signal area and participate in the *edge growing* process. An *edge aging* mechanism is introduced, to remove the edges made obsolete by the neuron movement, by setting an upper bound (*edge aging* parameter) for the edge ages. The above steps are repeated over the signal set till the adaptation or movement of neurons goes to zero and the closely connected neurons lie in the signal activity region.

For input feature vectors arriving from temporally connected data, the information present in the time history is not captured by the neural gas algorithm given above. Merge Neural Gas algorithm [20] combines the neural gas mechanism with explicit context representation which utilizes the temporal ordering present in the feature vectors of the frames (signals). We explain the changes made to the Neural Gas algorithm to pass on the temporal information. A new vector, *Context*, is defined for signals and neurons. The context vectors of the neurons are initialized randomly along with the feature vectors. In every iteration over the signal set, Context of the current signal(c_ξ) is set as a linear combination of the feature vector(f_w) and context vector(c_w) of the previous winner neuron [$c_\xi = (1-\beta).f_w + \beta.c_w$]. The context vector of the first signal in the sequence is set to zero. The mixing factor, β, decides the extent of context diversification (Here, β=0.55). The distance function is also modified to accommodate context influence(along with the usual feature vector) in the winner neuron selection [$d_i = (1-\alpha).dist(f_\xi, f_i) + \alpha.dist(c_\xi, c_i)$]. α determines the contribution of context distance. The knowledge about the previous signals(or frames) in the sequence is passed on using the context vector. The context and feature vectors of the neurons are adapted towards current signal based on their distance from the signal. The actual algorithm is in [20].

The value of α is critical in utilizing the temporal information present in the signals. A low value (α=0.02) is taken initially and as the neurons are adapted, the feature vectors of the neurons become more reliable and the entropy of the adaptation decreases (i.e., the movement of the neurons decreases). α is then increased (α=0.6) to allow more context(temporal) information to counteract the specialization of neurons only on the feature vectors of the signals. And as entropy increases α is reduced (α=0.4) to allow a fine tuning of the context influence and of the ordering. It has been observed that during the final stages, the context vector of the current signal converges to an encoding of the feature vectors of the previous sequence signals thus providing the temporal information needed at the current step. Cluster labelling for the frames is obtained in the final iteration of the algorithm using the winner neuron.

IV. Concept Acquisition: Chase video

Unsupervised clustering using the Merge Neural Gas algorithm is used on the feature vectors from the video, corresponding to object pairs that were in attentive focus around the same time. Salient objects in a scene are ordered by a computational model of bottom-up dynamic attention[18]. The most salient object is determined for each frame, and other objects that were salient within k frames before and after (we use $k = 10$) are considered as attended simultaneously. Dyadic feature vectors are computed for all object pairs in these $2k$ frames.

Owing to the randomized nature of the algorithm, the number of clusters varies from run to run. Clusters with less than ten frames are not considered. When the *aging* parameter was set to 30, the number of clusters came out to be four in 90% of the runs.

In order to validate these clusters with human concepts, we needed to obtain human labels (Ground Truth) for the actions. Taking the dominant actions from the commentary (*Come Close(CC), Move Away(MA), Chase1 or Chase2*), we asked three subjects (Male, Hindi-English / Telugu-English bilinguals, Age-22, 20 and 30) to label the scenes in the video. They were shown the video twice and in the third viewing they were asked to speak out one of three action labels (CC, MA, Chase) which was recorded. Given the label and the frame when this was uttered, the actual boundaries and participating

TABLE I

Clustering Accuracy: THE i^{th} ROW, j^{th} COLUMN GIVES THE NUMBER OF i^{th} ACTION LABELS IN j^{th} NG CLUSTER. % IS THE FRACTION OF VECTORS OF AN ACTION CORRECTLY CLASSIFIED TO THE TOTAL VECTORS OF THAT TYPE. TOTAL CLASSIFICATION ACCURACY(TCA) IS THE % OF TOTAL VECTORS CORRECTLY CLASSIFIED .

	C_1	C_2	C_3	C_4	Total	%	TCA
CC	399	6	10	29	444	90	
MA	16	311	5	48	380	82	84
Chase	21	59	149	154	383	79	

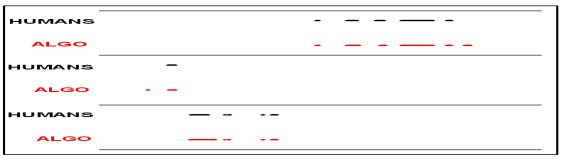

Fig. 4. Comparison of human and algorithm labelling of "come closer" action over a time line(x-axis) of first 1500 frames.

TABLE II

Hierarchical clustering: USING A LARGER NUMBER OF CLUSTERS REVEALS A SUB-CLASSIFICATION; E.G. FRAMES CLASSIFIED AS CC IN TABLE I, ARE NOW IN $C_1, C_5, or C_6$, REFLECTING TWO CASES OF $CC_{one-object-static}$, OR ONE CASE OF $CC_{both-moving}$.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
CC	201	3	9	20	189	21	1	0
MA	8	126	4	45	9	1	181	6
Chase	1	9	142	151	13	9	32	26

TABLE III

Relevance of Argument Order: VALUE AT i^{th} ROW, j^{th} COLUMN GIVES NUMBER OF VECTORS THAT WERE ORIGINALLY IN CLUSTER i AND NOW ASSIGNED TO CLUSTER j WHEN OBJECT ORDER WAS SWITCHED IN DYADIC FEATURE VECTORS. NOTE THAT C3 AND C4, THE CLUSTERS CORRESPONDING TO *Chase*, ARE FLIPPED.

	C_1	C_2	C_3	C_4
C_1	390	20	11	15
C_2	9	323	15	29
C_3	6	12	1	145
C_4	22	48	152	9

objects were assigned by inspection. In case of disagreement, we took the majority view.

To validate the clustering algorithm, we correlated the clusters learned with these human labellings, frame by frame. The distribution of human labels across the neural gas clusters when number of clusters was four is given in Table I. Clusters C1 and C2 had an overwhelming correlation with CC and MA. C3 and C4 were found to have a majority correlation with Chase. Inspecting the individual frames, it was seen that the reference object is leader in one chase cluster and chaser in the other, i.e. the argument order distinction - chase1(A,B) vs. chase2(B,A) - is discovered autonomously by the algorithm.

Fig.4 and Fig.5 present results along a time line for *Come-Closer* and *Chase* actions, each row reflects a different combination of agents (small square, big square, circle).

A surprising result was found when by experimenting with the *edge aging* parameter in the Merge Neural Gas algorithm. The number of clusters increase as aging parameter is decreased, and at one stage nine clusters were formed (edge aging parameter=16). The Total Classification Accuracy (TCA) was about 51 and we would have discarded the result, but inspecting the frames revealed that the clusters may be reflecting what appeared to be hierarchy of action types. Thus cluster C_1 from the earlier classification (majority correlation=CC) was broken up into C_1, C_5, C_6. C_1 was found to contain frames where both objects are moving towards each other whereas C_5 contains frames where the smaller object is stationary and the other moves closer. Thus *Come-Closer* and *Move-Away* appear to be sub-classified into 3 classes (two *one object static* cases, and one *both moving* case). This 'finer' classification is given in Table II, but since we have no human labellings at this fine level, we were not able to measure the clustering accuracy more precisely.

Fig. 5. Comparison of human and algorithm labelling of "chase" over first 1500 frames. Because of our choice of reference object, frames in first row are in C_3 and second row are in C_4.

A. Argument order in Action Schemas

In another experiment, we investigated the importance of argument ordering by re-classifying the same frames, but reversing the order of the objects used in the dyadic vector computation. Earlier, if the larger object was $arg1$ or reference object, now it became $arg2$ or non-reference object. If the corresponding concept changed, especially if it flipped, this would reflect a semantic necessity to preserve the argument order; otherwise the arguments were commutative. Using the coarser clusters, we observe that the argument order is immaterial since the majority relation is unchanged (black) for C1 and C2 (CC,MA respectively). On the other hand, both C3 and C4 (correlations with Chase) are flipped (Table III). Thus, the fact that argument order is important for *Chase* is learned implicitly within the action schema itself.

V. LANGUAGE ASSOCIATION

Having learned some action schemas, we now try to associate these with language labels by associating the actions with co-occurring narrative. Thirteen different commentaries on the Chase video were used. The commentaries vary considerably - e.g. for events in Fig.2, we have: "large square corners the little circle", "big square approaches little circle", "little circle is moving away from big square; and objects inside are moving closer together", "big block tries to go after little circle", etc. Learning the noun-object mappings in this type of input have been reported elsewhere[12], and now we use this knowledge to correlate linguistic and perceptual focus. Given two objects

TABLE IV
Association Results: MAXIMUM VALUES OF ASSOCIATION MEASURE FOR N-GRAMS (ASSOCIATION VALUE IN PARENTHESIS).

Clusters	Monograms	Bigrams
$Cluster1$	Move(0.05)	Move toward(0.11)
$Cluster2$	Come(0.06)	Move away(0.10)
$Cluster3$	Chase(0.67)	Chase around(0.30)
$Cluster4$	Chase(0.34)	Chase after(0.14)

TABLE V
PHRASE WITH HIGHEST ASSOCIATION MEASURE FOR EACH CLUSTER

Clusters	With Stemming	Without Stemming
C_1	move toward	move towards
C_2	move away	moving away
C_3	chase	chasing
C_4	chase	chases

that are mentioned in a sentence, only those actions involving these two objects were associated. As the number of objects increase, object pairings rise as $O(n^2)$ and this is clearly a computationally significant constraint. Since we are only interested in concepts involving two agents, sentences not referring to two nominals (or pronomial equivalents), are eliminated. Very frequently occurring words (e.g. a, an, the) are pruned. Assuming mutual exclusivity [15], we drop the lexical labels for the agents participating in the action; the mutual exclusivity principle holds that if an object has one name, it should not have another. In our case, if the name is that of the agent, it cannot be the name of the action as well.

In other work on perception-language association, complex association measures adapted from Machine Translation are often used [2]. However, given the excellent pruning already achieved by aligning arguments, we were able to obtain satisfactory results simply by maximizing the conditional probability ($P(p_j/C_i) = \frac{P(C_i/p_j)}{P(C_i)}\ P(p_j)$) over the set of phrase/n-gram p_j co-occurring with the action schema C_i. P(p_j) is computed as the ratio of frequency of the n-gram to the sum of frequencies of all n-grams in the commentary file. For 3-grams and above P(p_j) becomes unreliable because of the sparse commentary. So we consider only monograms and bigrams as labels for the clusters. We perform association both with and without stemming of the commentary (Porter stemming). We consider the coarser level clusters obtained by the unsupervised algorithm. Association results are listed in Table IV - note that English does not have one-word classifications for CC or MA, and the corresponding clusters show a stronger association with the 2-gram phrases. The best linguistic labels assigned to the clusters are listed in Table V; these constitute valid linguistic units of English describing the majority mappings for each of these actions.

VI. LANGUAGE PRODUCTION: DESCRIBING UNSEEN 3D VIDEO

Now that language labels, as well as action schemas (which includes the argument structure) have been learned from a multimodal stimulus (the Chase video), can these be used to generate linguistic fragments describing similar situations? Other 2D videos were tested for this, but here we report results from a 3D video of three persons running around in a field (Fig.6). In human classification of the action categories (into one of *CC, MA, Chase*), the dominant predicate in the video, (777 out of 991 frames), is Chase.

Fig. 6. *Test Video :* Scenes from the 3D video

TABLE VI
DISTRIBUTION OF CHASE FRAMES(GROUND TRUTH) FROM THE 3D VIDEO ACROSS THE NEURAL GAS CLUSTERS

	C_1	C_2	C_3	C_4	Total Chase Frames	%
Chase	13	15	496	253	777	96

In the image processing stage, the system learns the background over the initial frames based on which it segments out the foreground blobs. It is then able to track all the three agents using the Meanshift algorithm. Assuming camera height near eye level, the bottom-most point in each blob corresponds to that agent's contact with the ground, from which its depth can be determined within some scaling error (157 frames with extensive occlusion between agents were omitted). Given this depth, one can solve for the lateral position - thus, we are able to obtain, from a single view video, the (x, y) coordinates for each agent in each frame, within a constant scale. Based on this, the relative pose and motion parameters are computed for each agent pair, and therefrom the features as outlined earlier. Now these feature vectors are classified using the action schemas (coarse clusters) already obtained from the Chase video (2D). Each feature vector is assigned the same label as its nearest neuron's cluster. Table VI shows the result for the predicate Chase; this has a 63% match with cluster 3 and 34% match with cluster 4, both of which had been earlier associated with Chase. Remember these two clusters differ in whether the larger object is the leader or the chaser. Using the best-matching phrase from Table V, we can now generate predicates with arguments and preserving ordering (Figure 7); while these are not actual linguistic expressions, once the system has learned some morphology and syntax, this association can directly work with richer linguistic structures[8].

Fig. 7. *Commentary generated by Algorithm*: "Chase(Red,Green)", "Move Away(Red,Yellow)", "Move Away(Green,Yellow)"

VII. Discussion and Conclusion

We have outlined how our unsupervised approach learns action schemas of two-agent interactions, the arguments these actions may take, and also the linguistic predicates for such actions. The image schematic nature of the clusters are validated by producing commentary for a 3D video. The approach provided here underlines the role of concept argument structures in aligning with linguistic expressions, and that of bottom-up dynamic attention in pruning the visual input and in aligning linguistic focus.

An important assumption underlying the methodology is that the attention mechanism of our (computational) learner is directed towards the same objects that the original narrators had focused on. It is as if the listener knows, without explicitly tracking their gaze, that the objects that are salient to her are also those that were salient to the speaker. This assumption, labeled the *Perceptual Theory of Mind*[12], may constitute an important element in much discourse understanding.

While there is evidence that human infants are acquiring some action models (often called perceptual schemas) in the pre-linguistic stage, what are the computational imperatives for such an approach? In the past, computational efforts have often presented both language and visual inputs simultaneously [14], but these efforts required incorporating some constraints manually in the form of structured elements. This larger input space, including language as well as perception, increases the dimensionality of search enormously. Another reason why learning the perceptual schemata first may be easier, is that it reduces the language learning problem to one of associating tokens across modalities, a problem that has been well addressed in machine translation[2]. Further, knowing the arguments involved in an action schema severely restricts the linguistic search as well.

Speculating further on the role of semantics in language acquisition, one may suggest a mechanism for acquiring grammatical elements, which are meaningless if not learned together with their semantic pole [8]. Computationally, the grounded semantics underlying syntactical structures may actually make it easier to learn the associated syntax as well. That argument ordering may be important is highlighted in situations involving irreversible predicates (e.g. X chases Y). Other aspects such as particles "from" (participating here with "move away"), and also morphological elements like "-er" e.g. *chaser*, may be more tractably learned if combined with the semantic pole underlying these structures.

Once a few *basic* concepts are learned, other concepts can be learned without direct grounding, by using conceptual blending mechanisms on the concept itself. These operations are often triggered by linguistic cues, resulting in new concepts, as well as their labels being learned together, in a later stage. Indeed, the vast majority of our vocabularies are learned later purely from the linguistic input [3]. But this is only possible because of the grounded nature of the first few concepts, without which these later concepts cannot be grounded. Thus the perceptually grounded nature of the very first concepts are crucial to subsequent compositions. The linguistic aspects of these new structures may now be derived from these conceptual underpinnings, although this of course has been a matter of considerable debate.

Acknowledgements

We are grateful to Barbara Tversky and her group for comments on an earlier draft (as well as the video and commentaries). Dana Ballard was instrumental in our developing the dynamic attention model. We acknowledge support from the *Research I Foundation*

References

[1] Ballard, D. H., and Yu, C. A multimodal learning interface for word acquisition. In *International Conference on Acoustics,Speech and Signal Processing(ICASSP03)* (April 2003), vol. 5, pp. 784–7.

[2] Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D. A., Blei, D., and Jordan, M. Matching words and pictures. *Journal of Machine Learning Research 3* (2003), 1107–1135.

[3] Bloom, P. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA, 2000.

[4] Fleischman, M., Decamp, P., and Roy, D. Mining temporal patterns of movement for video content classification. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval* (New York, NY, USA, 2006), ACM, pp. 183–192.

[5] Fleischman, M., and Roy, D. Why verbs are harder to learn than nouns: Initial insights from a computational model of intention recognition in situated word learning. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (2005).

[6] Guha, P., and Mukerjee, A. Baby's day out: Attentive vision for pre-linguistic concepts and language acquisition. In *Proceedings 4th Workshop on Attention in Cognitive Systems WAPCV-2007* (2007), L. Paletta, Ed., pp. 81–94.

[7] Heider, F., and Simmel, M. An experimental study of apparent behavior. In *American Journal of Psychology* (1944), vol. 57, pp. 243–59.

[8] Langacker, R. W. *Grammar and Conceptualization*. Berlin/New York: Mouton de Gruyer, 1999.

[9] Mandler, J. M. *Foundations of Mind*. Oxford University Press, 2004.

[10] Martinetz, T., and Schulten, K. Topology representing networks. *Neural Networks 7*, 3 (1994), 507–522.

[11] Moeslund, T. B., and Granum, E. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding 81* (2001), 231–268.

[12] Mukerjee, A., and Sarkar, M. Perceptual theory of mind: An intermediary between visual salience and noun /verb acquisition. In *Proc. International Conference on Developmental Learning* (2006).

[13] Niebles, J. C., Wang, H., and Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. In *Proceedings of the British Machine Vision Conference* (2006).

[14] Regier, T. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Bradford Books, September 1996, p. 276.

[15] Regier, T. Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences 7* (2003), 263–268.

[16] Roy, D. Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences 9*, 8 (August 2005), 389–396.

[17] Roy, D., and Mukherjee, N. Towards situated speech understanding: visual context priming of language models. *Computer Speech and Language 19*, 2 (2005), 227–248.

[18] Singh, V. K., Maji, S., and Mukerjee, A. Confidence based updation of motion conspicuity in dynamic scenes. In *Third Canadian Conference on Computer and Robot Vision* (2006).

[19] Steels, L. Language learning and language contact. In *Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks, ECML-97* (1997), pp. 11 – 24.

[20] Strickert, M., and Hammer, B. Merge som for temporal data. *Neurocomputing 64* (2005), 39–71.

[21] Sugiura, K., and Iwahashi, N. Learning object-manipulation verbs for human-robot communication. In *WMISI '07: Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction* (New York, NY, USA, 2007), ACM, pp. 32–38.

[22] Wierzbicka, A. *The Semantics of Grammar*. John Benjamins, Studies in Language Companion Series 18, Amsterdam, 1988.