Embodied Active Vision in Language Learning and Grounding

Chen Yu

Indiana University, Bloomington IN 47401, USA,
`chenyu@indiana.edu`,
WWW home page: `http://www.indiana.edu/~dll/`

Abstract. Most cognitive studies of language acquisition in both natural systems and artificial systems have focused on the role of purely linguistic information as the central constraint. However, we argue that non-linguistic information, such as vision and talkers' attention, also plays a major role in language acquisition. To support this argument, this chapter reports two studies of embodied language learning – one on natural intelligence and one on artificial intelligence. First, we developed a novel method that seeks to describe the visual learning environment from a young child's point of view. A multi-camera sensing environment is built which consists of two head-mounted mini cameras that are placed on both the child's and the parent's foreheads respectively. The major result is that the child uses their body to constrain the visual information s/he perceives and by doing so adapts to an embodied solution to deal with the reference uncertainty problem in language learning. In our second study, we developed a learning system trained in an unsupervised mode in which users perform everyday tasks while providing natural language descriptions of their behaviors. The system collects acoustic signals in concert with user-centric multisensory information from non-speech modalities, such as user's perspective video, gaze positions, head directions and hand movements. A multimodal learning algorithm uses this data to first spot words from continuous speech and then associate action verbs and object names with their perceptually grounded meanings. Similar to human learners, the central ideas of our computational system are to make use of non-speech contextual information to facilitate word spotting, and utilize body movements as deictic references to associate temporally co-occurring data from different modalities and build a visually grounded lexicon.

1 Introduction

One of the important goals in cognitive science research is to understand human language learning systems and apply the findings of human cognitive systems to build artificial intelligence systems that can learn and use language in human-like ways. Learning the meanings of words poses a special challenge towards this goal, as illustrated in the following theoretical puzzle (Quine, 1960): Imagine that you are a stranger in a strange land with no knowledge of the language

or customs. A native says "Gavagai" while pointing at a rabbit running by in the distance. How can you determine the intended referent? Quine offered this puzzle as an example of reference uncertainty in mapping language to the physical world (what words in a language refer to). Quine argued that, given the novel word "Gavagai" and the object rabbit, there would be an infinite number of possible intended meanings - ranging from the basic level kind of rabbit, to a subordinate/superordinate kind, its color, fur, parts, or activity. Quine's example points up a fundamental problem in first language lexical acquisition - the ambiguity problem of word-to-world mapping.

A common conjecture about human lexical learning is that children map sounds to meanings by seeing an object while hearing an auditory word-form. The most popular mechanism of this word learning process is *associationism*. Most learning in this framework concentrates on statistical learning of co-occurring data from the linguistic modality and non-linguistic context (see a review by Plunkett, 1997). Smith (2000) argued that word learning trains children's attention so that they attend to the just right properties for the linguistic and world context. Nonetheless, a major advance in recent developmental research has been the documentation of the powerful role of social-interactional cues in guiding the learning and in linking the linguistic stream to objects and events in the world (Baldwin, 1993; Tomasello & Akhtar, 1995). Many studies (e.g., Baldwin, 1993; Woodward & Guajardo, 2002) have shown that there is much useful information in social interaction and that young learners are highly sensitive to that information. Often in this literature, children's sensitivities to social cues are interpreted in terms of (seen as diagnostic markers of) children's ability to infer the intentions of the speaker. This kind of social cognition is called "mind reading" by Baron-Cohen (1995). Bloom (2000) suggested that children's word learning in the second year of life actually draws extensively on their understanding of the thoughts of speakers. However, there is an alternative explanation of these findings to the proposals of "mind-reading". Smith (2000) has suggested that these results may be understood in terms of the child's learning of correlations among actions, gestures and words of the mature speaker, and intended referents. Smith (2000) argued that construing the problem in this way does not "explain away" notions of "mind-reading" but rather grounds those notions in the perceptual cues available in the real-time task that young learners must solve.

Meanwhile, Bertenthal, Campos, and Kermoian (1994) have shown how movement — crawling and walking over, under, and around obstacles - creates dynamic visual information crucial to children's developing knowledge about space. Researchers studying the role of social partners in development and problem solving also point to the body and active movement -points, head turns, and eye gaze - in social dynamics and particularly in establishing joint attention. Computational theorists and roboticists (e.g. Ballard, Hayhoe, Pook, & Rao, 1997; Steels & Vogt, 1997) have also demonstrated the computational advantages of what they call "active vision", how an observer - human or robot - is able to understand a visual environment more effectively and efficiently by interacting

with it. This is because perception and action form a closed loop; attentional acts are preparatory to and made manifest in action while also constraining perception in the next moment. Ballard and colleagues proposed a model of "embodied cognition" that operates at time scales of approximately one-third of a second and uses subtle orienting movements of the body during a variety of cognitive tasks as input to a computational model. At this "embodiment" level, the constraints of the body determine the nature of cognitive operations, and the body's pointing movements are used as deictic (pointing) references to bind objects in the physical environment to variables in cognitive programs of the brain.

In the present study, we apply embodied cognition in language learning. Our hypothesis is that momentary body movements may constrain and clean visual input to human or artificial agents situated in a linguistic environment and in doing so provide a unique embodied solution to the reference uncertainty problem. To support this argument, we have designed and implemented two studies – one on human learners and one on machine learners. The results from both studies consistently show the critical advantages of embodied learning.

2　Embodied Active Vision in Human Learning

The larger goal of this research enterprise is to understand the building blocks for fundamental cognitive capabilities and, in particular, to ground social interaction and the theory of mind in sensorimotor processes. To these ends, we have developed a new method for studying the structure of children's dynamic visual experiences as they relate to children's active participation in a physical and social world. In this paper, we report results from a study that implemented a sensing system for recording the visual input from both the child's point of view and the parent's viewpoint as they engage in toy play. With this new methodology, we compare and analyze the dynamic structure of visual information from these two views. The results show that the dynamic first-person perspective from a child is substantially different from either the parent's or the third-person (experimenter) view commonly used in developmental studies of both the learning environment and parent-child social interaction. The key differences are these: the child's view is much more dynamically variable, more tightly tied to the child's own goal-directed action, and more narrowly focused on the momentary object of interest – an embodied solution to the reference uncertainty problem.

2.1　Multi-Camera Sensing Environment

The method uses a multi-camera sensing system in a laboratory environment wherein children and parents are asked to freely interact with each other. As shown in Figure 1, participants interactions are recorded by three cameras from different perspectives - one head-mounted camera from the child's point of view to obtain an approximation of the child's visual field, one from the parent's viewpoint to obtain an approximation of the parent's visual field, and one from
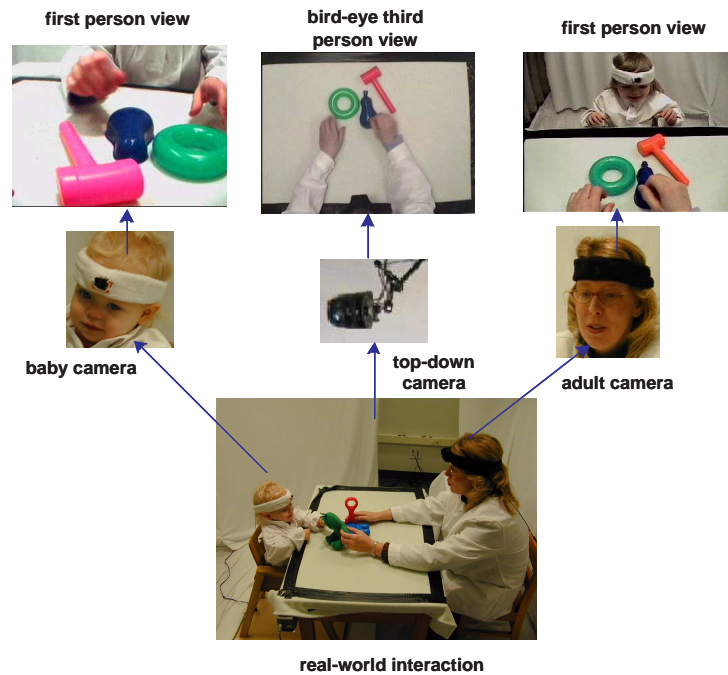
Fig. 1. Multi-camera sensing system. The child and the mother play with a set of toys at a table. Two mini cameras are placed onto the child's and the mother's heads respectively to collect visual information from two first-person views. A third camera mounted on the top of the table records the bird-eye view of the whole interaction.

a top-down third-person viewpoint that allows a clear observation of exactly what was on the table at any given moment (mostly the participants' hands and the objects being played with).

Head-Mounted Cameras. Two light-weight head-mounted mini cameras (one for the child and another for the parent) were used to record the first-person view from both the child and the parent's perspectives. These cameras were mounted on two everyday sports headbands, each of which was placed on one participant's forehead and close to his eyes. The angle of the camera was adjustable. The head camera field is approximately 90 degrees, which is comparable to the visual field of young learner, toddlers and adults. One possible concern in the use of a head camera is that the head camera image changes with changes in head movements not in eye-movements. This problem is reduced by the geometry of table-top play. In fact, Yoshida and Smith (2007) documented this in a head-camera study of toddlers by independently recording eye-gaze and showed that small shifts in eye-gaze direction unaccompanied by a head shift do not yield distinct table-top views. Indeed, in their study 90% of head camera video frames corresponded with independently coded eye positions.

Bird-Eye View Camera. A high-resolution camera was mounted right above the table and the table edges aligned with edges of the bird-eye image. This view provided visual information that was independent of gaze and head movements of a participant and therefore it recorded the whole interaction from

a third-person static view. An additional benefit of this camera lied in the high-quality video, which made our following image segmentation and object tracking software work more robustly compared with two head-mounted mini cameras. Those two were light-weighted but with a limited resolution and video quality due to the small size.
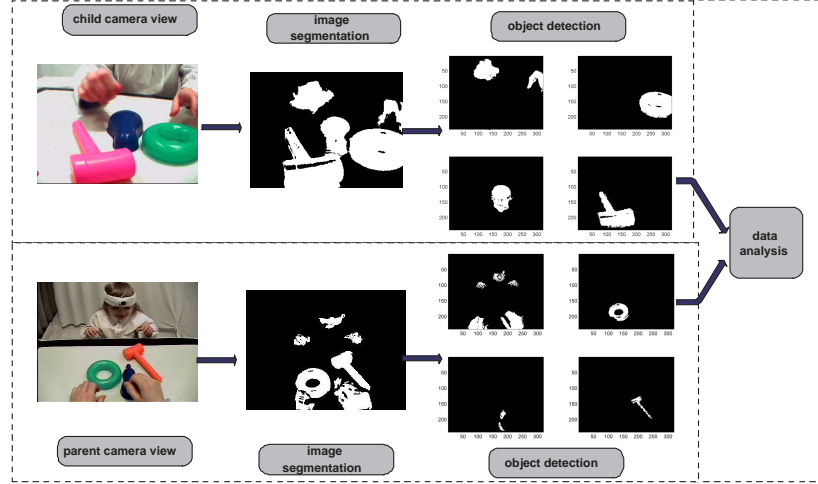
Fig. 2. Overview of data processing using computer vision techniques. We first remove background pixels from an image and then spot objects and hands in the image based on pre-trained object models. The visual information from two views is then aligned for further data analyses.

2.2 Image Segmentation and Object Detection

The recording rate for each camera is 10 frames per second. In total, we have collected approximately 10800 ($10 \times 60 \times 6 \times 3$) image frames from each inter-action. The resolution of image frames is 320×240.

The first goal of data processing is to automatically extract visual information, such as the locations and sizes of objects, hands, and faces, from sensory data in each of the three cameras. These are based on computer vision techniques, and include three major steps (see Figure 2). Given raw images from multiple cameras, the first step is to separate background pixels and object pixels. This step is not trivial in general because two first-view cameras attached on the heads of two participants moved around all the time during interaction causing moment-to-moment changes in visual background. However, since we designed the experimental setup (as described above) by covering the walls, the floor and the tabletop with white fabrics and asking participants to wear white cloth, we simply treat close-to-white pixels in an image as background. Occasionally, this approach also removes small portions of an object that have light reflections on them as well. (This problem can be fixed in step 3). The second step focuses on

the remaining non-background pixels and breaks them up into several blobs using a fast and simple segmentation algorithm. This algorithm first creates groups of adjacent pixels that have color values within a small threshold of each other. The algorithm then attempts to create larger groups from the initial groups by using a much tighter threshold. This follow-up step of the algorithm attempts to determine which portions of the image belong to the same object even if that object is broken up visually into multiple segments. For instance, a hand may decompose a single object into several blobs. The third step assigns each blob into an object category. In this object detection task, we used Gaussian mixture models to pre-train a model for each individual object. By applying each object model to a segmented image, a probabilistic map is generated for each object indicating the likelihood of each pixel in an image belongs to this special object. Next, by putting probabilistic maps of all the possible objects together, and by considering spatial coherence of an object, our object detection algorithm assign an object label for each blob in a segmented image as shown in Figure 2. As a result of the above steps, we extract useful information from image sequences, such as what objects are in the visual field at each moment, and what are the sizes of those objects, which will be used in the following data analyses.

3 Data Analyses and Results

The multi-camera sensing environment and computer vision software components enable fine-grained description of child-parent interaction from two different viewpoints. In this section, we report our preliminary results while focusing on comparing sensory data collected simultaneously from two views. We are particularly interested in the differences between what a child sees and what the mature partner sees.
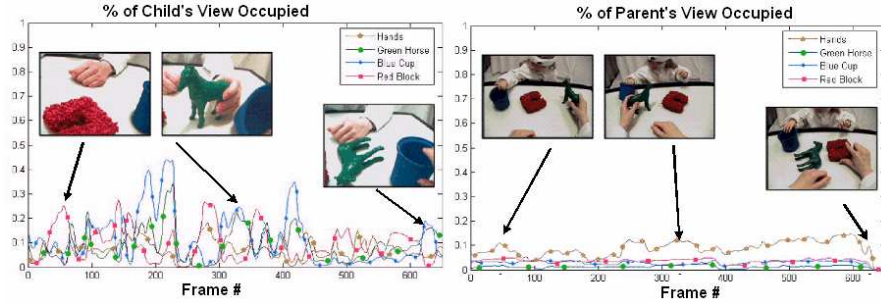

Fig. 3. A comparison of the child's and the parent's visual fields. Each curve represents a proportion of an object in the visual field over the whole trial. The total time in a trial is about 1 minute (600 frames). The three snapshots show the image frames from which the visual field information was extracted.

Figure 3 shows the proportion of each object or hand in one's visual field over a whole trial (three snapshots taken from the same moments from these two views). Clearly, the child's visual field is substantially different from the

parent's. Objects and hands occupy the majority of the child's visual field and the whole field changes dramatically moment by moment. In light of this general observation, we developed several metrics to quantify three aspects of the differences between these two views.

First, we measure the composition of visual field shown in Figure 4 (a). From the child's perspective, objects occupy about 20% of his visual field. In contrast, they take just less than 10% of the parent's visual field. Although the proportions of hands and faces are similar between these two views, a closer look of data suggests that the mother's face rarely occurs in the child's visual field while the mother's and the child's hands occupy a significant proportion (15%-35%) in some image frames. From the mother's viewpoint, the child's face is always around the center of the field while the hands of both participants occur frequently but occupy just a small proportion of visual field.

Second, Figure 4(b) compares the salience of the dominating object in two views. The dominating object for a frame is defined as the object that takes the largest proportion of visual field. Our hypothesis is that the child's view may provide a unique window of the world by filtering irrelevant information (through movement of the body close to the object) enabling the child to focus on one object (or one event) at a single moment. To support this argument, the first metric used here is the percentage of the dominating object in the visual field at each moment. In the child's view, the dominating object takes 12% of the visual field on average while it occupies just less than 4% of the parent's field. The second metric measures the ratio of the dominating object vs. other objects in the same visual field, in terms of the occupied proportion in an image frame. A higher ratio would suggest that the dominating object is more salient and distinct among all the objects in the scene. Our results show a big difference between two views. More than 30% of frames, there is one dominating object in the child's view which is much larger than other objects (ratio > 0.7). In contrast, less than 10% of time, the same phenomena happens in the parent's view.
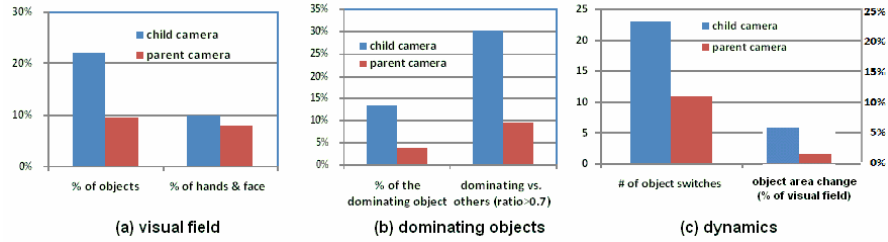
Fig. 4. We quantify and compare visual information from two views in three ways.

This result suggests not only that children and parents have different views of the environment but also that the child's view may provide more constrained and clean input to facilitate learning processes which don't need to handle a huge amount of irrelevant data because there is just one object (or event) in view at a time. We also note that this phenomenon doesn't happen randomly

and accidentally. Instead the child most often intentionally moves his body close to the dominating object and/or uses his hands to bring the object closer to his eyes which cause one object to dominate the visual field. Thus, the child's own action has direct influences on his visual perception and most-likely also on the underlying learning processes that may be tied to these perception-action loops.

The third measure is the dynamics of visual field, shown in Figure 4(c). The dominating object may change from moment to moment, and also the locations, appearance and the size of other objects in the visual field may change as well. Thus, we first calculated the number of times that the dominating object changed. From the child's viewpoint, there are on average 23 such object switches in a single trial (about 1 minute or 600 frames). There are only 11 per trial from the parent's view. These results together with the measures in Figure 4(b) suggest that children tend to move their head and body frequently to switch attended objects, attending at each moment to just one object. Parents, on the other hand, don't switch attended objects very often and all the objects on the table are in their visual field almost all the time.

The dynamics of their visual fields in terms of the change of objects in visual field makes the same point. In the child's view, on average, in each frame, 6% of the visual field consists of new objects, objects that are different from the just previous frame to frame. Only less than 2% of the parent's visual field changes this way frame to frame. over time. The child's view is more dynamic and such offers potentially more spatio-temporal regularities that may be utilized by lead young learners to pay attention to the more informative (from their point of view!) aspects of a cluttered environment.

There are two practical reasons that the child's view is quite different from the parent's view. First, because they are small, their head is close to the tabletop. Therefore, they perceive a "zoom-in", more detailed, and more narrowed view than taller parents. Second, at the behavioral level, children move objects and their own hands close to their eyes while adults rarely do that. Both explanations above can account for dramatic differences between these two views. Both factors highlight the crucial role of the body in human development and learning. The body constraints and narrows visual information perceived by a young learner. One challenge that young children face is the uncertainty and ambiguity inherent to real-world learning contexts: learners need to select the features that are reliably associated with an object from all possible visual features and they need to select the relevant object (at the moment) from among all possible referents on a table. In marked contrast to the mature partner's view, the visual data from the child's first-person view camera suggests a visual field filtered and narrowed by the child's own action. Whereas parents may selectively attend through internal processes that increase and decrease the weights of received sensory information, young children may selectively attend by using the external actions of their own body. This information reduction through their bodily actions may remove a certain degree of ambiguity from the child's learning environment and by doing so provide an advantage to bootstrap learning. This suggests that an adult view of the complexity of learning tasks may often be fundamentally wrong. Young

children may not need to deal with all the same complexity inherent in an adult's viewpoint - some of them that complexity may be automatically solved by bodily action and the corresponding sensory constraints. Thus, the word learning problem from the child learner's viewpoint is significantly simplified (and quite different from the experimenter's viewpoint) due to the embodiment constraint.

4   A Multimodal Learning System

Our studies on human language learners point to a promising direction for building anthropomorphic machines that learn and use language in human-like ways. More specifically, we take a quite different approach compared with traditional speech and language systems. The central idea is that the computational system needs to have sensorimotor experiences by interacting with the physical world. Our solution is to attach different kinds of sensors to a real person to share his/her sensorimotor experiences as shown in Figure 5. Those sensors include a head-mounted CCD camera to capture a first-person point of view, a microphone to sense acoustic signals, an eye tracker to track the course of eye movements that indicate the agent's attention, and position sensors attached to the head and hands of the agent to simulate proprioception in the sense of motion. The functions of those sensors are similar to human sensory systems and they allow the computational system to collect user-centric multisensory data to simulate the development of human-like perceptual capabilities. In the learning phase, the human agent performs some everyday tasks, such as making a sandwich, pouring some drinks or stapling a letter, while describing his/her actions verbally. We collect acoustic signals in concert with user-centric multisensory information from non-speech modalities, such as user's perspective video, gaze positions, head directions and hand movements. A multimodal learning algorithm is developed that first spots words from continuous speech and then builds the grounded semantics by associating object names and action verbs with visual perception and body movements. In this way, the computational system can share the lexicon with a human teacher shown in Figure 5.

To learn words from this input, the computer learner must solve three fundamental problems: (1) visual object segmentation and categorization to identify potential meanings from non-linguistic contextual information, (2) speech segmentation and word spotting to extract the sound patterns of the individual words which might have grounded meanings, and (3) association between spoken words and their meanings. To address those problems, our model includes the following components shown in Figure 6:

- **Attention detection** finds where and when a caregiver looks at the objects in the visual scene based on his or her gaze and head movements.
- **Visual processing** extracts visual features of the objects that the speaker is attending to. Those features consist of color, shape and texture properties of visual objects and are used to categorize the objects into semantic groups.
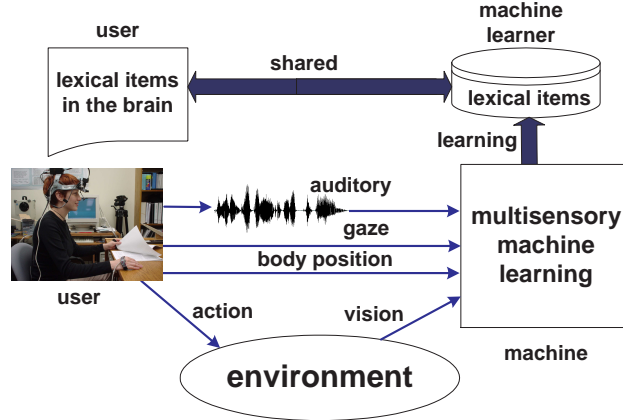
Fig. 5. The computational system shares sensorimotor experiences as well as linguistic labels with the speaker. In this way, the model and the language teacher can share the same meanings of spoken words.

- **Speech processing** includes two parts. One is to convert acoustic signals into discrete phoneme representations. The other is to compare phoneme sequences to find similar substrings and then cluster those subsequences.
- **Word discovery and word-meaning association** is the crucial step in which information from different modalities is integrated. The central idea is that extralinguistic information provides a context when a spoken utterance is produced. This contextual information is used to discover isolated spoken words from fluent speech and then map them to their perceptually grounded meanings extracted from visual perception.

Due to space limitations, the following sections will focus on the two most important components – attention detection and word-meaning association.

4.1 Estimating focus of attention

Eye movements are closely linked with visual attention. This gives rise to the idea of utilizing eye gaze and head direction to detect the speaker's focus of attention. We developed a velocity-based method to model eye movements using a hidden Markov model representation that has been widely used in speech recognition with great success (Rabiner & Juang, 1989). A hidden Markov model consists of a set of N states $S = \{s_1, s_2, s_3, ..., s_N\}$, the transition probability matrix $A = a_{ij}$, where a_{ij} is the transition probability of taking the transition from state s_i to state s_j, prior probabilities for the initial state π_i, and output probabilities of each state $b_i(O(t)) = P\{O(t)|s(t) = s_i\}$. Salvucci et al.(Salvucci & Anderson, 1998) first proposed a HMM-based fixation identification method that uses probabilistic analysis to determine the most likely identifications of a given protocol. Our approach is different from theirs in two ways. First, we use training data to estimate the transition probabilities instead of setting pre-determined values. Second, we notice that head movements provide valuable cues to model focus of attention. This is because when users look toward an object, they always orient their heads toward the object of interest so as to make it in

the center of their visual fields. As a result of the above analysis, head positions are integrated with eye positions as the observations of the HMM.
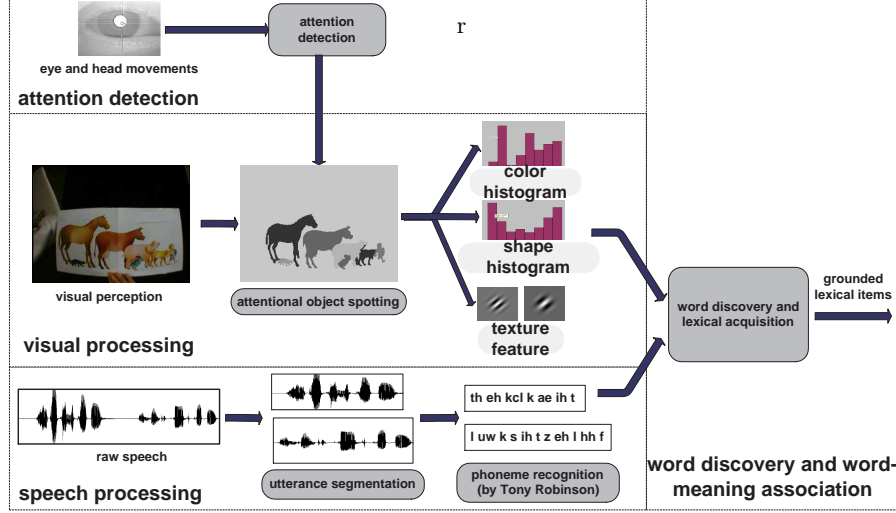
Fig. 6. The system first estimates speakers' focus of attention, then utilizes spatial-temporal correlations of multisensory input at attentional points in time to associate spoken words with their perceptually grounded meanings.

A 2-state HMM is used in our system for eye fixation finding. One state corresponds to saccade and the other represents fixation. The observations of HMM are 2-dimensional vectors consisting of the magnitudes of the velocities of head rotations in three dimensions and the magnitudes of velocities of eye movements. We model the probability densities of the observations using a two-dimensional Gaussian. As learning results, the saccade state contains an observation distribution centered around high velocities and the fixation state represents the data whose distribution is centered around low velocities. The transition probabilities for each state represent the likelihood of remaining in that state or making a transition to another state.

4.2 Word-Meaning Association

In this step, the co-occurrence of multimodal data selects meaningful semantics that associate spoken words with their grounded meanings. We take a novel view of this problem as being analogous to the word alignment problem in machine translation. For that problem, given texts in two languages (e.g. English and French), computational linguistic techniques can estimate the probability that an English word will be translated into any particular French word and then align the words in an English sentence with the words in its French translation. Similarly, for our problem, if different meanings can be viewed as elements of a "meaning language", associating meanings with object names and action verbs

can be viewed as the problem of identifying word correspondences between English and "meaning language". In light of this, a technique from machine translation can address this problem. The probability of each word is expressed as a mixture model that consists of the conditional probabilities of each word given its possible meanings. In this way, an Expectation-Maximization (EM) algorithm can find the reliable associations of spoken words and their grounded meanings that will maximize the probabilities.

The general setting is as follows: suppose we have a word set $X = \{w_1, w_2, ..., w_N\}$ and a meaning set $Y = \{m_1, m_2, ..., m_M\}$, where N is the number of word-like units and M is the number of perceptually grounded meanings. Let S be the number of spoken utterances. All data are in a set $\chi = \{(S_w^{(s)}, S_m^{(s)}), 1 \leq s \leq S\}$, where each spoken utterance $S_w^{(s)}$ consists of r words $w_{u(1)}, w_{u(2)}, ..., w_{u(r)}$, and $u(i)$ can be selected from 1 to N. Similarly, the corresponding contextual information $S_m^{(s)}$ include l possible meanings $m_{v(1)}, m_{v(2)}, ..., m_{v(l)}$ and the value of $v(j)$ is from 1 to M. We assume that every word w_n can be associated with a meaning m_m. Given a data set χ, we want to maximize the likelihood of generating the "meaning" corpus given English descriptions can be expressed as:

$$P(S_m^{(1)}, S_m^{(2)}, ..., S_m^{(S)} | S_w^{(1)}, S_w^{(2)}, ..., S_w^{(S)}) = \prod_{s=1}^{S} P(S_m^{(s)} | S_w^{(s)}) \tag{1}$$

We use the model similar to that of Brown et al. (Brown, Pietra, Pietra, & Mercer, 1994). The joint likelihood of meanings and an alignment given spoken utterances:

$$P(S_m^{(s)} | S_w^{(s)}) = \sum_a P(S_m^{(s)}, a | S_w^{(s)}) \tag{2}$$

$$= \frac{\epsilon}{(r+1)^l} \sum_{a_1=1}^{r} \sum_{a_2=1}^{r} ... \sum_{a_l=1}^{r} \prod_{j=1}^{l} t(m_{v(j)} | w_{a_{v(j)}}) \tag{3}$$

$$= \frac{\epsilon}{(r+1)^l} \prod_{j=1}^{l} \sum_{i=0}^{r} t(m_{v(j)} | w_{u(i)}) \tag{4}$$

where the alignment $a_{v(j)}, 1 \leq j \leq l$ can take on any value from 0 to r and indicate which word is aligned with jth meaning. $t(m_{v(j)} | w_{u(i)})$ is the association probability for a word-meaning pair and ϵ is a small constant.

To more directly demonstrate the role of embodied visual cues in language learning, we processed the data by another method in which the inputs of eye gaze and head movements were removed, and only audio-visual data were used for learning. Speech segmentation accuracy measures whether the beginning and the end of phoneme strings of word-like units are word boundaries. Word-meaning association accuracy (precision) measures the percentage of successfully segmented words that are correctly associated with their meanings. Considering that the system processes raw sensory data, and our embodied learning method works in an unsupervised mode without manually encoding any
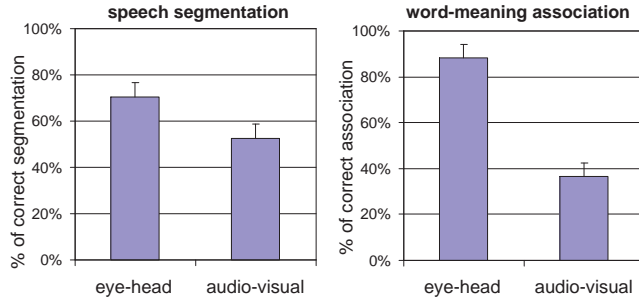
Fig. 7. A comparison of performance of the eye-head-cued method and the audio-visual approach.

linguistic information, the accuracies for both speech segmentation and word meaning association are impressive. Clearly, this embodied approach reduces the amount of information available to the learner, and it forces the model to consider all the possible meanings in a scene instead of just attended objects. In all other respects, this approach shares the same implemented components with the eye-head-cued approach. Figure 7 shows the comparison of these two methods. The eye-head-cued approach outperforms the audio-visual approach in both speech segmentation ($t(5) = 6.94$, $p < 0.0001$) and word-meaning association ($t(5) = 23.2$, $p < 0.0001$). The significant difference lies in the fact that there exist a multitude of co-occurring word-object pairs in natural environments that learning agents are situated in, and the inference of referential intentions through body movements plays a key role in discovering which co-occurrences are relevant.

To our knowledge, this work is the first model of word learning which not only learns lexical items from raw multisensory signals, but also explores the computational role of social cognitive skills in lexical acquisition. In addition, the results obtained are very much in line with the results obtained from human subjects, suggesting that not only is our model cognitively plausible, but the role of multimodal interaction can be appreciated by both human learners and by the computational model Yu, Ballard, and Aslin (2005).

5 General Discussions and Conclusions

5.1 Multimodal Learning

Recent studies in human development and machine intelligence show that the world and social signals encoded in multiple modalities play a vital role in language learning. For example, young children are highly sensitive to correlations among words and the physical properties of the world. They are also sensitive to social cues and are able to use them in ways that suggest an understanding of speaker's intent. We argue that social information can only be made manifest in correlations that arise from the physical embodiment of the mature (the mother) and immature partners (the learner) in real time. For example, the mother "jiggles" an object, the learner looks and simultaneously the mother provides the name. These time-locked "social" correlations play two roles. First,

they add multi-modal correlations that enhance and select some physical correlations making them more salient and thus learnable. Second, the computational system described above demonstrates that body movements play a crucial role in creating correlations between words and world, correlations that yield word-world mappings on the learner's part that match those intended by the speaker. Our studies show that the coupled world-word maps between the speaker and the learner -what some might call the learner's ability to infer the referential intent of the speaker - are made from simple associations in real time and the accrued results over time of learning those statistics. Critically, these statistics yield the coupled world-word maps only when they include body movements such as direction of eye gaze and points.

The present work also leads to two potentially important findings in human learning. First, our results suggest the importance of spatial information. Children need to not only share visual attention with parents at the right moment; they also need to perceive the right information at the moment. Spatio-temporal synchrony encoded in sensorimotor interaction may be provide this. Second, hands (and other body parts, such as the orientation of the body trunk) play a crucial role in signaling social cues to the other social partner. The parent's eyes are rarely in the child's visual field but the parent's and the child's own hands occupy a big proportion of the child's visual field. Moreover, the change of the child's visual field can be caused by gaze and head movement, but this change can be caused by both his own hand movements and the social partner's hand movements. In these ways, hand movements directly and significantly changes the child's view.

5.2   A New Window of the World

The first-person view is visual experience as the learner sees it and thus changes with every shift in eye gaze, every head turn, every observed hand action on an object. This view is profoundly different from that of an external observer, the third-person view, who watches the learner perform in some environment precisely because the first person view changes moment-to-moment with the learner's own movements. The systematic study of this first person view in both human learning and machine intelligence — of the dynamic visual world through the developing child's eyes – seems likely to reveal new insights into the regularities on which learning is based and on the role of action in creating those regularities. The present findings suggest that the visual information from a child's point of view is dramatically different from the parent's (or an experimenter's) viewpoint. This means analyses of third-person views from an adult perspective may be missing the most significant visual information to a young child's learning.

In artificial intelligence, our system demonstrates a new approach to developing human-computer interfaces, in which computers seamlessly integrate in our everyday lives and are able to learn lexical items by sharing user-centric multi-sensory information. The inference of speaker's referential intentions from their body movements provides constraints to avoid the large amount of irrelevant

computation and can be directly applied as deictic reference to associate words with perceptually grounded referents in the physical environment.

5.3 Human and Machine Learning

The two studies in this chapter also demonstrate that the breakthroughs in one field can bootstrap the findings in another field. Human and machine learning research shares the same goal – understanding existing intelligent systems and developing artificial systems that can simulate human intelligence. Therefore, these two fields can benefit from each other in at least two important ways. First, the findings from one field can provide useful insights to the other field. More specifically, the findings from human learning can guide us to develop intelligent machines. Second, the advanced techniques in machine intelligence can provide useful tools to analyze behavioral data and in doing so allow us to better understanding human learning. In this way, these two lines of research can co-evolve and co-develop because they intend to understand the core problems in learning and intelligence – no matter if it is human intelligence or machine intelligence. The two studies in this chapter represent the first efforts toward this goal, showing that this kind of interdisciplinary studies can indeed lead to interesting findings.

References

Baldwin, D. (1993). Early referential understanding: Infant's ability to recognize referential acts for what they are. *Developmental psychology, 29*, 832-843.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*, 1311-1328.

Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind.* Cambridge: MIT Press.

Bertenthal, B., Campos, J., & Kermoian, R. (1994). An epigenetic perspective on the development of self-produced locomotion and its consequences. *Current Directions in Psychological Science, 3*, 140-145.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: The MIT Press.

Brown, P. F., Pietra, S., Pietra, V., & Mercer, R. L. (1994). The mathematics of statistical machine translation:parameter estimation. *Computational Linguistics, 19*(2), 263-311.

Plunkett, K. (1997). Theories of early language acquisition. *Trends in cognitive sciences, 1*, 146-153.

Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.

Rabiner, L. R., & Juang, B. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Salvucci, D. D., & Anderson, J. (1998). Tracking eye movement protocols with cognitive process models. In *Proceedings of the twentieth annual conference of the cognitive science society* (p. 923-928). LEA: Mahwah, NJ.

Smith, L. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (p. 51-80). Oxford: Oxford University Press.

Steels, L., & Vogt, P. (1997). Grounding adaptive language game in robotic agents. In C. Husbands & I. Harvey (Eds.), *Proc. of the 4th european conference on artificial life.* London: MIT Press.

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development, 10*, 201-224.

Woodward, A., & Guajardo, J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development, 17*, 1061-1084.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29*(6), 961-1005.