A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions

CHEN YU
University of Rochester

and
DANA H. BALLARD
University of Rochester

We present a multimodal interface that learns words from natural interactions with users. In light of the studies of human language development, the learning system is trained in unsupervised mode in which users perform everyday tasks while providing natural language descriptions of their behaviors. 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 associates action verbs and object names with their perceptually grounded meanings. The central ideas 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 hypothesized lexical items. From those items, an EM-based method is developed to select correct word-meaning pairs. Successful learning is demonstrated in the experiments of three natural tasks: "unscrewing a jar", "stapling a letter" and "pouring water".

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning—*Language acquisition*; I.2.0 [**Artificial Intelligence**]: General—*Cognitive simulation*; H.5.2 [**Information interfaces and representation**]: User Interfaces—*Theory and methods*

General Terms: Human Factors, Experimentation

Additional Key Words and Phrases: Multimodal learning, multimodal interaction, cognitive modeling

1. INTRODUCTION

The next generation of computers is expected to interact and communicate with users in a cooperative and natural manner while users engage in everyday activities. By being situated in users' environments, intelligent computers should have basic perceptual abilities, such as understanding what people are talking about (speech recognition), what they are looking at (visual object recognition) and what they are doing (action recognition). Furthermore, similar to human counterparts, computers should acquire and then use the knowledge of associations between different perceptual inputs. For instance, spoken words of object names (sensed from auditory perception) are naturally correlated with visual appearances of the corresponding objects obtained from visual perception. Once machines have that knowledge and those abilities, they can demonstrate many human-like behaviors and perform many helpful acts. In the scenario of making a peanut-butter sandwich, for example, when a user asks for a piece of bread verbally, a computer can understand that the spoken word "bread" refers to some flat square piece on the kitchen table. Therefore, with an actuator such as a robotic arm, the machine can first locate the position of the

bread, then grasp and deliver it to the user. In another context, a computer may detect the user's attention and notice that the attentional object is a peanut butter jar, it can then utter the object name and provide information related to peanut butter by speech, such as a set of recipes or nutritional values. In a third example, a computer may be able to recognize what the user is doing and verbally describe what it sees. The ability to generate verbal descriptions of user's behaviors is a precursor to making computers communicate with users naturally. In this way, computers will seamlessly integrate into our everyday lives, and work as intelligent observers and human-like assistants.

To progress toward the goal of anthropomorphic interfaces, computers need to not only recognize the sound patterns of spoken words but also associate them with their perceptually grounded meanings. Two research fields are closely related to this topic: speech recognition and multimodal human-computer interfaces. Unfortunately, both of them only address some parts of the problem. They cannot provide a solution to the whole issue.

Most existing speech recognition systems can not achieve the goal because they purely rely on statistical models of speech and language, such as hidden Markov models [Rabiner and Juang 1989] and hybrid connectionist models [Lippmann 1989]. Typically, an automatic speech recognition system consists of a set of modules: acoustic feature extraction, acoustic modeling, word modeling and language modeling. The parameters of acoustic models are estimated using training speech data. Word models and a language model are trained using text corpora. After training, the system can decode speech signals into recognized word sequences using acoustic models, language models and word network. This kind of systems has two inherent disadvantages. First, they require a training phase in which large amounts of spoken utterances paired with manually labeled transcriptions are needed to train the model parameters. This training procedure is time-consuming and needs human expertise to label spoken data. Second, these systems transform acoustic signals to symbolic representations (texts) without regard to their perceptually grounded meanings. Humans need to interpret the meanings of these symbols based on our own knowledge. For instance, a speech recognition system can map the sound pattern "jar" to the string "jar", but it does not know its meaning.

In multimodal human-computer interface studies, researchers mainly focus on the design of multimodal systems with performance advantages over unimodal ones in the context of different types of human-computer interaction [Oviatt 2002]. The technical issue here is multimodal integration – how to integrate signals in different modalities. There are two types of multimodal integration, one is to merge signals at the sensory level and the other at a semantic level. The first approach is most often used in such applications that the data is closely coupled in time, such as speech and lip movements. At each timestamp, several features extracted from different modalities are merged to form a higher-dimensional representation, which is then used as input of the classification system usually based on multiple HMMs or temporal neural networks. Multimodal systems using semantic fusion include individual recognizers and a sequential integration process. These individual recognizers can be trained using unimodal data, which can then be integrated directly without re-training. Integration is thus an assembling process that occurs after each unimodal processing system has already made decisions based on the individual inputs. However, no matter based on feature or semantic fusion, most systems do not have *learning* ability in the sense that developers need to encode knowledge into some symbolic representations or probabilistic models during the training phase. Once the systems are trained, they are not

able to automatically gain additional knowledge even though they are situated in physical environments and can obtain multisensory information.

We argue that the shortcomings described above lie in the fact that sensory perception and knowledge acquisition of machine are quite different from those of human counterparts. For instance, humans learn language based on their sensorimotor experiences with the physical environment. We learn words by sensing the environment through our perceptual systems, which do not provide the labeled or preprocessed data. Different levels of abstraction are necessary to efficiently encode those sensorimotor experiences, and one vital role of human brain is to map those embodied experiences with linguistic labels (symbolic representations). Therefore, to communicate with humans in daily life, a challenge in machine intelligence is how to acquire the semantics of words in a language from cognitive and perceptual experiences. This challenge is relevant to the symbol grounding problem [Harnad 1990]: establishing correspondences between internal symbolic representations in an intelligent system situated in the physical world (e.g., a robot or an embodied agent) and sensory data collected from the environment. We believe that computationally modeling how humans ground semantics is a key to understanding our own minds and ultimately creating embodied learning machines.

This paper describes a multimodal learning system that is able to learn perceptually grounded meanings of words from users' everyday activities. The only requirement is that users need to describe their behaviors verbally while performing those day-to-day tasks. To learn a word (shown in Figure 1), the system needs to discover its sound pattern from continuous speech, recognize its meaning from non-speech context, and associate these two. Since no manually labeled data is involved in the learning process, the range of problems we need to address in this kind of word learning is substantial. To make concrete progress, this paper focuses on how to associate visual representations of objects with their spoken names and map body movements to action verbs.

In our system, perceptual representations are extracted from sensory data and used as perceptually grounded meanings of spoken words. This is based on evidence that from an early age, human language learners are able to form perceptually-based categorical representations [Quinn et al. 1993]. Those categories are highlighted by the use of common words to refer to them. Thus, the meaning of the word "dog" corresponds to the category of dogs, which is a mental representation in the brain. Furthermore, [Schyns and Rodet] argued that the representations of object categories emerge from the features that are perceptually learned from visual input during the developmental course of object recognition and categorization. In this way, object naming by young children is essentially about mapping words to selected perceptual properties. Most researchers agree that young language learners generalize names to new instances on the basis of some similarity but there are many debates about the nature of similarity (see a review in Landau et al. 1998). It has been shown that shape is generally attended to for solid rigid objects, and children attend to other specific properties, such as texture,size or color, of the objects that have eyes or are not rigid [Smith et al. 1996]. In light of the perceptual nature of human categorization, our system represents object meanings as perceptual features consisting of shape, color and texture features extracted from the visual appearances of objects. The categories of objects are formed by clustering those perceptual features into groups. Our system then chooses the centroid of each category in the perceptual feature space as a representation of the meaning of this category, and associate this feature representation with linguistic

labels. The meanings of actions verbs are described in terms of motion profiles in our system, which do not encapsulate inferences about causality ,function and force dynamics (see [Siskind 2001] for an good example). We understand that those meanings for object names and action verbs (mental representations in our computational system) are simplified and may not be the exact same with the concepts in the brain (mental representations in the user's brain) because it depends on how machines judge the content of the user's mental states when he/she utters the speech. In addition, many human concepts cannot be simply characterized in easy perceptual terms (see further discussions about concepts from different views in [Gopnik and Meltzoof 1997; Keil 1989]). However, as long as we agree that meanings are some mental entities in the user's brain and that the cognitive structures in the user's brain are connected to his/her perceptual mechanisms, then it follows that meanings should be at least partially perceptually grounded. Since we focus on automatic language learning but not concept learning in this work, a hypothesis here is that the form we store perceptions has the same form as the meanings of words [Gardenfors 1999]. Therefore, we use the form of perceptual representation that can be directly extracted from sensory data to represent meanings.

To learn perceptually grounded semantics, the essential ideas of our system are to identify the sound patterns of individual words from continuous speech using non-linguistic contextual information and employ body movements as deictic references to discover word-meaning associations. Our work suggests a new trend in developing human-computer interfaces that can automatically learn spoken language by sharing user-centric multisensory information. This advent represents the beginning of an ongoing progression toward computational systems capable of human-like sensory perception [Weng et al. 2001].

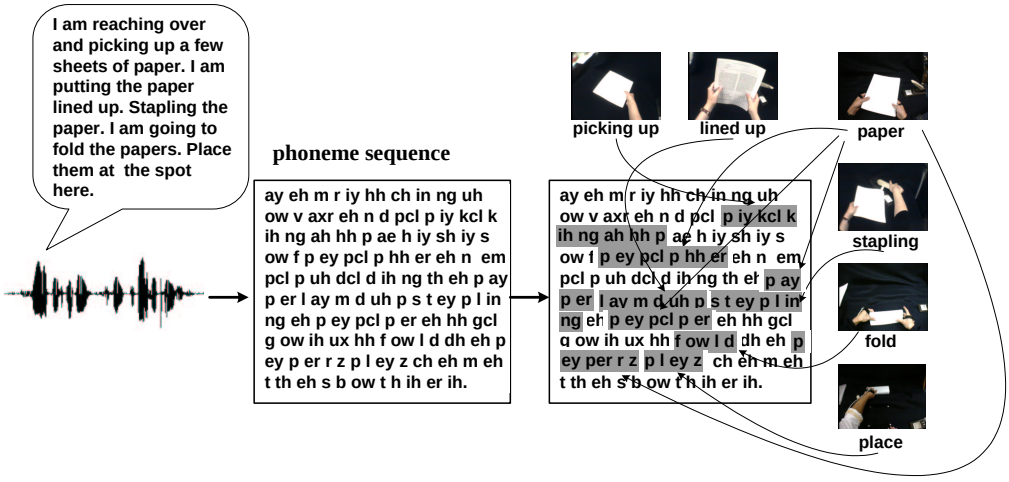


Fig. 1. **The problems in word learning.** The raw speech is first converted to phoneme sequences. The goal of our method is to discover phoneme substrings that correspond to the sound patterns of words and then infer the grounded meanings of those words from non-speech modalities.

2. BACKGROUND

Language is about symbols and humans ground those symbols in sensorimotor experiences during their development [Lakoff and Johnson 1980]. To develop a multimodal learning interface for word acquisition, it is helpful to make use of our knowledge of human language development to guide the design of our approach. English-learning infants first display some ability to segment words at about 7.5 months [Jusczyk and Aslin 1995]. By 24 months, the speed and accuracy with which infants identify words in fluent speech is similar to that of native adult listeners. A number of relevant cues have been found that are correlated with the presence of word boundaries and can potentially signal word boundaries in continuous speech (see [Jusczyk 1997] for a review). Around 6 to 12 months is the stage of grasping the first words. A predominant proportion of most children's first vocabulary (the first 100 words or so), in various languages and under varying child-rearing conditions, consist of object names, such as food, clothing and toys. The second large category is the set of verbs that is mainly limited to action terms. Gillette et al. [Gillette et al. 1999] showed that learnability of a word is primarily based upon its imageability or concreteness. Therefore, most object names and action verbs are learned before other words because they are more observable and concrete. Next, infants move to the stage of vocabulary spurt or naming explosion, in which they start learning large amounts of words much more rapidly than before. At the meanwhile, grammar gradually emerges from the lexicon, both of which share the same mental-neural mechanisms [Bates and Goodman 1999]. Many of the later learned words correspond to abstract notions (e.g., noun:"idea", verb:"think") and are not directly grounded in embodied experiences. However, Lakoff and Johnson [Lakoff and Johnson 1980] proposed that all human understanding is based on metaphorical extension of how we perceive our own bodies and their interactions with the physical world. Thus, the initial and imageable words directly grounded in physical embodiment serve as a foundation for the acquisition of abstract words and syntax that become indirectly grounded through their relations to those grounded words. Therefore, the initial stage of language acquisition, in which infants deal primarily with the grounding problem, is critical in this semantic bootstrapping procedure because it provides a sensorimotor basis for further development.

The experimental studies have yielded insights into perceptual abilities of young children and provided informative constraints in building computational systems that can acquire language automatically. Recent computational models address the problems of both speech segmentation and lexical learning. A good survey of the related computational studies of speech segmentation can be found in [Brent 1999], in which several methods are explained, their performance in computer simulations is summarized, and behavioral evidence bearing on them is discussed. Among them, Brent and Cartwright [Brent and Cartwright 1996] have encoded information of distributional regularity and phonotactic constraints in their computational model. Distributional regularity means that sound sequences occurring frequently and in a variety of contexts are better candidates for the lexicon than those that occur rarely or in few contexts. The phonotactic constraints include both the requirement that every word must have a vowel and the observation that languages impose constraints on word-initial and word-final consonant clusters. Most computational studies, however, use phonetic transcriptions of text as input and do not deal with raw speech. From a computational perspective, they simplified the problem by not coping with the acoustic variability of spoken words in different contexts and by various talkers. As

a result, their methods cannot be directly applied to develop computational systems that acquire lexicons from raw speech.

Siskind [Siskind 1995] developed a mathematical model of lexical learning based on cross-situational learning and the principle of contrast, which learned word-meaning associations when presented with paired sequences of pre-segmented tokens and semantic representations. Regier's work [Regier 1996] was about modeling how some lexical items describing spatial relations might develop in different languages. Bailey [Bailey 1997] proposed a computational model that learns to not only produce verb labels for actions but also carry out actions specified by verbs that it has learned. A good review of word learning models can be found in [Regier 2003]. Different from most other symbolic models of vocabulary acquisition, physical embodiment has been appreciated by the works of [Roy 2002; Roy and Pentland 2002] and [Steels and Vogt 1997]. Steels and Vogt showed how a coherent lexicon may spontaneously emerge in a group of robots engaged in language games and how a lexicon may adapt to cope with new meanings that arise. Roy and Pentland [Roy and Pentland 2002] implemented a model of early language learning which can learn words and their semantics from raw sensory input. They used the temporal correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. However, the associated visual and audio corpses are collected separately from different experimental setups in Roy's system. Specifically, audio data are gathered from infant-caregiver interactions while visual data of individual objects are captured by a CCD camera on a robot. Thus, audio and visual inputs are manually correlated based on the co-occurrence assumption, which claims that words are always uttered when their referents are perceived. Roy's work is groundbreaking but leaves two important areas for improvement. The first is that the co-occurrence assumption has not been verified by experimental studies of human language learners (e.g., infants learning their native language [Bloom 2000]). We argue that this assumption is not reliable and appropriate for modeling human language acquisition and statistical learning of audio-visual data is unlikely to be the whole story for automatic language acquisition. The second issue is that Roy's work does not include the intentional signals of the speaker when he/she utters the speech. We show that they can provide pivotal constraints to improve performance.

3. A MULTIMODAL LEARNING INTERFACE

Recent psycholinguistic studies (e.g., [Baldwin et al. 1996]; [Bloom 2000]; [Tomasello 2000]) have shown that a major source of constraint in language acquisition involves social cognitive skills, such as children's ability to infer the intentions of adults as adults act and speak to them. These kinds of social cognition are called mind reading by [Baron-Cohen 1995]. Bloom [Bloom 2000] argued that children's word learning actually draws extensively on their understanding of the thoughts of speakers. His claim has been supported by the experiments in which young children were able to figure out what adults were intending to refer to by speech. In a complementary study of embodied cognition, Ballard and colleagues [Ballard et al. 1997] proposed that orienting movements of the body play a crucial role in cognition and form a useful computational level, termed the embodiment level. At this level, the constraints of the body determine the nature of cognitive operations, and the body's pointing movements are used as deictic references to bind objects in the physical environment to cognitive programs of our brains. Also, in the studies of speech production, Meyer et al. [Meyer et al. 1998] showed that the speakers' eye move-

ments are tightly linked to their speech output. They found that when speakers were asked to describe a set of objects from a picture, they usually looked at each new object before mentioning it, and their gazes remained on the object until they are about to say the last word about it.

By putting together the findings from these cognitive studies, we propose that speakers' body movements, such as eye movements, head movements and hand movements, can reveal their referential intentions in verbal utterances, which could play a significant role in automatic language acquisition in both computational systems and human counterparts [Yu et al. 2003; Yu and Ballard 2003]. To support this idea, we provide an implemented system to demonstrate how inferences of speaker's referential intentions from their body movements, which we term *embodied intention*, can facilitate acquiring grounded lexical items. In our multimodal learning interface, a speaker's referential intentions are estimated and utilized to facilitate lexical learning in two ways. First, possible referential objects in time provide cues for word spotting from a continuous speech stream. Speech segmentation without prior language knowledge is a challenging problem and has been addressed by solely using linguistic information. In contrast, our method emphasizes the importance of non-linguistic contexts in which spoken words are uttered. We propose that the sound patterns frequently appearing in the same context are likely to have grounded meanings related to this context. Thus, by finding frequently uttered sound patterns in a specific context (e.g., an object that users intentionally attend to), the system discovers word-like sound units as candidates for building lexicons. Second, a difficult task of word learning is to figure out which entities specific words refer to from a multitude of co-occurrences between spoken words (from auditory perception) and things in the world (from non-auditory modalities, such as visual perception). This is accomplished in our system by utilizing speakers' intentional body movements as deictic references to establish associations between spoken words and their perceptually grounded meanings.

To ground language, 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 2. 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. To learn words from user's spoken descriptions, three fundamental problems needed to be addressed are: (1) action recognition and object recognition to provide grounded meanings of words encoded in non-speech contextual information, (2) speech segmentation and word spotting to extract the sound patterns that correspond to words, (3) association between

spoken words and their perceptually grounded meanings.

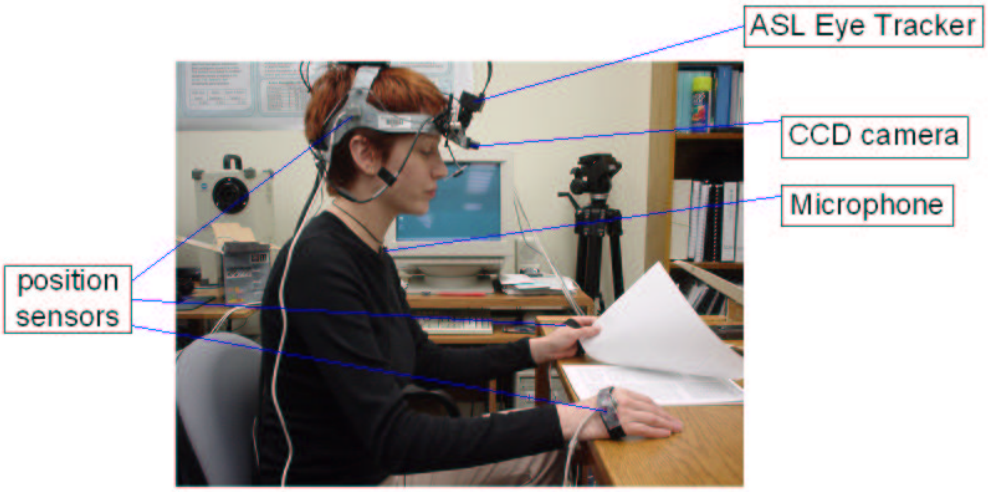


Fig. 2. The learning system shares multisensory information with a real agent in a first-person sense. This allows the association of coincident signals in different modalities.

4. REPRESENTING AND CLUSTERING NON-SPEECH PERCEPTUAL INPUTS

The non-speech inputs of the system consist of visual data from a head-mounted camera, head and hand positions in concert with gaze-in-head data. Those data provide contexts in which spoken utterances are produced. Thus, the possible meanings of spoken words that users utter are encoded in those contexts, and we need to extract those meanings from raw sensory inputs. Specifically, the system should spot and recognize actions from user's body movements, and discover the objects of user interest. In implementation, we observe that in accomplishing well-learned tasks, the user's focus of attention is linked with body movements. In light of this, our method first uses eye and head movements as cues to estimate the user's focus of attention. Attention, as represented by gaze fixation, is then utilized for spotting the target objects of user interest. Attention switches are calculated and used to segment a sequence of hand movements into action units which are then categorized by Hidden Markov Models (HMMs). The results are two temporal sequences of perceptually grounded meanings (objects and actions) as depicted by the box labeled "contextual information" in Figure 9.

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 and 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 and 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 his 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 HMM.

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. The parameters of HMMs needing to be estimated comprise the observation and transition probabilities. Specifically, we need to compute the means (μ_{j_1}, μ_{j_2}) and variances $(\sigma_{j_1}, \sigma_{j_2})$ of two-dimensional Gaussian for s_j state and the transition probabilities between two states. The estimation problem concerns how to adjust the model λ to maximize $P(O \mid \lambda)$ given an observation sequence O of gaze and head motions. We can initialize the model with flat probabilities, then the forward-backward algorithm [Rabiner and Juang 1989] allows us to evaluate this probability. Using the actual evidence from the training data, a new estimate for the respective output probability can be assigned:

$$\bar{\mu}_j = \frac{\sum_{t=1}^{T} \gamma_t(j) O_t}{\sum_{t=1}^{T} \gamma_t(j)} \tag{1}$$

and

$$\bar{\sigma}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)(O_t - \bar{\mu}_j)(O_t - \bar{\mu}_j)^T}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2}$$

where $\gamma_t(j)$ is defined as the posterior probability of being in state s_j at time t given the observation sequence and the model.

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. An example of the results of eye data analysis is shown in Figure 3.

4.2 Attentional Object Spotting

Knowing attentional states allows for automatic object spotting by integrating visual information with eye gaze data. For each attentional point in time, the object of user interest is discovered from the snapshot of the scene. Multiple visual features are then extracted from the visual appearance of the object which are used for object categorization. Figure 4 shows an overview of our approach [Yu et al. 2002] .

4.2.1 *Object Spotting.* Attentional object spotting consists of two steps. First, the snapshots of the scene are segmented into blobs using ratio-cut [Wang and Siskind 2003].

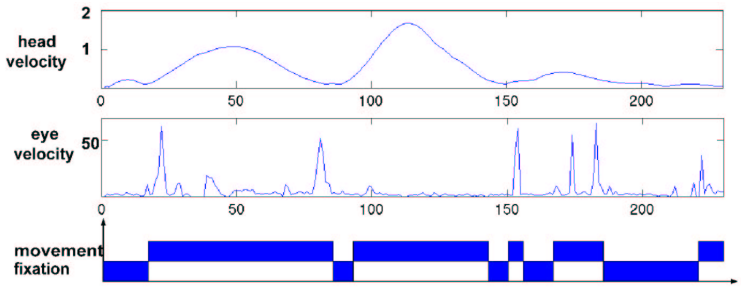


Fig. 3. **Eye fixation finding. The top plot:** The speed profile of head movements. **The middle plot:** Point-to-point magnitude of velocities of eye positions. **The bottom plot:** A temporal state sequence of HMM (the label "fixation" indicates the fixation state and the label "movement" represents the saccade state).

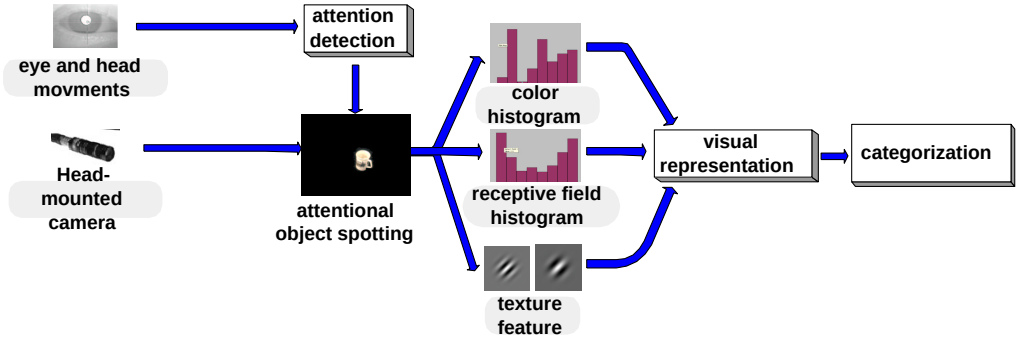


Fig. 4. The overview of attentional object spotting

The result of image segmentation is illustrated in Figure 6(b) and only blobs larger than a threshold are used. Next, we group those blobs into several semantic objects. Our approach starts with the original image, uses gaze positions as seeds and repeatly merges the most similar regions to form new groups until all the blobs are labeled. Eye gaze in each attentional time is then utilized as a cue to extract the object of user interest from all the detected objects.

We use color as the similarity feature for merging regions. $L*a*b$ color space is adopted to overcome undesirable effects caused by varied lighting conditions and achieve more robust illumination-invariant segmentation. $L*a*b$ color consists of a luminance or lightness component (L*) and two chromatic components: the a* component (from green to red) and the b* component (from blue to yellow). To this effect, we compute in the $L*a*b$ color space the similarity distance between two blobs and employ the histogram intersection method proposed by [Swain and Ballard 1991]. If C_A and C_B denote the color histograms of two regions A and B, their histogram intersection is defined as:

$$h(A,B) = \frac{\sum_{i=1}^{n} min(C_A^i, C_B^i)}{\sum_{i=1}^{n}(C_A^i + C_B^i)} \tag{3}$$

where n is the number of bin in color histogram, and $0 < h(A,B) < 0.5$. Two neighboring regions are merged into a new region if the histogram intersection $h(A,B)$ is between a threshold $t_c(0 < t_c < 0.5)$ and 0.5. While this similarity measure is fairly simple, it

Algorithm: object segmentation based on gaze fixations
Initialization:
Compute the color histogram of each region.
Label seed regions according to the positions of gaze fixations.
Merge seed regions that are neighbors to each other and are close with respect to their similarity.
Put neighboring regions of seed regions in the SSL.
Merging:
While the SSL is not empty
Remove the top region A from SSL.
Compare the similarity between A and all the regions in S_A and find the closest seed region B.
Merge the regions A and B and compute the color histogram of new region $I = A \cup B$.
Test each neighboring region A_i of A:
If A_i is labeled as a seed region
Merge the region with I if they are similar.
Otherwise
Add the region to the SSL according to its color similarity with I, $h(A_i, I)$.

Fig. 5. The algorithm for merging blobs

is remarkably effective in determining color similarity between regions of multi-colored objects.

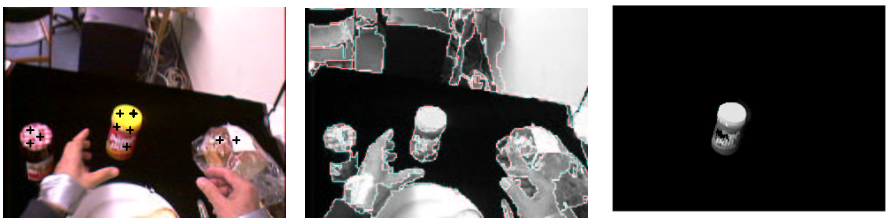
The approach of merging blobs is based on a set of regions selected by a user's gaze fixations, termed seed regions. We start with a number of seed regions $S_1, S_2, ..., S_n$, in which n is the number of regions that the user was attending to. Given those seed regions, the merging process then finds a grouping of the blobs into semantic objects with the constraint that the regions of a visual object are chosen to be as homogeneous as possible. The process evolves inductively from the seed regions. Each step involves the addition of one blob to one of the seed regions and the merging of neighbor regions based on their similarities.

In the implementation, we make use of a Sequentially Sorted List (SSL) [Adams and Bischof 1994] that is a linked list of blobs ordered according to some attribute. In each step of our method, we consider the blob at the beginning of the list. When adding a new blob to the list, we place it according to its value of the ordering attribute so that the list is always sorted based on the attribute. Let $S_A = \{S_A^1, S_A^2, ..., S_A^n\}$ be the set of immediate neighbors of the blob A, which are seed regions. For all the regions in S_A, the seed region that is closest to A is defined as:

$$B = \arg\max_i h(A, S_A^i); 1 \leq i \leq n \tag{4}$$

where $h(A, S_A^i)$ is the similarity distance between region A and S_A^i based on the selected similarity feature. The ordering attribute of region A is then defined as $h(A, B)$. The merging procedure is illustrated in Figure 5. Figure 6 shows how these steps are combined to get an attentional object.

4.2.2 *Object Representation and Categorization.* The visual representation of the extracted object contains color, shape and texture features. Based on the works of [Mel 1997], we construct the visual features of objects which are large in number, invariant to different viewpoints, and driven by multiple visual cues. Specifically, 64-dimensional color features are extracted by a color indexing method [Swain and Ballard 1991], and 48-dimensional shape features are represented by calculating histograms of local shape properties [Schiele and Crowley 2000]. The Gabor filters with three scales and five orientations are applied



(a) (b) (c)

Fig. 6. **Left:** The snapshot image with eye positions (black crosses). **Middle:** The results of low-level image segmentation. **Right:** Combining eye position data with the segmentation to extract an attended object.

to the segmented image. It is assumed that the local texture regions are spatially homogeneous, and the mean and the standard deviation of the magnitude of the transform coefficients are used to represent an object in a 48-dimensional texture feature vector. The feature representations consisting of a total of 160 dimensions are formed by combining color, shape and texture features, which provide fundamental advantages for fast, inexpensive recognition. Most pattern recognition algorithms, however, do not work efficiently in high dimensional spaces because of the inherent sparsity of the data. This problem has been traditionally referred to as the dimensionality curse. In our system, we reduced the 160-dimensional feature vectors into the vectors of dimensionality 15 by principle component analysis (PCA), which represents the data in a lower dimensional subspace by pruning away those dimensions with the least variance. Next, since the feature vectors extracted from visual appearances of attentional objects do not occupy a discrete space, we vector quantize them into clusters by applying a hierarchical agglomerative clustering algorithm [Hartigan 1975]. Finally, we select a prototype to represent perceptual features of each cluster.

4.3 Segmenting and Clustering Motion Sequences

Recent results in visual psychophysics [Land et al. 1999; Hayhoe 2000; Land and Hayhoe 2001] indicate that in natural circumstances, the eyes, the head, and hands are in continual motion in the context of ongoing behavior. This requires the coordination of these movements in both time and space. Land et al. [Land et al. 1999] found that during the performance of a well-learned task (making tea), the eyes closely monitor every step of the process although the actions proceed with little conscious involvement. Hayhoe [Hayhoe 2000] has shown that eye and head movements are closely related to the requirements of motor tasks and almost every action in an action sequence is guided and checked by vision, with eye and head movements usually preceding motor actions. Moreover, their studies suggested that the eyes always look directly at the objects being manipulated. In our experiments, we confirm the conclusions by Hayhoe and Land. For example, in the action of “picking up a cup”, the subject first moves the eyes and rotates the head to look toward the cup while keeping the eye gaze at the center of view. The hand then begins to move toward the cup. Driven by the upper body movement, the head also moves toward the location while the hand is moving. When the arm reaches the target place, the eyes are fixating on it to guide the action of grasping.

Despite the recent discoveries of the coordination of eye, head and hand movements in

cognitive studies, little work has been done in utilizing these results for machine understanding of human behavior. In this work, our hypothesis is that eye and head movements, as an integral part of the motor program of humans, provide important information for action recognition of human activities. We test this hypothesis by developing a method that segments action sequences based on the dynamic properties of eye gaze and head direction, and applies Dynamic Time Warping (DTW) and HMM to cluster temporal sequences of human motion [Yu and Ballard 2002a; 2002b].

Humans perceive an action sequence as several action units [Kuniyoshi and Inoue 1993]. This gives rise to the idea that the segmentation of a continuous action stream into action primitives is the first step toward understanding human behaviors. With the ability to track the course of gaze and head movements, our approach uses gaze and head cues to detect user-centric attention switches that can then be utilized to segment human action sequences.

We observe that actions can occur in two situations: during eye fixations and during head fixations. For example, in a "picking up" action, the performer focuses on the object first firstly, then the motor system moves a hand to approach it. During the procedure of approaching and grasping, the head moves toward the object as the result of upper body movements, but eye gaze remains stationary on the target object. The second case includes such actions as "pouring water" in which the head fixates on the object involved in the action. During the head fixation, eye-movement recordings show that there can be a number of eye fixations. For example, when the performer is pouring water, he spends five fixations on the different parts of the cup and one look-ahead fixation to the location where he will place the waterpot after pouring. In this situation, the head fixation is a better cue than eye fixations to segment the actions.

Based on the above analysis, we develop an algorithm for action segmentation, which consists of the following three steps:

(1) **Head fixation finding** is based on the orientations of the head. We use 3D orientations to calculate the speed profile of the head, as shown in the first two rows of Figure 7.

(2) **Eye fixation finding** is accomplished by a velocity-threshold-based algorithm. A sample of the results of eye data analysis is shown in the third and fourth rows of Figure 7.

(3) **Action Segmentation** is achieved by analyzing head and eye fixations, and partitioning the sequence of hand positions into action segments (shown in the bottom row of Figure 7) based on the following three cases:
   —A head fixation may contain one or multiple eye fixations. This corresponds to actions, such as "unscrewing". "Action 3" in the bottom row of Figure 7 represents this kind of action.
   —During the head movement, the performer fixates on the specific object. This situation corresponds to actions, such as "picking up". "Action 1" and "Action 2" in the bottom row of Figure 7 represent this class of actions.
   —During the head movement, eyes are also moving. It is most probable that the performer is switching attention after the completion of the current action.

We collect the raw position (x, y, z) and the rotation (h, p, r) data of each action unit from which feature vectors are extracted for recognition. We want to recognize the types of motion not the accurate trajectory of the hand because the same action performed by different people varies. Even in different instances of a simple action of "picking up" performed

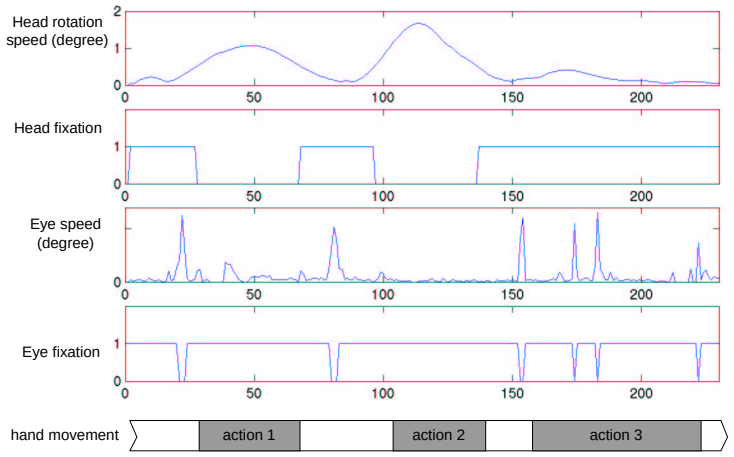


Fig. 7. **Segmenting actions based on head and eye fixations. The first two rows:** point-to-point speeds of head data and the corresponding fixation groups (1–fixating, 0–moving).**The third and fourth rows:** eye movement speeds and the eye fixation groups (1–fixating, 0–moving) after removing saccade points. **The bottom row:** the results of action segmentation by integrating eye and head fixations.

by the same person, the hand goes roughly in different trajectories. This indicates that we can not directly use the raw position data to be the features of the actions. As pointed out by Campbell et al. [Campbell et al. 1996], features designed to be invariant to shift and rotation perform better in the presence of shifted and rotated input. The feature vectors should be chosen so that large changes in the action trajectory produce relatively small excursions in the feature space, while the different types of motion produce relatively large excursions. In the context of our experiment, we calculated three element feature vectors consisting of the hand's speed on the table plane ($d\sqrt{x^2+y^2}$), the speed in the z-axis, and the speed of rotation in the 3 dimensions ($d\sqrt{h^2+p^2+r^2}$).

Let S denote a hand motion trajectory that is a multivariate time series spanning n time steps such that $S = \{s_t \mid 1 \leq t \leq n\}$. s_t is a vector of values containing one element for the value of each of the component univariate time series at time t. Given a set of m multivariate time series of hand motion, we want to obtain in an unsupervised manner a partition of these time series into subsets such that each cluster corresponds to a qualitatively different regime. Our clustering approach is based on the combination of HMM (described briefly in Section 4.1) and Dynamic Time Warping [Oates et al. 1999]. Given two time series S_1 and S_2, DTW finds the warping of the time dimension in S_1, which minimizes the difference between two series.

We model the probability of individual observation (a time series S) as generated by a finite mixture model of K component HMMs [Smyth 1997]:

$$f(S) = \sum_{k=1}^{K} p_k(S|c_k)p(c_k) \tag{5}$$

where $p(c_k)$ is the prior probability of kth HMM and $p_k(S|c_k)$ is the generative probability given the kth HMM with its transition matrix, observation density parameters, and initial

state probabilities. $p_k(S|c_k)$ can be computed via the forward part of the forward-backward procedure. Assume that the number of clusters K is known, the algorithm for clustering sequences into K groups can be described in terms of three steps:

—given m time series, construct a complete pairwise distance matrix by invoking DTW $m(m-1)/2$ times. Use the distance matrix to cluster the sequences into K groups by employing a hierarchical agglomerative clustering algorithm [Hartigan 1975].

—fit one HMM for each individual group and train the parameters of the HMM. $p(c_k)$ is initialized to M_k/M where M_k is the number of sequences which belong to cluster k.

—iteratively reestimate the parameters of all the k HMMs in the Baum-welch fashion using all of the sequences [Rabiner and Juang 1989]. The weight that a sequence S has in the reestimation of kth HMM is proportional to the log-likelihood probability of the sequence given that model $\log p_k(S|c_k)$. Thus, sequences with bigger generative probabilities for a HMM have greater influence in reestimating the parameters of that HMM.

The intuition of the procedure is as follows: since the Baum-Welch algorithm is hill-climbing the likelihood surface, the initial conditions critically influence the final results. Therefore, DTW-based clustering is used to get a better estimate of the initial parameters of HMMs so that the Baum-Welch procedure will not converge to a local maximum only. In the reestimation, sequences that are more likely generated by a specific model cause the parameters of that HMM to change in such a way that it further fits for modeling a specific group of sequences.

5. SPEECH PROCESSING

This section presents the methods of phoneme recognition and phoneme string comparison [Ballard and Yu 2003], which provide a basis for word-meaning association.

5.1 Phoneme Recognition

An endpoint detection algorithm was implemented to segment a speech stream into several spoken utterances. Then the speaker-independent phoneme recognition system developed by Robinson [Robinson 1994] is employed to convert spoken utterances into phoneme sequences. The method is based on Recurrent Neural Networks (RNN) that perform the mapping from a sequence of the acoustic features extracted from raw speech to a sequence of phonemes. The training data of RNN are from the TIMIT database — phonetically transcribed American English speech — which consists of read sentences spoken by 630 speakers from eight dialect regions of the United States. To train the networks, each sentence is presented to the recurrent back-propagation procedure. The target outputs are set using the phoneme transcriptions provided in the TIMIT database. Once trained, a dynamic programming match is made to find the most probable phoneme sequence of a spoken utterance (e.g. the boxes labeled with “phoneme strings” in Figure 9).

5.2 Comparing Phoneme Sequences

The comparison of phoneme sequences has two purposes in our system: one is to find the longest similar substrings of two phonetic sequences (word-like units spotting described in Subsection 6.1), and the other is to cluster segmented utterances represented by phoneme sequences into groups (word-like units clustering presented in Subsection 6.2). In both

Algorithm: phonetic string comparison
for $i = 0$ to m **do**
  $s_{i0} = 0$
end for
for $j = 0$ to n **do**
  $s_{0j} = 0$
end for
for $i = 0$ to m **do**
  for $j = i - r$ to $i + r$ **do**
    $S_{ij} = max(S_{i-1,j-3} + w[a_i, a_{j-2}] + \frac{1}{2}w[a_i, b_{j-1}] + \frac{1}{2}w[a_i, b_j]$
      $S_{i-1,j-2} + w[a_i, b_{j-1}] + \frac{1}{2}w[a_i, b_j]$,
      $S_{i-1,j} + w_{ins}[a_i]$,
      $S_{i-1,j-1} + w[a_i, b_j]$,
      $S_{i,j-1} + w_{del}[b_j]$,
      $S_{i-2,j-1} + w[a_{i-1}, a_j] + \frac{1}{2}w[a_i, b_j]$,
      $S_{i-3,j-1} + w[a_{i-2}, a_j] + \frac{1}{2}w[a_{i-1}, b_j] + \frac{1}{2}w[a_i, b_j])$
  end for
end for

Fig. 8. The algorithm for computing the similarity of two phonetic strings

cases, an algorithm of the alignment of phoneme sequences is a necessary step. Given raw speech input, the specific requirement here is to cope with the acoustic variability of spoken words in different contexts and by various talkers. Due to this variation, the outputs of the phoneme recognizer described above are noisy phoneme strings that are different from phonetic transcriptions of text. In this context, the goal of phonetic string matching is to identify sequences that might be different actual strings, but have similar pronunciations.

5.2.1 *Similarity between individual phonemes.* To align phonetic sequences, we first need a metric for measuring distances between phonemes. We represent a phoneme by a 12-dimensional binary vector in which every entry stands for a single articulatory feature called a distinctive feature. Those distinctive features are indispensable attributes of a phoneme that are required to differentiate one phoneme from another in English [Ladefoged 1993]. In a feature vector, the number one represents the presence of a feature in a phoneme and zero represents the absence of that feature. When two phonemes differ by only one distinctive feature, they are known as being minimally distinct from each other. For instance, phonemes /p/ and /b/ are minimally distinct because the only feature that distinguishes them is "voicing". We compute the distance $d(i, j)$ between two individual phonemes as the Hamming distance which sums up all value differences for each of the 12 features in two vectors. The underlying assumption of this metric is that the number of binary features in which two given sounds differ is a good indication of their proximity. Moreover, phonological rules can often be expressed as a modification of a limited number of feature values. Therefore, sounds that differ in a small number of features are more likely to be related.

We compute the similarity matrix that consists of $n \times n$ elements where n is the number of phonemes. Each element is assigned to a score which represents the similarity of two phonemes. The diagonal elements are set to a positive value r as the rewards of matching (with the same phoneme). The other elements in the matrix are assigned to negative values $-d(i, j)$ which correspond to the distances of distinctive features between two phonemes.

5.2.2 *Alignment of two phonetic sequences.* The outputs of the phoneme recognizer are phonetic strings with timestamps of the beginning and the end of each phoneme. We subsample the phonetic strings so that symbols in the resulting strings contain the same duration. The concept of similarity is then applied to compare phonetic strings. A similarity scoring scheme assigns positive scores to pairs of matching segments and negative scores to pairs of dissimilar segments. The optimal alignment is the one that maximizes the overall score. The advantage of the similarity approach is that it implicitly includes the length information in comparing the segments. Fundamental to the algorithm is the notion of string-changing operations of Dynamic Programming (DP). To determine the extent to which two phonetic strings differ from each other, we define a set of primitive string operations, such as insertion and deletion. By applying those string operations, one phonetic string is aligned with the other. Also, the cost of each operation allows the measurement of the similarity of two phonetic strings as the sum of the cost of individual string operations in alignment and the reward of matching symbols. To identify the phonetic strings that may be of similar pronunciation, the method needs to consider both the duration and the similarity of phonemes. Thus, each phonetic string is subject not only to alternation by the usual additive random error but also to variations in speed (the duration of the phoneme being uttered). Such variations can be considered as compression and expansion of phoneme with respect to the time axis. In addition, additive random error may also be introduced by interpolating or deleting original sounds. One step toward dealing with such additional difficulties is to perform the comparison in a way that allows for deletion and insertion operations as well as compression and expansion ones. In the case of an extraneous sound that does not delay the normal speech but merely conceals a bit of it, deletion and insertion operations permit the concealed bit to be deleted and the extraneous sound to be inserted, which is a more realistic and perhaps more desirable explanation than that permitted by additive random error.

The details of the phoneme comparison method are as follows: given two phoneme sequences $a_1, a_2, ..., a_m$ and $b_1, b_2, ..., b_n$, of length m and n respectively, to find the optimal alignment of two sequences using dynamic programming, we construct an m-by-n matrix where the (i^{th}, j^{th}) element of the matrix contains the similarity score $S(a_i, b_j)$ that corresponds to the shortest possible time-warping between the initial subsequences of a and b containing i and j elements, respectively. $S(a_i, b_j)$ can be recurrently calculated in an ascending order with respect to coordinates i and j, starting from the initial condition at $(1, 1)$ up to (m, n). One additional restriction is applied on the warping process:

$$j - r \leq i \leq j + r \tag{6}$$

where r is an appropriate positive integer called window length. This adjustment window condition avoids undesirable alignment caused by a too excessive timing difference.

Let w be the metric of the similarity score and $w_{del}[a_i] = min(w[a_i, a_{i-1}], w[a_i, a_{i+1}])$ and $w_{ins}[b_j] = min(w[b_j, b_{j-1}], w[b_j, b_{j+1}])$. Figure 8 contains our DP algorithm to compute the similarity score of two phonetic strings.

6. WORD LEARNING

At this point, we can describe our approach to integrating multimodal data for word acquisition [Ballard and Yu 2003]. The system comprises two basic steps: speech segmentation shown in Figure 9 and lexical acquisition illustrated in Figure 11.

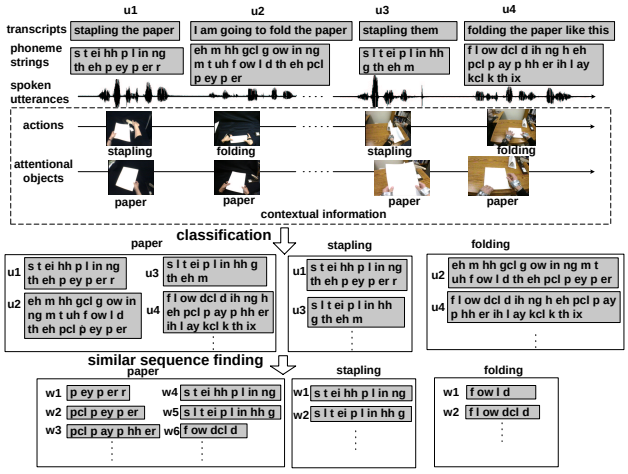


Fig. 9. **Word-like unit segmentation**. Spoken utterances are categorized into several bins that correspond to temporally co-occurring actions and attentional objects. Then we compare any pair of spoken utterances in each bin to find the similar subsequences that are treated as word-like units.

6.1 Word-like Unit Spotting

Figure 9 illustrates our approach to spotting word-like units in which the central idea is to utilize non-speech contextual information to facilitate word spotting. The reason for using the term "word-like units" is that some actions are verbally described by verb phrases (e.g. "line up") but not single action verbs. The inputs shown in Figure 9 are phoneme sequences (u_1, u_2, u_3, u_4) and possible meanings of words (objects and actions) extracted from non-speech perceptual inputs, which are temporally co-occurring with speech. First, those phoneme utterances are categorized into several bins based on their possible associated meanings. For each meaning, we find the corresponding phoneme sequences uttered in temporal proximity, and then categorize them into the same bin labeled by that meaning. For instance, u_1 and u_3 are temporally correlated with the action "stapling", so they are grouped in the same bin labeled by the action "stapling". Note that, since one utterance could be temporally correlated with multiple meanings grounded in different modalities, it is possible that an utterance is selected and classified in different bins. For example, the utterance "stapling a few sheets of paper" is produced when a user performs the action of "stapling" and looks toward the object "paper". In this case, the utterance is put into two bins: one corresponding to the object "paper" and the other labeled by the action "stapling". Next, based on the method described in Subsection 5.2, we compute the similar substrings between any two phoneme sequences in each bin to obtain word-like units. Figure 10 shows an example of extracting word-like units from the utterances u_2 and u_4 that are in the bin of the action "folding".

6.2 Word-like Unit Clustering

Extracted phoneme substrings of word-like units are clustered by a hierarchical agglomerative clustering algorithm that is implemented based on the method described in Subsection 5.2. The centroid of each cluster is then found and adopted as a prototype to represent this cluster. Those prototype strings are mapped back to continuous speech stream

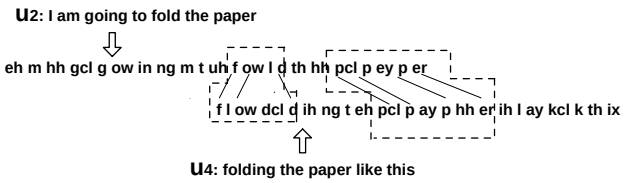


Fig. 10. **An example of word-like unit spotting**. The similar substrings of two sequences are /f ow l d/ (fold), /f l ow dcl d/ (fold), /pcl p ey p er/ (paper) and /pcl p ay p hh er/ (paper).

as shown in Figure 11, which are associated with their possible meanings to build hypothesized lexical items. Among them, some are correct ones, such as /s t ei hh p l in ng/ (stapling) associated the action of "stapling", and some are incorrect, such as /s t ei hh p l in ng/ (stapling) paired with the object "paper". Now that we have hypothesized word-meaning pairs, the next step is to select reliable and correct lexical items.

6.3 Multimodal Integration

In the final 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 looked 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.

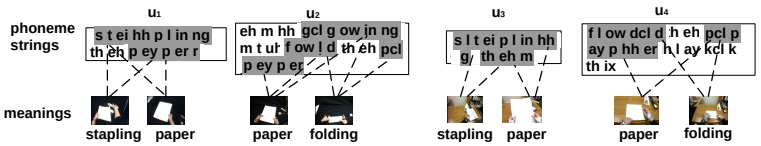


Fig. 11. **Word learning.** The word-like units in each spoken utterance and co-occurring meanings are temporally associated to build possible lexical items.

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 utterance and 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:

$$P(S_m^{(1)}, S_m^{(2)}, ..., S_m^{(S)} | S_w^{(1)}, S_w^{(2)}, ..., S_w^{(S)}) = \prod_{s=1}^{S} P(S_m^{(s)} | S_w^{(s)}) \tag{7}$$

We use the model similar to that of Brown et al. [Brown et al. 1993]. The joint likelihood of meanings and an alignment given spoken utterances:

$$\begin{aligned} P(S_m^{(s)} | S_w^{(s)}) &= \sum_a P(S_m^{(s)}, a | S_w^{(s)}) \\ &= \frac{\epsilon}{(r+1)^l} \sum_{a_1}^{r} \sum_{a_2}^{r} ... \sum_{a_l}^{r} \prod_{j=1}^{l} t(m_{v(j)} | w_{a_{v(j)}}) \\ &= \frac{\epsilon}{(r+1)^l} \prod_{j=1}^{l} \sum_{i=0}^{r} t(m_{v(j)} | w_{u(i)}) \end{aligned} \tag{8}$$

where the alignment $a_{v(j)}, 1 \leq j \leq l$ can taken any value from 0 to r which indicates 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.

We wish to find the association probabilities so as to maximize $P(S_m^{(s)}|S_w^{(s)})$ subject to the constraints that for each word w_n:

$$\sum_{m=1}^{M} t(m_m | w_n) = 1 \tag{9}$$

Therefore, we introduce Lagrange multipliers λ_n and seek an unconstrained maximization:

$$L = \sum_{s=1}^{S} \log P(S_m^{(s)} | S_w^{(s)}) + \sum_{n=1}^{N} \lambda_n (\sum_{m=1}^{M} t(m_m | w_n) - 1) \tag{10}$$

We then compute derivatives of the above objective function with respect to the multipliers λ_n and the unknown parameters $t(m_m|w_n)$ and set them to be zeros. As a result, we can express:

$$\lambda_n = \sum_{m=1}^{M} \sum_{s=1}^{S} c(m_m | w_n, S_m^{(s)}, S_w^{(s)}) \tag{11}$$

$$t(m_m | w_n) = \lambda_n^{-1} \sum_{s=1}^{S} c(m_m | w_n, S_m^{(s)}, S_w^{(s)}) \tag{12}$$

where

$$\begin{aligned} c(m_m | w_n, S_m^{(s)}, S_w^{(s)}) = \frac{t(m_m|w_n)}{t(m_m|w_{u(1)}) + ... + t(m_m|w_{u(r)})} \times \\ \sum_{j=1}^{l} \delta(m, v(j)) \sum_{i=1}^{r} \delta(n, u(i)) \end{aligned} \tag{13}$$

The EM-based algorithm sets an initial $t(m_m|w_n)$ to be flat distribution and performs the E-step and the M-step successively until convergence. In E-step, we compute $c(m_m|w_n, S_m^{(s)}, S_w^{(s)})$ by Equation (13). In M-step, we reestimate both the Lagrange multipliers and the association probabilities using Equation (11) and (12).

When the association probabilities converge, we obtain a set of $t(m_m|w_n)$ and need to select correct lexical items from many possible word-meaning associations. Compared with the training corpus in machine translation, our experimental data is sparse and consequently causes some words to have inappropriately high probabilities to associate the meanings. This is because those words occur very infrequently and are in a few specific contexts. We therefore use two constraints for selection. First, only words that occur more than a pre-defined times are considered. Moreover, for each meaning m_m, the system selects all the words with the probability $t(m_m|w_n)$ greater than a pre-defined threshold. In this way, one meaning can be associated with multiple words. This is because people may use different names to refer to the same object and the spoken form of an action verb can be expressed differently. For instance, the phoneme strings of both "staple" and "stapling" correspond to the action of stapling. In this way, the system is developed to learn all the spoken words that have high probabilities in association with a meaning.

7. EXPERIMENTAL RESULTS

A Polhemus 3D tracker was utilized to acquire 6-DOF hand and head positions at $40Hz$. The performer wore a head-mounted eye tracker from Applied Science Laboratories(ASL). The headband of the ASL held a miniature "scene-camera" to the left of the performer's head that provided the video of the scene from a first-person perspective. The video signals were sampled at the resolution of 320 columns by 240 rows of pixels at the frequency of $15Hz$. The gaze positions on the image plane were reported at the frequency of $60Hz$. Before computing feature vectors for HMMs, all position signals passed through a 6th order Butterworth filter with the cut-off frequency of $5Hz$. The acoustic signals were recorded using a headset microphone at a rate of 16 kHz with 16-bit resolution. In this study, we limited user activities to those on a table. The three activities that users were performing were: "stapling a letter", "pouring water" and "unscrewing a jar". Figure 12 shows snapshots captured from the head-mounted camera when a user performed three tasks. Six users participated in the experiment. They were asked to perform each task nine times while describing what they were doing verbally. We collected multisensory data when they performed the task, which were used as training data for our computational model. Several examples of verbal transcription and detected meanings are showed in Appendix A.

The action sequences in the experiments consist of several motion types: "pick up", "line up", "staple", "fold", "place", "unscrew" and "pour". The objects that are referred to by speech are: "cup", "jar", "waterpot" and "paper". For the evaluation purpose, we manually annotated speech data and calculate the frequencies of words. We have collected 963 spoken utterances and on average, a spoken utterance approximately contains 6 words, which illustrates the necessity of word segmentation from connected speech. Among all these words, most frequently occurring words, such as "and", "are", "going to", are not action verbs and object names that we want to spot and associate with their perceptual grounded meanings. This further demonstrates the difficulty of learning lexical items from naturally co-occurring data. These annotations were only used for the evaluation purpose

and our model did not use them as extra information.

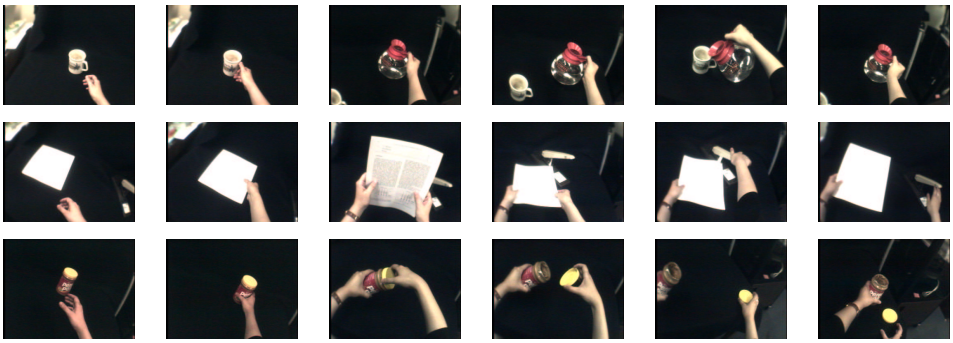


Fig. 12. The snapshots of three continuous action sequences in our experiments. **Top row:** pouring water. **Middle row:** stapling a letter. **Bottom row:** unscrewing a jar.

To evaluate the results of the experiments, we define the following measures on word-like units and grounded lexical items.

—**Semantic accuracy** measures the recognition accuracy of processing non-linguistic information, which consists of clustering the feature sequences of human body movements as well as categorizing visual features of attentional objects.

—**Segmentation accuracy** measures whether the beginning and the end of phonetic strings of word-like units are word boundaries. For example, the string /k ah p/ is a positive instance corresponding to the word "cup" while the string /k ah p i/ is negative. The phrases with correct boundaries are also treated as position instances for two reasons. One is that those phrases do not break word boundaries but only combine some words together. The other reason is that some phrases correspond to concrete grounded meanings, which are exactly spoken units we want to extract. For instance, the phrases, such as "pick up" or "line up", specify some human actions.

—**Word learning accuracy (precision)** measures the percentage of successfully segmented words that are correctly associated with their meanings.

—**Lexical spotting accuracy (recall)** measures the percentage of word-meaning pairs that are spotted by the model. This measure provides a quantitive indication about the percentage of grounded lexical items that can be successfully found.

Table I shows the results of four measures. The recognition rate of the phoneme recognizer we used is 75% because it does not encode any language model and word model. Based on this result, the overall accuracy of speech segmentation is 69.6%. Naturally, an improved phoneme recognizer based on a language model would improve the overall results, but the intent here is to study the model-independent learning method. The error in word learning is mainly caused by a few words (such as "several" and "here") that frequently occur in some contexts but do not have grounded meanings. The overall accuracy of lexical spotting is 83.9%, which demonstrates that by inferring speakers' referential intents, the stable links between words and meanings could be easily spotted and established.

Table I. Results of word acquisition

detected meanings	discovered spoken words /phoneme string/ [text]	semantics	speech segmentation	word learning	lexical spotting
pick up	/p ih kcl k ah p/ [pick up] /p ih kcl k ih ng ah p/ [picking up] /g r ae s pcl p/ [grasp]	96.5%	72.5%	87.5%	72.6%
place	/p l ey s/ [place] / p l ey s ih ng/ [placing] /p uh t/ [put]	93.9%	66.9%	81.2%	69.2%
line up	/l ay n ah pcl p/ [line up] /l ay n ih ng ah pcl p/ [lining up] /l ay n hh m ah pcl p/ [line them up]	75.6%	70.3%	86.6%	83.5%
staple	/s t ey pcl p/ [staple] /s t ey pcl p ih ng/ [stapling]	86.9%	70.6%	85.3%	90.6%
fold	/f ow l d/ [fold] /f ow l d ih ng/ [folding]	86.3%	69.8%	89.2%	87.7%
unscrew	/ ah n s kcl k r uw/ [unscrew] /ow p ah n/ [open]	90.6%	73.8%	91.6%	80.6%
pour	/pcl p ao r/ [pour] /pcl p ao r ih ng/ [pouring]	86.7%	65.3%	91.9%	85.5%
paper	pcl p ey p er/ [paper] /sh iy tcl t/sheet	96.7%	73.9%	86.6%	82.1%
jar	/j aa r/ [jar] /pcl p iy n ah t b ah tcl t er / [peanut butter] /l ih d/ [lid]	91.3%	62.9%	92.1%	76.6%
cup	/k ah p/ [cup]	92.9%	68.3%	87.3%	76.9%
waterpot	/ w ao tcl t er pcl p aa t / [waterpot] /pcl p aa t/ [pot] /kcl k ao f iy pcl p aa t/ [coffee pot]	87.5%	71.9%	85.6%	82.3%
overall		90.2%	69.6%	87.9%	82.6%

Considering that the system processes natural speech and our method works in unsupervised mode without manually encoding labels for multisensory information, the accuracies for both speech segmentation and word learning are impressive.

8. CONCLUSION

This paper presents a multimodal learning interface for word acquisition. The system is able to learn the sound patterns of words and their semantics while users perform everyday tasks and provide spoken descriptions of their behaviors. Compared to previous works, the novelty of our approach arises from the following aspects. First, our system shares user-centric multisensory information with a real agent and grounds semantics directly from egocentric experience without manual transcriptions and human involvement. Second, both words and their perceptually grounded meanings are acquired from sensory inputs. Furthermore, grounded meanings are represented by perceptual features but not abstract symbols, which provides a sensorimotor basis for machines and people to com-

municate with each other through language. From the perspective of machine learning, we propose a new approach of unsupervised learning using multisensory information. Furthermore, we argue that the solely statistical learning of co-occurring data is less likely to explain the whole story of language acquisition. The inference of speaker's referential intentions from their body movements provides constraints to avoid the large amount of irrelevant computations and can be directly applied as deictic reference to associate words with perceptually grounded referents in the physical environment. From an engineering perspective, 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 multisensory information.

9. ACKNOWLEDGMENTS

The authors wish to express their thanks to Richard Aslin, Mary Hayhoe and Robert Jacobs for fruitful discussions. Lucian Galescu provided valuable suggestions about speech perception. Brian Sullivan was a great help in building the experimental system. This material is based upon work supported by the NSF under research grant number EIA-0080124 and by the NIH/PHS under research grant number 2 R01 RR09283.

Appendix A

The tables II III and IV show several examples of transcripts and contextual information extracted from visual perception and body movements. Note that our computational system actually processed continuous speech signals instead of (segmented) transcripts.

Table II. Two sample transcripts of the tasks of stapling papers

transcripts	actions	attentional objects
first, I reach over and pick up some papers	pick up	paper
and I line them up	line up	paper
and I staple them	staple	paper
push the arm down	staple	paper, stapler
now I fold the paper	fold	paper
fold the bottom-up first	fold	paper
then I fold the top over	fold	paper
smooth creases	fold	paper
and I place the paper up here	place	paper
I am picking up several pieces of papers	pick up	paper
I am lining up the paper	line up	paper
now I will staple it	staple	paper, stapler
now I am going to fold the paper	fold	paper
fold the bottom an the top	fold	paper
finally I will place it at its location here	place	paper

Table III. Two sample transcripts of the tasks of pouring water

transcripts	actions	attentional objects
I am going to pick up the cup	pick up	cup
and put it on its spot	place	cup
now I am going to pick up the waterpot	pick up	waterpot
and pour the water into the cup like this	pour	waterpot, cup
after pouring I am going to	none	cup
place the waterpot on its spot	place	waterpot
I will pick up the cup	pick up	cup
place the white cup down on its spot	place	cup
pick up the waterpot and move it toward the cup	pick up	waterpot,cup
pouring myself some water	pour	waterpot, cup
then placing the waterpot into its target area	place	waterpot

Table IV. Two sample transcripts of the tasks of unscrewing a jar

transcripts	actions	attentional objects
I am picking up a peanut butter jar	pick up	jar
now I am unscrewing the lid	unscrew	jar
placing the lid on its spot	place	jar
and placing the jar on its spot which is labeled over there	place	jar
I pick up a jar of peanut butter	pick up	jar
open the peanut butter jar	unscrew	jar
unscrew the lid of the jar	unscrew	jar
place the lid there and	place	jar
place the jar at its location here	place	jar

REFERENCES

ADAMS, R. AND BISCHOF, L. 1994. Seeded region growing. *IEEE Transaction on Pattern Analysis and Machine Intelligence 16*, 6 (June), 641–647.

BAILEY, D. 1997. When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs. Ph.D. thesis, Computer Science Division,University of California Berkeley.

BALDWIN, D. A., MARKMAN, E. M., BILL, B., DESJARDINS, R. N., IRWIN, J. M., AND TIDBALL, G. 1996. Infant's reliance on a social criterion for establishing word-object relations. *Child development 67*, 3135–3153.

BALLARD, D. H., HAYHOE, M. M., POOK, P. K., AND RAO, R. P. N. 1997. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences 20*, 1311–1328.

BALLARD, D. H. AND YU, C. 2003. A multimodal learning interface for word acquisition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Hong Kong.

BARON-COHEN, S. 1995. *Mindblindness: an essay on autism and theory of mind*. MIT Press, Cambridge.

BATES, E. AND GOODMAN, J. C. 1999. On the emergence of grammar from the lexicon. In *The Emergence of Language*, B. M. Whinney, Ed. Lawrence Erlbaum Associates.

BLOOM, P. 2000. *How children learn the meanings of words*. The MIT Press, Cambridge, MA.

BRENT, M. R. 1999. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive science 3*, 8, 294–301.

BRENT, M. R. AND CARTWRIGHT, T. A. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition 61*, 93–125.

BROWN, P. F., PIETRA, S., PIETRA, V., AND MERCER, R. L. 1993. The mathematics of statistical machine translation:parameter estimation. *Computational Linguistics 19*, 2.

CAMPBELL, L., BECKER, D., AZARBAYEJANI, A., BOBICK, A., AND PENTLAND, A. 1996. Invariant features for 3-d gesture recognition. In *Second International Workshop on Face and Gesture Recognition*. Killington, VT, 157–162.

GARDENFORS, P. 1999. Some tenets of cognitive semantics. John Benjamins.

GILLETTE, J., GLEITMAN, H., GLEITMAN, L., AND LEDERER, A. 1999. Human simulations of vocabulary learning. *Cognition 73*, 135–176.

GOPNIK AND MELTZOOF. 1997. *Words, Thought and Theories*. MIT.

HARNAD, S. 1990. The symbol grounding problem. *physica D 42*, 335–346.

HARTIGAN, J. 1975. *Clustering Algorithms*. Wiley, New York.

HAYHOE, M. 2000. Visual routines: A functional account of vision. *Visual Cognition 7*, 43–64.

JUSCZYK, P. W. 1997. *The discovery of spoken language*. MIT Press, Cambridge, MA.

JUSCZYK, P. W. AND ASLIN, R. N. 1995. Infants detection of the sound patterns of words in fluent speech. *Cognitive Psychology*.

KEIL. 1989. *concepts, kinds, and cognitive development*. MIT.

KUNIYOSHI, Y. AND INOUE, H. 1993. Qualitative recognition of ongoing human action sequences. In *Proc. IJCAI93*. Chambery, France, 1600–1609.

LADEFOGED, P. 1993. *A Course in Phonetics*. Harcourt Brace Jovanovich, Orlando, FL.

LAKOFF, G. AND JOHNSON, M. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

LAND, M., MENNIE, N., AND RUSTED, J. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception 28*, 1311–1328.

LAND, M. F. AND HAYHOE, M. 2001. In what ways do eye movements contribute to everyday activities? *Vision Research 41*, 3559–3565.

LANDAU, B., SMITH, L., AND JONES, S. 1998. Object perception and object naming in early development. *Trends in cognitive science*.

LIPPMANN, R. P. 1989. Review of neural networks for speech recognition. *Neural computation 1*, 1, 1–38.

MEL, B. W. 1997. Seemore: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation 9*, 777–804.

MEYER, A. S., SLEIDERINK, A. M., AND LEVELT, W. J. 1998. Viewing and naming objects: eye movements during noun phrase production. *Cognition 66*, B25–B33.

OATES, T., FIROIU, L., AND COHEN, P. R. 1999. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*. Stockholm, Sweden, 17–21.

OVIATT, S. 2002. Multimodal interfaces. In *Handbook of Human-Computer Interaction*, J. Jacko and A. Sears, Eds. Lawrence Erlbaum, New Jersey.

QUINN, P., EIMAS, P., AND ROSENKRANTZ, S. 1993. Evidence for representations of perceptually similar natural categories by 3-month old and 4-month old infants. *Perception 22*, 463–375.

RABINER, L. R. AND JUANG, B. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2, 257–286.

REGIER, T. 1996. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press, MA.

REGIER, T. 2003. Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences 7*, 263–268.

ROBINSON, T. 1994. An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks 5*, 2, 298–305.

ROY, D. 2002. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication 4*, 1.

ROY, D. AND PENTLAND, A. 2002. Learning words from sights and sounds: A computational model. *Cognitive Science 26,* 1, 113–146.

SALVUCCI, D. D. AND ANDERSON, J. 1998. Tracking eye movement protocols with cognitive process models. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. LEA: Mahwah, NJ, 923–928.

SCHIELE, B. AND CROWLEY, J. L. 2000. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision 36,* 1, 31–50.

SCHYNS, P. AND RODET, L. 1997. Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition 23,* 681–696.

SISKIND, J. M. 1995. Grounding language in perception. *artificial Intelligence Review 8,* 371–391.

SISKIND, J. M. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research 15,* 31=90.

SMITH, L., JONES, S., AND LANDAU, B. 1996. Naming in young children: A dumb attentional mechanism? *Congition 60,* 2.

SMYTH, P. 1997. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing Systems,* M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. The MIT Press, 648.

STEELS, L. AND VOGT, P. 1997. Grounding adaptive language game in robotic agents. In *Proc. of the 4th European Conference on Artificial Life,* C. Husbands and I. Harvey, Eds. MIT Press, London.

SWAIN, M. J. AND BALLARD, D. 1991. Color indexing. *International Journal of Computer Vision 7,* 11–32.

TOMASELLO, M. 2000. Perceiving intentions and learning words in the second year of life. In *Language Acquisition and Conceptual Development,* M. Bowerman and S. Levinson, Eds. Cambridge University Press.

WANG, S. AND SISKIND, J. M. 2003. Image segmentation with ratio cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25,* 6, 675–690,.

WENG, J., MCCLELLAND, J., PENTLAND, A., SPORNS, O., STOCKMAN, I., SUR, M., AND THELEN, E. 2001. Artificial intelligence: Autonomous mental development by robots and animals. *Science 291,* 5504, 599–600.

YU, C. AND BALLARD, D. H. 2002a. Learning to recognize human action sequences. In *IEEE Proceedings of the 2nd International Conference on Development and Learning*. Boston, MA, 28–34.

YU, C. AND BALLARD, D. H. 2002b. Understanding human bahaviors based on eye-head-hand coordination. In *2nd Workshop on Biologically Motivated Computer Vision*. Tubingen, Germany.

YU, C. AND BALLARD, D. H. 2003. Exploring the role of attention in modeling embodied language acquisition. In *Proceedings of the Fifth International Conference on Cognitive Modeling*. Bamburg, Germany.

YU, C., BALLARD, D. H., AND ASLIN, R. N. 2003. The role of embodied intention in early lexical acquisition. In *Proceedings the Twenty Fifth Cognitive Science Society Annual Meetings*. Boston, MA.

YU, C., BALLARD, D. H., AND ZHU, S. 2002. Attentional object spotting by integrating multisensory input. In *Proceedings of the 4th International Conference on Multimodal Interface*. Pittsburg, PA.