Visual Context Driven Semantic Priming of Speech Recognition and Understanding

Deb Roy and Niloy Mukherjee

***Abstract*— Fuse is a spoken language understanding system that integrates visual context into early stages of speech recognition. Given a visual scene and a spoken description, the system finds the object in the scene that best fits the meaning of the description. To solve this task, Fuse performs speech recognition and visually-grounded language understanding. Rather than treat these two problems separately, knowledge of the visual semantics of language and the specific contents of the visual scene are fused into the speech recognition process. The system effectively anticipates various ways a person might describe any object in the scene, and uses these predictions to bias the speech recognizer towards likely sequences of words. A dynamic model of visual attention is used to focus processing on likely objects within the scene as spoken utterances are processed. Visual attention and language prediction reinforce another and converge on interpretations of incoming speech signals which are most consistent with visual context. In evaluations, the introduction of visual context into the speech recognition process results in significantly improved speech recognition and understanding accuracy. The underlying principles of this model may be applied to a wide range of speech understanding problems including mobile and assistive technologies in which contextual information can be sensed by the system and semantically interpreted to bias processing.**

I. Introduction

Modularity is a central principal in the design of complex systems, and is often postulated in theories of human cognition [1], [2]. Modules operate as encapsulated "black boxes" that can only access other modules through well-defined interfaces. Access to internal data structures and processing within modules is privileged. Studies of human behavior, however, sometimes reveal surprising breaches of modularity. For example, recent psycholinguistic experiments have shown that acoustic and syntactic aspects of online spoken language comprehension are influenced by visual context. During interpretation of speech, partially heard utterances have been shown to incrementally steer the hearer's visual attention [3], and vice versa, visual context has been shown to influence speech processing [4], [5]. Motivated by these findings, we have developed a spoken language understanding system in which visual context primes early stages of speech processing, resulting in significantly improved speech recognition and understanding accuracy.

The development of robots provides an exemplery problem that lends itself to modular design. In practically all robots, the perceptual, planning, motor control, and speech systems (if any) operate independently and are integrated through relatively high level interfaces. In this paper, we consider the design of a speech understanding system that will eventually provide speech processing capabilities for an interactive conversational robot [6]. A straight forward approach is to take an off-the-shelf speech recognition system and connect its output to other modules of the robot. We argue, however, that by treating the speech recognizer as a black box that is unaware of the contents of other modules, valuable information is lost. Since high accuracy speech recognition in natural conditions remains unattainable, leveraging information from other channels can be of immense value.

We will consider the problem of understanding spoken utterances that make reference to objects in a scene. When an utterance is known to refer to an object in the immediate environment, the hearer can use knowledge of the shared environment to anticipate words and phrases that the speaker is likely to choose. A difficulty in this approach is that there are typically numerous potential referents in most environments. The hearer does not know, a priori, which referent the speaker intends to describe (otherwise there would be no need to listen to the speech!). Our approach is to jointly infer the most likely words in the utterance along with the identity of the intended referent.

This approach has been implemented in an on-line, real-time, multimodal processing system. Visual scene analysis reaches into the core of the speech recognition search algorithm and steers search paths towards more likely word sequences. The semantic content of partially decoded spoken utterances, in complement, feed back to the visual system and drive a dynamic model of visual attention. As processing proceeds, linguistic and visual information mutually reinforce each other, sharpening both linguistic and visual hypotheses as sensory evidence accumulates. We show that in contrast to modular approaches to integration, early integration leads to substantial improvement in speech recognition accuracy. We believe that the strategic introduction of cross-module bridges may be an important design principal in a wide range of applications beyond the specific system presented.

After providing some background remarks, Section introduces the task we used for our current work on contextualized speech understanding. This section provides a self-contained overview of our approach to integration of visual context into speech recognition. Subsequent chapters provide details on aspects of this approach, followed by experimental evaluations.

II. Background

Integration of spoken and visual input has been investigated in a wide range of tasks. It is useful to distinguish two broad classes of tasks. Let S and V denote the speech and visual

input signals, respectively. The speech signal's primary role is to encode sequences of words. Prosodic aspects of speech also encode affective, syntactic, and stress information. All information in S convey the speaker's intent. In contrast, V may carry two distinct kinds of information, depending on the task. Consider first the problem of audiovisual lipreading. In this task, visual input typically consists of images of the speaker's lips as they speak. In this case, the basic kind of information carried in V is the same as S: words. The visual channel provides complimentary or redundant aspects of the surface form of words. This complementarity of encodings of word surface forms can be leveraged to increase recognition accuracy. For lipreading, we can say that $V = V_i$, where i reminds us that the purpose of the visual channel is to *indicate*. The lips are part of the speakers way of conveying his/her intention. A related problem that has received significant attention is the integration of speech with visually observed gestures made either by hand or using a mouse. Although hand gestures are very different in nature from the motion of lips, broadly speaking, both belong to the same class of $V = V_i$ since gestures also play the role of indicating the speaker's intentions.

In contrast, consider the problem of building a speech understanding system for robot in which the visual input comes from a camera mounted in the robot, looking out into the robot's environment. The speaker asks the robot to pick up a red block. The visual channel might capture the speaker, complete with lip movements and other body gestures. However, the visual signal will also contain information about the robot's *context*, which in this case may include a red block. We indicate this kind of visual information by saying $V = V_i + V_c$ where V_c denotes contextual information captured in the visual signal. If the speaker is not in view, then $V = V_c$. The contents of V_c are fundamentally different from V_i since S may be *about* aspects of V_c but not V_i[1].

The focus of this paper is for a task in which $V = V_c$, i.e, the visual input contains purely contextual information. In contrast to lipreading and gesture understanding problems, we will instead investigate the semantic referential content of the visual signal and how it can be integrated with S in useful ways for a real-time multimodal understanding system.

Most previous work on integrating visual context with speech / language understanding involves modular, late integration. SAM (speech activated manipulator) [7] is a robotic system with sensory capabilities that interacts with a human conversation partner through spoken language dialog. Speech recognition and visual analysis are integrated at a relatively late stage through an augmented transition network that operates on a frame-based knowledge representation. Crangle and Suppes [8] have proposed an approach to verbal interaction with an instructable robot based on a unification grammar formalism. They have examined the use of explicit verbal instructions to teach robots new procedures and have also studied ways a robot could learn from corrective user commands containing qualitative spatial expressions. Although speech may provide linguistic input to their framework, there is no mechanism for propagating semantic information to early speech processing due to the modular design of their model. Wachsmuth and Sagerer (2002) presents a probabilistic decoding scheme that takes the speech signal and an image or image sequence as input. The speech signal is decoded independent of the decoding of the image data. A Bayesian network integrates speech and image representations to generate a representation of the speaker's intention. In summary, each of these systems integrates spoken language with visual context, but the conversion of speech to text occurs in a contextual vacuum. In contrast, we have explored one way to push context into speech recognition.

In our own previous work [9], we have developed a trainable spoken language understanding system that selects individual objects on a table top based on referring spoken language expressions. The system uses speech recognition output and image representations generated by a visual analysis module to robustly parse the speech in real time and points to an object that best fits the description. In contrast to these previous works, the approach presented here demonstrates integration of visual context into core of the speech recognition search process, a process that is treated as modular and isolated from visual influence in previous work. To make the early link between visual context and speech, we rely on visually-grounded models of word semantics that we developed in a previous system [10]. For other approaches to visually-grounding word meaning, see also [11]–[16].

III. Overview

To study the role of visual context in spoken language comprehension, we developed a simple scene description task. Participants in a data collection study were asked to verbally describe objects in scenes consisting of oversized lego blocks (Figure 1). No restrictions were placed on the vocabulary, style, or length of description. Typical descriptions ranged from simple phrases such as, "The green one in front" to more complex referential utterances such as, "The large green block beneath the smaller red and yellow ones". The result of this data collection was a set of images and spoken descriptions to objects in the images. We addressed the problem of visually-grounded speech understanding: given a spoken description, find the object in the scene that best fits the description.

In a modular approach, speech recognition and visual analysis would be performed separately and combined by an integrator that does not affect the internal operations of earlier stages of processing. In previous work, we have followed this modular approach [9], as have others (cf. [7], [17]).

In this section, we provide an overview of an alternate approach in which the core speech recognition process is altered by knowledge of visual context.

A. The Role of Language Modeling in Speech Recognition

Speech recognition is most commonly formulated in a maximum likelihood framework [18]. Given an observed spoken utterance, X, we wish to choose a word string $\widehat{W}$

[1]One can imagine rare exceptions to this. A person, while waving their arm in a strange gesture, might say, "It hurts when I do *this*", where, "this" refers to the gesture. This is harder to do with lips, and for our current purposes, we set these exceptions aside.

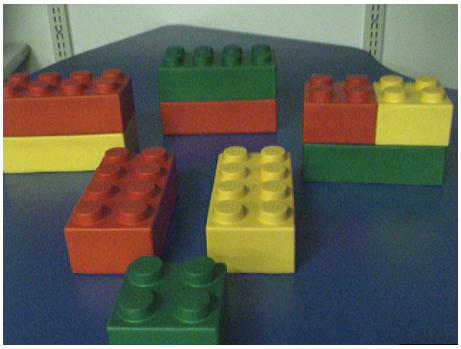

Fig. 1. A typical visual scene in the current experimental task.

$$\widehat{W} = \underset{W}{\text{argmax}} \; P(X|W)P(W) \tag{1}$$

The terms $P(X|W)$ and $P(W)$ correspond to an acoustic model and language model, respectively. In conventional speech recognition systems, the acoustic model captures the acoustic properties of speech and provides the probability of speech observation given hypothesized word sequences. In audio-visual speech recognition systems, speech observations include both acoustic and visual information. The acoustic model that provides $P(X|W)$ is generalized to also model visual information. Referring to our discussion in Section II, this is an instance of using V_i, visual information that conveys the speaker's intent.

The language model, $P(W)$, provides probabilities of word strings W based on *context*. In practically all speech recognition systems, this context is a function of the history of words that the speaker has uttered. In contrast, our approach to visual integration is to dynamically modify $P(W)$ on the basis of visual context (V_c). By doing so, the search process which is central to speech recognition is influenced by visual context. This cross-modal coupling provides an example of early cross-modal integration. In contrast to late integration, early integration effectively reaches into the speech decoder and effects how the decoder interprets acoustic speech signals.

Since our focus will be on dynamic language models, we provide a brief overview of the most widely used statistical language model, known as the *n-gram* which will serve as a basis for our cross-modal extension. The n-gram model assigns probabilities to hypothesized word sequences. The model implicitly captures syntactic, semantic, and contextual knowledge. The probability of a word sequence $W = w_1, w_2, \ldots, w_k$ which we denote as w_1^k, can be expressed as a product of conditional probabilities:

$$P(w_1^k) = P(w_1)P(w_2|w_1)\cdots P(w_k|w_1^{k-1}) \tag{2}$$

In the $P(w_k|w_1^{k-1})$, w_1^{k-1} is called the history and w_k the prediction. In the n-gram approach, two histories are treated as identical when they end in the same $n-1$ words. For example, with $n = 2$, we obtain a bigram language model:

$$P(w_1^k) = P(w_1)P(w_2|w_1)\cdots P(w_k|w_{k-1}) \tag{3}$$

Many extensions to basic n-gram language models have been proposed such as variable length histories [19], long distance dependencies [20] (for a review, see [21]). Stochastic context-free grammars provide an alternative to n-grams that does not make Markov assumptions [22]. Our goal is to introduce a form of visually-driven semantic priming into the statistical language model of a real-time speech recognizer. In principal, any of the n-gram extensions mentioned above can be augmented with visual context in the way that we propose. For simplicity, we have chosen to work with the bigram language model which has sufficient modeling power for the present scene description task.

The parameters of a bigram model are usually estimated from a large text corpus. Given a training corpus of size T words in which word w occurs $|w|$ times, the maximum likelihood estimate of $P(w)$ is $|w|/T$. The maximum likelihood estimates for the conditional terms $P(w_i|w_{i-1})$ are given by $|w_{i-1}, w_i|/|w_i|$ where $|w_{i-1}, w_i|$ is the number of times the sequence w_{i-1}, w_i occurs in the training corpus.

Words may be clustered into equivalence classes leading to n-gram class models [23]. For example, if the distribution of words in the neighborhood of *Monday* and *Tuesday* are believed to be similar, the words can be clustered, and treated as equivalent for language modeling. The principal benefit of creating word classes is that we are able to make better use of limited training data to make predictions for word histories that are not encountered in training. We can partition a vocabulary into word classes using a function which maps each word w_i to its corresponding class $c(w_i)$. For bigram class models,

$$P(w_i|w_{i-1}) = P(w_i|c(w_i))P(c_i|c_{i-1}) \tag{4}$$

Standard word bigrams are a special case of bigram class models in which each word is mapped into its own unique word class.

B. Visual-Context Sensitive Language Models

Figure 2 provides an overview of our approach to integrating visual context with speech recognition and understanding in a model called Fuse. The remainder of this section sketches the main ideas underlying the approach. Following sections provide details of implementation and evaluation.

As shown in Figure 2, input to Fuse consists of a speech signal and an image. Figure 1 is representative of images in the current task, captured by a color video camera. The speech signal is recorded from a head-worn microphone. The spoken utterances used for evaluations consisted of naturally spoken, fluent speech.

The visual scene analysis module detects objects in the scene and extracts a set of visual features that represent individual objects, and inter-object spatial relations. The results of the scene analysis are accessible by two modules: a language model, and a visual attention model. As the speech signal is processed, both the language and attention models are dynamically updated. Working together, the models steer the interpretation of the speech signal based on visual context.

To understand the main processing loop in Figure 2 and the role of the language model and visual attention model, we will

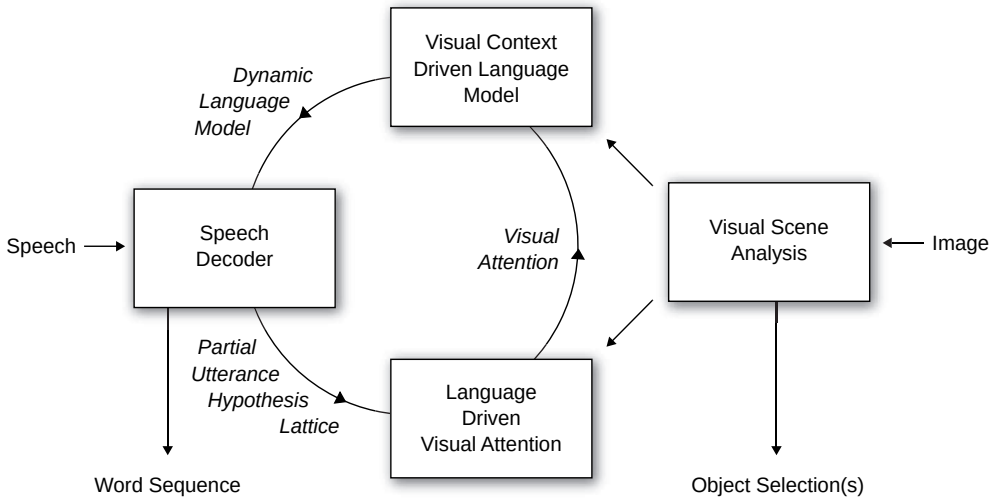

Fig. 2. Overview of the Fuse architecture.

work through a simple example. Let us consider a situation in which a speaker says, “The red block on the left” in the context of a scene containing four blocks: a red one and a blue one on the left, and a red one and blue one on the right. For a moment, let us ignore the dynamic language model connected to the speech recognizer, and instead assume a standard static bigram language model. As the first portion of the input utterance is processed, let us further assume that the speech recognizer correctly recovers the first two words of the utterance, “the red”. In actuality, the output of the speech recognizer will be a lattice that encodes multiple word hypotheses, but to keep the example simple, we only consider a single word sequence.

The partially decoded word sequence is fed to the visual attention module which also receives the output of the visual scene analyzer. Visual attention is modeled as a probability mass function (pmf) over the set of objects in the scene. Initially, before speech recognition begins, the pmf is non-informative and associates equal probability to all objects in the scene. When the words “the red” are fed into the visual attention module, the pmf is updated so that most of the probability mass is shifted to the red objects in the scene. In effect, the visual attention of the system shifts to the red objects. The attention module uses a set of visually-grounded semantic models to convert the word sequence into the pmf (Section VI).

The visual attention pmf, which now favors the two red objects in the scene, is transmitted to the language model. The language model may be thought of as a linguistic description generator. For each object in the scene, the model generates a set of referring expressions that a person might use to describe the object. For the red block on the left, the model might generate a set of descriptions including “the red block”, “the large red block”, the “the red block on the left”, and so forth. Each description is assigned a likelihood that depends on how well the description matches the visual attributes of the object, but also based on syntactic and contextual factors. The likelihoods of the descriptions for each object are multiplied by the probability assigned to that object by the visual attention pmf. The resulting mixture of descriptions is converted into a statistical language model which is used by the speech recognizer. In effect, visual attention steers the speech recognizer to interpret the input speech signal as a description of objects that have captured more of the system’s attention.

To summarize, as acoustic evidence is incrementally processed, the visual attention pmf evolves. The dynamic pmf in turn biases the language model of the speech recognizer. As more of the utterance is processed, the visual attention becomes progressively sharpened towards potential referents in the scene.

Several details have been simplified in this overview. One complication is introduced with utterances with relative spatial clauses such as, “The red block to the left of the large blue one”. In this class of utterances, the visual attention must be reset mid-way through processing to refocused from one object to another. Another complication arises from the fact that the output of the speech recognizer at any moment is not a single word sequence, but rather a lattice that encodes multiple (potentially thousands) of alternative word hypotheses. These and other aspects of Fuse are explained in the following sections which provide detailed descriptions of each component of the system.

IV. Visual Scene Analysis

The visual scene analysis module segments objects in an input scene and computes visual properties of individual objects, and spatial relations between pairs of objects. The resulting representation of the scene is used by both the language model and visual attention model.

Objects are segmented based on color. A simple statistical color model is created objects by training Gaussian mixture models on sample images of the objects. We assume that objects will be single-colored, greatly simplifying the segmentation process. The Expectation Maximization (EM) algorithm is used to estimate both the mixture weights and the underlying Gaussian parameters for each color model. The color models

are used as a Bayes classifier to label each 5x5 pixel region of an input region. Regions of the image that do not match any object color model are classified as background using a fixed threshold. Objects are found by extracting connected foreground regions of consistent color.

A set of visual properties are computed for each object found in the segmentation step, and for spatial relations between each pair of objects. These properties and relations constitute the complete representation of a visual scene. The features attempt to capture aspects of the scene that are likely to be referred to in natural spoken descriptions. The following visual features are extracted:

- **Color** is represented by the mean RGB value of the 10x10 pixel region in the center of the object.
- **Shape** is represented by five geometric features computed on the bounding box of each object: height, width, height-to-width ratio, ratio of the larger to the smaller dimension (height / width), and bounding box area [10].
- **Position** is represented by the horizontal and vertical position of the of center of the region.
- **Spatial relations** are encoded by a set of three spatial features suggested in [12] that are measured between pairs of objects. The first feature is the angle (relative to the horizon) of the line connecting the centers of area of an object pair. The second feature is the shortest distance between the edges of the objects. The third feature measures the angle (relative to the horizon) of the line which connects the two most proximal points of the objects.

To summarize, each object is represented by a ten-dimensional feature vector (3 color features, 5 shape, and 2 position). The spatial relation between each pair of objects is represented by 3 additional spatial features. In real time operation, the visual analysis system captures and processes video frames at a rate of 15Hz. When Fuse detects the onset of a spoken utterances, the visual frame co-occurring with the start of the utterance is captured, and the results visual features are used to provide context for processing of the entire spoken utterance (changes made to the scene once the utterance has begun are ignored).

V. Speech Decoding

The role of the speech decoder is to find word sequences that best explain acoustic input. Since the main contribution of the Fuse architecture lies in the treatment of language modeling, this section briefly summarizes the speech decoder. The decoding strategy and algorithms are all based on previously published work. The decoder has been tested on standard speech recognition test corpora and performs competitively with other research platforms, and thus serves as a useful baseline for the experiments presented here.

Speech is represented using a 24-band Mel-scaled cepstral acoustic representation [24]. Words are modeled by concatenating context sensitive phoneme (triphone) models based on continuous-density three-state, Hidden Markov Models [25]. Speech decoding is accomplished using a time-synchronous Viterbi beam search [25].

VI. Visual Context Driven Language Model

The language model is designed to "second guess" what the speaker is likely to say, assuming he/she will speak a description of an object in the current visual scene. If the language model is able to accurately anticipate the speaker's words, the model can bias the speech decoder towards more likely interpretations of the incoming speech signal. There are several sources of uncertainty in predicting how a person will describe objects in the scene:

1) The identity of the target item is unknown, so the language model must consider descriptions that fit all objects in the scene.
2) People may use different words to refer to the same attributes. For example, one person might call an object blue, while another speaker will call it purple.
3) Speakers may use different combinations of words to refer to the same object; "the blue one", the "the tall block", and "the cube to the left of the red one" may all refer to the same referent.

To address the first sources of uncertainty, descriptions are generated, in turn, for each object in the current scene. For each object, multiple descriptions are generated to account for variations due to factors (2) and (3). The potentially large set of resulting descriptions are then weighted and combined to create an n-gram language model that is used by the speech decoder. Although the descriptions stay fixed during the processing of an utterance, the relative weighting of individual descriptions is dynamically updated using the visual attention model that is described in Section VII. As a result, the n-gram language model is not only influenced by visual context as recorded at the onset of the utterance, but further evolves online as the utterance is processed.

The method for generating descriptions is adapted from the trainable object description system described in [10]. In this previous work, we developed learning algorithms that take as input synthetic visual scenes paired with natural language descriptions of objects. The trained description system consists of a set of visually-grounded word models that are grouped into word classes. A two-layer stochastic finite state machine (SFSM) organizes the word classes into a structured graph. Any path through the SFSM generates a sequence of word classes. By selecting a single word from each word class (using the visual models linked to each word), the word class sequence is converted into a word sequence that constitutes a description of an object. Although the original system was designed to operate with synthetic images, transfer of the algorithms to the current constrained visual task was straightforward since similar features are extracted from the camera-based images as were from the synthetic images.

A. The Description Model

At the heart of the description model is a set of *visually-grounded word models*. Each word model consists of a phonemic transcription paired with a statistical model that represents the visual semantics associated with the word. The phonemic transcriptions are used by the speech decoder for acoustic matching with the speech signal. A visual model consists of a

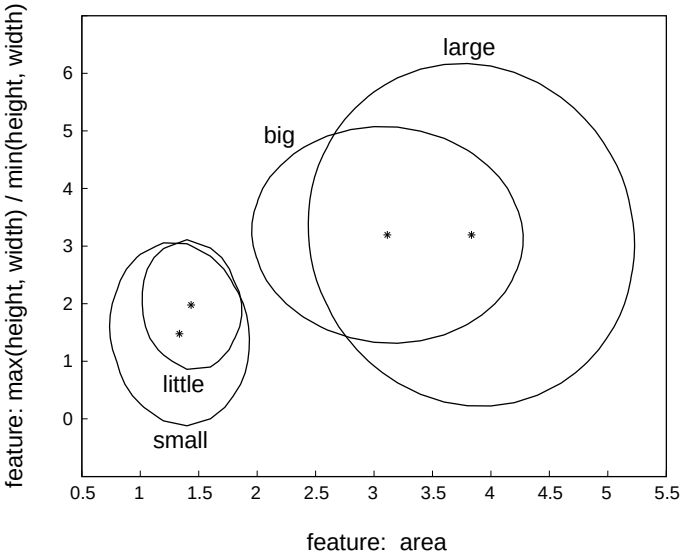

Fig. 3. Example of a word class with four members. Each ellipse indicates an equal probability density contour associated with a visual model (a full-covariance Gaussian distribution), centered on its mean which is indicated with a small asterisk. An automatic feature selection algorithm determined the two visual features used for defining this set of four words.

multidimensional Gaussian distribution defined over a subset of the 10 visual features described in Section refsec:vision. Grounded words are clustered into word classes based on semantic and syntactic relatedness. A two-layer stochastic finite state machine (SFSM) models word order.

All parameters of the description model are learned from examples of objects embedded in scenes that are labeled with descriptive phrases. A set of 60 training examples were collected from eight participants, resulting in a total of 480 examples in the training dataset.

Learning algorithms that we have previously developed [10] were used to train all components of the model. Since the training methods have been previously described, we summarize the contents of the trained model and how the model is used to generate descriptions.

Figure 3 shows the visual models associated with the members of a word class in Fuse. The figure shows that two geometric features (area, and ratio of dimensions) have been selected as the defining visual attributes for this cluster of words. The overlapping distributions show the relation between the words *big* and *large*, and the near-antonyms *little* and *small*. As we shall see, word classes and their associated visual models are used as Bayes classifiers in order to generate labels for novel objects.

Word order is modeled through bigrams that specify transition probabilities between words and word classes. Figure 4 shows a subset of phrase level bigrams in the form of a transition network. Each arc is labeled with the transition probability between the connected words / word classes. Any path through this network constitutes a possible description of an object. For instance, *the red block* and *the leftmost large one* are word sequences that may be generated by this network. A higher-order phrase network (Figure 5) models relative spatial phrases. The phrase nodes in this network each embed a copy of the phrase network and are connected by relative spatial terms. This phrase network can generate sequences such as *the large green block beneath the red one.*

B. Mixtures of Descriptions for Language Modeling

In our implementation, the speech recognizer requires a language model consisting of a set of word bigram transition probabilities. As Equation 4 shows, the word bigram can be obtained from the product of word class transition probabilities $P(c_i|c_{i-1})$ and class conditional word probabilities $P(w_i|c_i)$. The word class transition probabilities are fully determined from training data (Figure reffig:fsm1) and remain static during speech processing. Thus, the expected order of word classes, and transition probabilities between classes is not expected to change as a function of visual context. The probabilities of words *within* each word class, on the other hand, do depend on context. As a simple example, if there are no blue objects in the scene, the probability for the word blue should be reduced relative to other words in its class. To capture this intuition, class conditional word probabilities are dynamically estimated as a function of the scene and visual attention using a six step process:

1) *Enumerate all left-to-right paths through grammar*
   All distinct paths connecting the $start$ and end nodes of the transition network are enumerated. Loops are avoided, resulting in only left-to-right paths. This process leads to a set of N sequences, $\{C_1, C_2, \ldots, C_N\}$. Each sequence C_i consist of a ordered set of T_i word classes:

$$C_i = c_i^1, c_i^2, \ldots, c_i^{T_i} \tag{5}$$

   These sequences constitute the set of syntactic frames embedded in the transition network.
2) *Map word classes to words*
   Each grounded word class is visually grounded in a set of visual models, one model associated with each word in the class. These models can be treated as a standard Bayes classifier [26] to classify objects based on their measured visual attributes. For example, consider the word class shown in Figure 3. To use this word class as a Bayes classifier to label an object, the two features of the object associated with visual models must be measured. Each of the visual models are then evaluated at the measured values, and the model with the highest value (probability density) is selected as the best match to the object. The word associated with that model is thus the best choice within the word class for describing the object. The mapping from word class to word is object dependent; different words may become most activated within a class depending on the visual properties of the object. We denote the word sequence generated by using the word class sequence C_i to describe object O_j as:

$$W_i^j = w_{ij}^1, w_{ij}^2, \ldots, w_{ij}^{T_i} \tag{6}$$

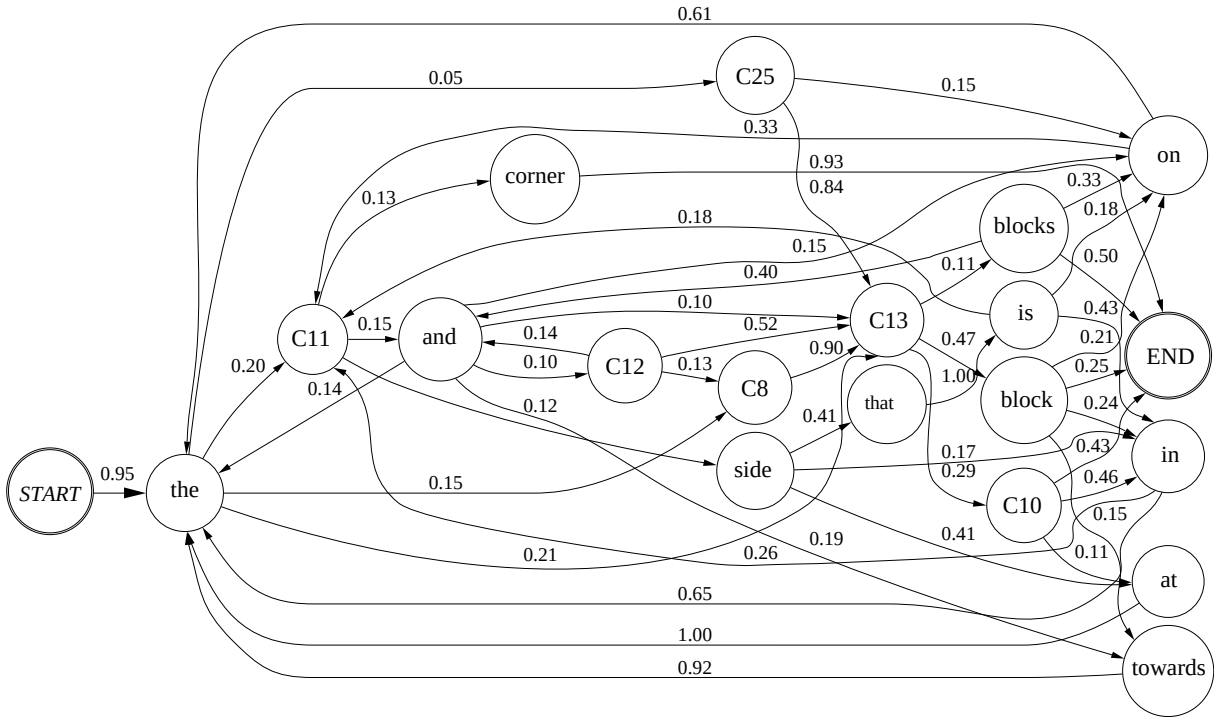

Fig. 4. The probabilistic grammar used to generate descriptions of objects. Nodes include individual ungrounded words and grounded word classes. To allow legibility, the full grammar used in experiments has been pruned for the figure (18 of 55 nodes are shown).

For a scene with M objects, this mapping process results in $N \times M$ word sequences (N descriptions for each of M objects).

3) *For each description, compute its descriptive fitness*
Each description can be evaluated for how well it visually matches its target object by computing the product of the word conditional probabilities of the observed object properties, which is equivalently expressed as a sum of log probabilities:

$$fit(W_i^j, O_j) = \frac{\Sigma_{t=1}^{T_i} \log p(O_j|w_{ij}^t)}{G(C_i)} \quad (7)$$

Where $G(C_i)$ is the number of visually grounded word classes in the sequence C_i, and $p(O_j|w_{ij}^t))$ evaluated the visual model associated with word w_{ij}^t for the visual features of object O_j . For ungrounded words, $p(O_j|w_{ij}^t))$ is set to 1.0. The denominator term normalizes effects due to the length of the description.
The fitness function measures how well a descriptive phrase matches the properties of the target object, but it does not account for contextual effects due to other objects in the scene. For example, a description that matches the target well may also describe a non-target equally well. To capture contextual effects, we define a context-sensitive fitness, which is assigned to the source word class sequence:

$$\psi(C_i, O_j) = fit(W_i^j, O_j) - max_{k \neq j} fit(W_i^j, O_k) \quad (8)$$

4) *Compute object-conditional word predictions*
For a given object and word class sequence, object conditional probabilities are assigned to each visually grounded word:

$$P(w|O_i, c(w)) = \frac{p(O_i|w) \sum_{C_j, w \epsilon C_j} \psi(C_j, O_i)}{\sum_{k=1}^{M} p(O_k|w) \sum_{C_j, w \epsilon C_j} \psi(C_j, O_k)} \quad (9)$$

Where $c(w)$ is the word class to which w belongs. The context-sensitive fitness scores $\psi(C_j, O_i)$ scale each visually based probability density $p(O_i|w)$ depending on how well the overall syntactic frame C_j is able to generate an unambiguous description of O_i.

5) *Mix word predictions using visual attention*
The final step is to mix the influences of all objects in the scene:

$$P(w|c(w)) = \sum_{i=1}^{M} P(w|O_i, c(w)) P(O_i) \quad (10)$$

Relative emphasis of objects is controlled by Fuse's visual attention state, $P(O_i)$, described in the next section.

Using these five steps, a set of class conditional word probabilities are generated that represent the system's anticipation of words the speaker will use, given the contents of the visual scene, and the system's current visual attention state. Referring back to Equation 4, we can see that the dynamic formulation of class conditional probability estimates $P(w|c(w))$ in Equation 10 can be directly inserted into the computation of bigrams that feed into the speech recognizer. As certain objects in the scene capture more of Fuse's attention, the words that better

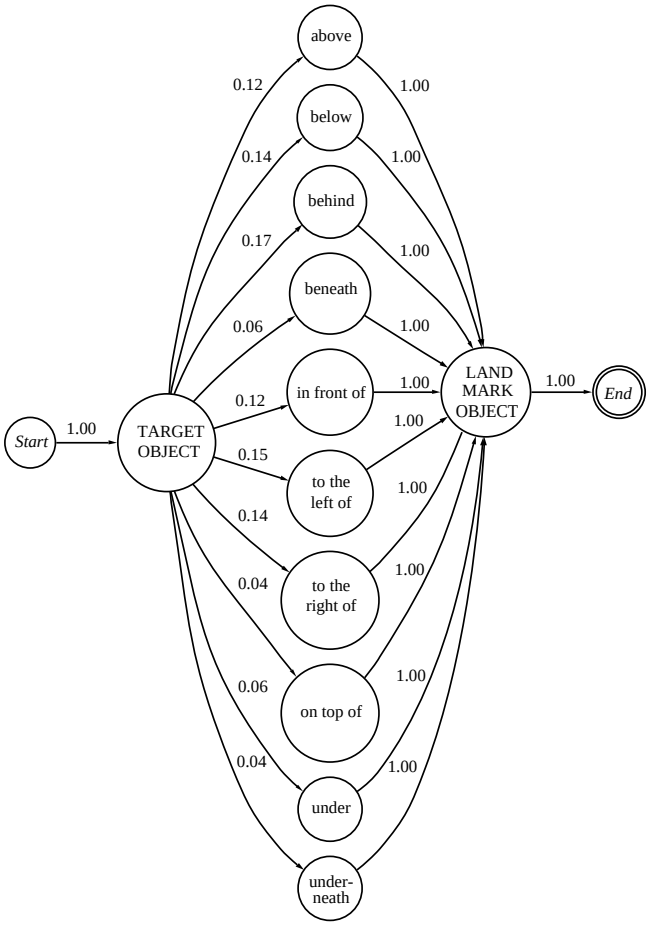

Fig. 5. The probabilistic grammar used to generate descriptions with relative spatial clauses.

describe those objects become more probable and thus steer the speech recognizer towards those parts of the vocabulary. The expected *order* of words as specified by the bigram class probability transitions remains static throughout this process.

C. Relative Spatial Clauses

Phrase bigrams are used to model the use of relative spatial clauses. For example, "The red block beneath the small green one" contains references two objects, the target and a *landmark* ("the small green one"). The spatial relation "beneath" describes the relation between target and landmark. The class bigram network is augmented with spatial relation terms as shown in Figure 5.

Spatial connective terms may consist of multiple words (e.g., "to the left of") but are tokenized and treated as a single acoustic unit during speech decoding. Each spatial term is grounded in a Guassian model that is defined over the spatial features described in Section IV. To combine the spatial grammar (Figure 5) with the simpler single-object description grammar (Figure 4), an additional transition probability is needed: the probability that a description will contain a relative spatial phrase. This is estimated based on frequency of occurrence of relative phrases in the training data.

VII. Language Driven Visual Attention

As Fuse processes incoming speech and generates partial word sequences, a model of visual attention is incrementally updated to reflect the system's current belief of the intended referent object. Attention consists of a probability mass function (pmf) spread over the objects in the scene. This pmf is used to mix object-dependent description bigrams into a single weighted bigram. Thus, as speech is processed, the evolving distribution of attention shifts the weight of bigrams to favor descriptions of objects that hold more attention. The Visual Attention model enables the early integration of visual context to provide dynamic incremental estimation of the priors associated with the interpolated class conditional probabilities. In other words, the model uses the visual context to immediately determine the attention distribution spread over the objects in the current scene.

The speech decoder used in Fuse is based on a single pass Viterbi beam search [25]. In this strategy, multiple word sequences are considered during a forward pass, and in a backward pass the best word sequence is selected. In the following, we show how the visual attention model, $P(O_i)$, is computed for a partial word sequence. Separate attentional pmf's are maintained for each parallel word sequence hypothesis. The average pmf over all search paths of the decoder may be interpreted as the system's overall attention at any given point of time.

At the start of each utterance, before any words have been processed, visual attention is shared equally by all M objects in the scene:

$$P(O_i)[0] = \frac{1}{M} \tag{11}$$

The index marks that the this is the pmf when 0 words have been processed. As each new word w_n posited in one of the search paths of the speech decoder, the path-dependent attention pmf is incrementally updated using one of three update rules depending on the new word:

1) w_n is a visually-grounded word (e.g., "large", "blue", etc.). In this case, the update rule is:

$$P(O_i)[n] = \frac{p(O_i|w_n)P(O_i)[n-1]}{\sum_{j=1}^{M} p(O_j|w_n)P(O_j)[n-1]} \tag{12}$$

As a result of this udpate rule, the visual models corresponding to modifier terms of a single object are multiplied.

2) w_n is a visually-grounded spatial relation (e.g., "above", "beneath", etc.). The update rule is:

$$P(O_i)[n] = \frac{\sum_{j=1,j\neq i}^{M} p(O_i|w_n, O_j)P(O_j)[n-1]}{\sum_{k=1}^{M}\sum_{j=1,j\neq k}^{M} p(O_k|w_n, O_j)P(O_j)[n-1]} \tag{13}$$

where $P(O_j|w, O_i)$ is derived from visual models of spatial relations

in which O_i is the target object and w is the relative spatial term. This update rule causes the attention of the system to shift to objects that hold the spatial relation

indicated by w_n relative to whatever object has been described by the partial word sequence $w_1 \ldots w_{n-1}$.

3) w_n is a visually ungrounded word (e.g., "the", "by", etc.). The update rule is:

$$P(O_i)[n] = \gamma P(O_i)[n-1] \quad (14)$$

where γ is a constant likelihood assigned to all ungrounded words. Visually grounded words thus have no effect on visual attention.

Using these three update rules, Fuse maintains separate attentional state pmf's for each search path of the decoder.

VIII. Visually-Grounded Speech Recognition and Understanding

Processing in Fuse is initiated by the detection of a spoken utterance. A forward pass maintains multiple word sequence hypotheses in a search trellis. Following standard speech recognition practice, a beam is used to limit the number of active paths at any point in the forward pass. The visual attention model biases the search to word sequences that semantically match the properties and spatial configurations of objects in the co-occurring visual scene. Once the entire utterance has been processed (i.e., the forward pass is complete), the backchaining is used to recover the most likely word sequence.

Fuse is able to understand two classes of referring expressions which we refer to as simple and complex [10]. Simple expressions refer to single objects without use of spatial relations, and are fully modeled by the transition network shown in Figure 4. Complex expressions include relative spatial clauses and are modeled by the network shown in Figure 5.

Once the forward pass of the Viterbi beam search is complete, the best word sequence is extracted using dynamic programming [25]. We denote this word string as $W = w_1 \ldots w_N$. In the case of a simple referring expression, Fuse selects the object with greatest visual attention:

$$\underset{i}{\text{argmax}} \quad P(O_i)[N] \quad (15)$$

For complex referring expressions, we can segment W into three sub-sequences, $W = w_1 \ldots w_{m-1}, w_m, w_{m+1} \ldots w_N$ where w_m is a relative spatial term, $w_1 \ldots w_{m-1}$ describes the target object, and $w_{m+1} \ldots w_N$ describes a landmark object. Fuse selects O_i based on:

$$\underset{i}{\text{argmax}} \quad P(O_i)[m-1] \sum_{j=1, j\neq i}^{M} p(O_j|w_m, O_i) P(O_j)[N] \quad (16)$$

where $p(O_j|w_m, O_i)$ is derived from the visual model associated with the relative spatial term w_m.

A. A Detailed Example of Visually-Steered Speech Processing

To make the interaction between visual attention and speech processing more concrete, we take a closer look at an example. Table I shows the transcription of a sample utterance, the output of the speech decoder using standard bigrams without use of the visual attention model, and the decoder's output using visual attention.

Errors from the decoder are underlined, and omitted words are indicated by square parentheses. Corrections due to visual context are shown in italics. The introduction of visual context in this case makes two important differences. First, the word *lower* is corrected to *large*, and the incorrectly decoded words *to me* are changed to *beneath*.

The evolution of visual attention is illustrated for this example in Figure 6. Each plot shows the spread of attention across the ten objects after integrating the words shown to the left of that plot. The word sequence that was selected as most probably during the backward pass of the decoder is shown. Ungrounded words are shown in parentheses and do not effect the attention pmf. Attention vectors are normalized within each plot so only relative values with plots are significant. As evidence for the target object accumulate from the first part of the utterance, "The large green block in the far right", the pmf becomes progressively sharper with most probability mass focused on Object 8. When the relative spatial term "beneath" is incorporated, visual attention is captured almost equally by Objects 9 and 10 which are the two smaller blocks above Object 8. Thus, the grounded model associated with "beneath" has caused attention to shift appropriately. The remainder of this utterance refers to two objects. Fuse is designed on the assumption that the remaining phrase will refer to only a single object. Due to the soft assignment of visual attention, however, Fuse is able to robustly deal with the phrase "the yellow block and the red block" by assigning roughly equal attention to both landmark objects. When Equation 16 is applied to example, the correct object (Object 8) is selected by Fuse.

B. Experimental Evaluation

A corpus of 990 utterances paired with corresponding visual camera images was collected from eight speakers. Each utterance describes one object in a scene containing ten objects. To evaluate Fuse, we used a leave-one-speaker-out train and test procedure. For each speaker, their data was held out and the remaining data was used to train word class bigrams.

Speech recognition and understanding errors are shown in Tables II and III, respectively. Speech recognition errors are measured using the standard NIST measurement package [] which aligns decoder output with transcripts with equal penalty for word insertions, deletions, and substitutions. Averaged across all eight speakers, the word recognition error rate is reduced by 31% when visual context is used. This result shows that early integration of visual context has significant impact on the recognition of speech that refers to the contents of the scene.

The effects of visual context on speech understanding are even more significant. Since each visual scene had 10 objects, random selection would lead to an average error rate of 90%. The first column of Table III shows that even without visual context, i.e., using a standard speech recognizer, the system works quite well, with an average error rate of 24% (i.e., the system chooses the correct object 76% of the time). This system is similar to that described previously in [9]. The

Transcript	The large green block on the far right beneath the yellow block and the red block.
No visual context	[The] lower green block in the far right to me [the] yellow block in the red block
Visual context	The *large* green block in the far right *beneath* the yellow block in the red block

TABLE I

A EXAMPLE OF SPEECH TRANSCRIPTION WITHOUT THE USE OF VISUAL CONTEXT, AND IMPROVED OUTPUT FROM FUSE WITH VISUAL CONTEXT. DELETION ERRORS ARE MARKED IN SQUARE PARENTHESES AND SUBSTITUTION ERRORS ARE UNDERLINED.

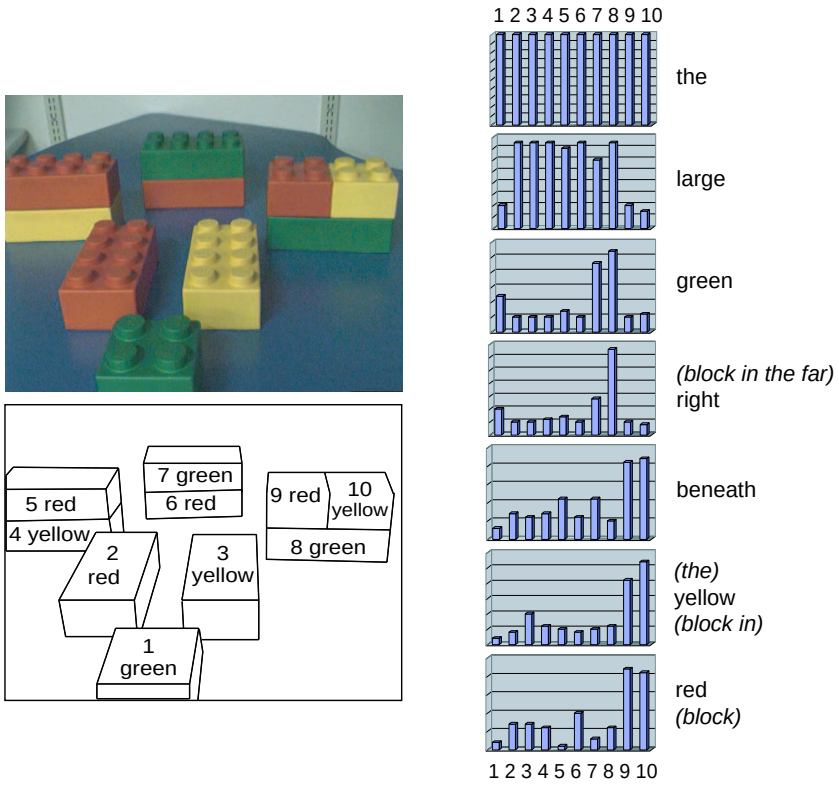

Fig. 6. Evolution of attention during processing of the utterance, "The large green block in the far right beneath the yellow block and the red block".

second column of Table III shows the change in understanding errors once visual attention is integrated into the speech decoding process. On average, the number of understanding errors drops by 41%, so that Fuse chooses the correct object 86% of the time. Thus, the early influences of vision on speech processing flow through the system and have substantial effects on overall understanding performance.

C. Analysis of Errors: Suggestions for Future Directions

We have observed five significant causes if speech understanding errors, each of which suggests extensions to the current Fuse architecture:

- Speech end point detection errors: The speech segmentation module in our real time speech recognition system occasionally merges utterances that should have been processed separately. Later stages of Fuse are designed on the assumption that only one referring expression is contained in the utterance. A possible extension is to integrate speech segmentation with semantic analysis.
- Descriptions with more than one landmark object: We assume that a complex referring expression consists of a target object description, and optionally a landmark object description with connective relative spatial term or phrase. Thus, Fuse cannot always handle cases where the referring expressions contain descriptions of more than one landmark objects in conjunction or groups of landmark objects (although the example in Section VIII-A demonstrates that sometimes this problem can be overcome in the current approach). This shortcoming suggests

Speaker	No Visual Context	With Visual Context
1	28.2	21.7
2	24.6	14.3
3	26.9	17.2
4	23.7	16.6
5	19.2	14.5
6	21.3	13.3
7	24.3	17.1
8	26.0	18.8
Ave	24.3	16.7

TABLE II

SPEECH RECOGNITION WORD ERROR RATES (%). AVERAGED ACROSS ALL EIGHT SPEAKERS, THE INTRODUCTION OF VISUAL CONTEXT REDUCED THE WORD ERROR RATE BY 31%.

Speaker	No Visual Context	With Visual Context
1	27.4	17.6
2	25.5	12.1
3	27.8	14.8
4	23.3	17.0
5	23.0	13.2
6	23.5	13.9
7	23.8	13.1
8	21.2	12.6
Ave	24.4	14.3

TABLE III

SPEECH UNDERSTANDING ACCURACY RESULTS (%). AVERAGED ACROSS ALL EIGHT SPEAKERS, THE EARLY INTEGRATION OF VISUAL CONTEXT REDUCED THE LANGUAGE UNDERSTANDING ERROR RATE BY 41%.

the use of more complex grammars, and treatment of semantic composition that goes beyond the multiplication of probability densities. For some steps in this direction, see [27].

- Error Propagation: Due to the feed-forward design of the visual attention update algorithm, errors that creep in during initial stages of decoding are propagated throughout the entire utterance. To remedy this, and other related problems, the notion of confidence might be introduced to the visual attention model. For example, the number of active search paths within the Viterbi beam search, which is often used as a source for estimating acoustic confidence in speech recognizers [28], might similarly be used as the basis for estimating confidence of the visual attention pmf. When confidence is low, the effects of attention could be discounted.
- Visual Segmentation Errors: Some errors in understanding occur due to imperfect image segmentation performed by the visual analysis system. Such segmentations may merge more than one objects or divide an object into two or more parts. These cause mismatches among descriptions and the corresponding objects. This problem suggests early integration of speech into visual processing, the complement of the integration we have explored in Fuse. Referring back to Figure 2, this suggests that the visual scene analysis module might be brought into the processing loop. If the speech decoder confidently reports the phrase "the two blue blocks on the right", this might help the visual analyzer decide between interpreting a stack of blocks as a single block versus two.
- Visual-Semantics Acquisition: Some errors are due to poor visually-grounded models that did not generalize to test data. A simple fix might be to collect more training data. In the long term, we believe that robust visual models must be dynamic to account for context-sensitive shifts of word usage, as well as speaker-dependent shifts of word usage. We are currently investigating dynamic grounded models to address this issue.

IX. CONCLUSIONS AND FUTURE DIRECTIONS

We have presented an implemented model that integrates visual context into the speech recognition and understanding process. In contrast to previous work, Fuse makes use of context at the earliest stages of speech processing, resulting in improved performance in an object selection task. The main idea that this work demonstrates is the payoff of strategically breaking modular boundaries in language processing. A key to achieving this cross-module integration is a model of how natural language semantics relates to visual features of a scene.

Looking ahead, we plan to expand this work along two directions. First, Fuse will be integrated into an interactive manipulator robot [6]. Fuse will have access to representations in the robot's visual system and also its planning and memory systems, leading to an enriched encoding of context to help guide speech processing. Second, we plan to extend Fuse to work with non-visual context cues such as geographical position and time of day in order to build context-aware assistive communication devices [29].

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0083032.

REFERENCES

[1] J. Fodor, *The Modularity of Mind*. MIT Press, 1983.
[2] L. A. Hirschfeld and S. A. Gelman, Eds., *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, 1994.
[3] M. J. Spivey, M. J. Tyler, K. M. Eberhard, and M. K. Tanenhaus, "Linguistically mediated visual search," *Psychological Science*, vol. 12, no. 4, pp. 282–286, 2001.
[4] M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy, "Integration of visual and linguistic information during spoken language comprehension," *Science*, vol. 268, pp. 1632–1634, 1995.
[5] M. J. Spivey-Knowlton, M. K. Tanenhaus, K. M. Eberhard, and J. C. Sedivy, "Integration of visuaspatial and linguistic information: Language comprehension in real time and real space," in *Representation and processing of spatial expressions*, P. Oliver and K.-P. Gapp, Eds. Erlbaum, 1998.
[6] D. Roy, K.-Y. Hsiao, and N. Mavridis, "Coupling robot perception and on-line simulation: Towards grounding conversational semantics," forthcoming, 2003.
[7] M. K. Brown, B. M. Buntschuh, and J. G. Wilpon, "SAM: A perceptive spoken language understanding robot," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22 . IEEE Transactions 22, pp. 1390–1402, 1992.
[8] C. Crangle and P. Suppes, *Language and Learning for Robots*. Stanford, CA: CSLI Publications, 1994.
[9] D. Roy, P. Gorniak, N. Mukherjee, and J. Juster, "A trainable spoken language understanding system for visual object selection," in *International Conference of Spoken Language Processing*, 2002.
[10] D. Roy, "Learning visually-grounded words and syntax for a scene description task," *Computer Speech and Language*, vol. 16(3), 2002.

[11] J. M. Lammens, "A computational model of color perception and color naming," Ph.D. dissertation, State University of New York, 1994.
[12] T. Regier, *The human semantic potential*. Cambridge, MA: MIT Press, 1996.
[13] T. Regier and L. Carlson, "Grounding spatial language in perception: An empirical and computational investigation," *Journal of Experimental Psychology*, vol. 130, no. 2, pp. 273–298, 2001.
[14] J. Siskind, "Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic," *Journal of Artificial Intelligence Research*, vol. 15, pp. 31–90, 2001.
[15] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.
[16] L. Steels, "Language games for autonomous robots," *IEEE Intelligent Systems*, vol. 16, no. 5, pp. 16–22, 2001.
[17] D. Perzanowski, A. Schultz, W. Adams, K. Wauchope, E. Marsh, and M. Bugajska, "Interbot: A multi-modal interface to mobile robots," in *Proceedings of Language Technologies 2001*, Carnegie Mellon University, 2001.
[18] L. Bahl, F. Jelinek, and R. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Journal of Pattern Analysis and Machine Intelligence*, vol. 2, no. 5, pp. 179–190, 1983.
[19] T. Niesler and P. Woodland, "Variable-length category n-gram language models," *Computer Speech and Language*, vol. 21, pp. 1–26, 1999.
[20] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixture vs. dynamic cache models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 30–39, 1999.
[21] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.
[22] F. Jelinek, J. Lafferty, and R. Mercer, *Speech recognition and understanding: Recent advances, trends, and applications*. Springer-Verlag, 1992, ch. Basic methods of probabilistic context-free grammars, pp. 345–360.
[23] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
[24] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
[25] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.
[26] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Wiley, 1973.
[27] P. Gorniak and D. Roy, "Grounded semantic composition for visual scenes," forthcoming, 2003.
[28] R. Rose, "Word spotting from continuous speech utterances," in *Automatic Speech and Speaker Recognition*, C. Lee, F. K. Soong, and K. Paliwal, Eds. Kluwer Academic, 1996, ch. 13, pp. 303–329.
[29] E. Dominowska, D. Roy, and R. Patel, "An adaptive context-sensitive communication aid," in *Proceedings of the CSUN International Conference on Technology and Persons with Disabilities*, Northridge, CA, 2002.