A Visual Context-Aware Multimodal System For Spoken Language Processing

Niloy Mukherjee, Deb Roy

The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Cambridge, MA 02142
niloy@media.mit.edu, dkroy@media.mit.edu

Abstract

Recent psycholinguistic experiments show that acoustic and syntactic aspects of online speech processing are influenced by visual context through cross-modal influences. During interpretation of speech, visual context seems to steer speech processing and vice versa. We present a real-time multimodal system motivated by these findings that performs early integration of visual contextual information to recognize the most likely word sequences in spoken language utterances. The system first acquires a grammar and a visually grounded lexicon from a "show-and-tell" procedure where the training input consists of camera images consisting of sets of objects paired with verbal object descriptions. Given a new scene, the system generates a dynamic visually-grounded language model and drives a dynamic model of visual attention to steer speech recognition search paths towards more likely word sequences.

1. Introduction

Recent psycholinguistic experiments [1] have shown that acoustic and syntactic aspects of online spoken language comprehension are influenced by visual context. During interpretation of speech, partially recognized utterances seem to incrementally steer the hearer's visual attention [2] and vice versa, visual context steers speech-processing [3]. We consider here the problem in which spoken language is used to make reference to objects in a physical environment. This problem highlights the importance of applying contextual knowledge from the environment to anticipate words and phrases that the spoken utterance is likely to contain.

We describe an on-line, real-time, multimodal processing system that performs speech recognition using visually steered dynamic language models. The system processes referring expressions such as "the large green block beneath the red and the yellow blocks". Figure 1 provides an overview of how the system integrates speech with visual context. The visual scene analysis module detects objects in a given scene and extracts a set of visual features representing individual objects and inter-object spatial relations. The language model generation component acquires visually-grounded word semantics from a set of visual features and generates a dynamic class conditional language model [4, 5] each time a new scene is presented. For each object in the scene, the model estimates the class-conditional word likelihoods using the acquired visual semantics. During speech recognition, partially decoded utterances feed back to a visual attention module, which distributes visual attention over the set of objects by updating a probability mass function over objects. As decoding proceeds, visual and linguistic information reinforce each other, to recognize the mostly likely word sequence in the utterance.

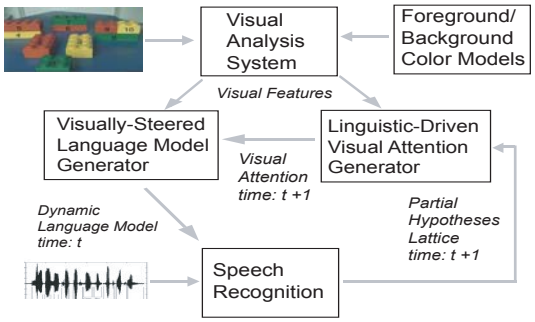

Figure 1: *The schematic system overview.*

To study the role of visual context in spoken language comprehension, we developed a scene description task similar to [6]. Participants in a data collection study were asked to verbally describe objects in scenes comprising of large Lego bricks (Figure 2) without restrictions on the vocabulary, style or length of descriptions. The system acquired a grammar and visually relevant vocabulary from the training input consisting of a set of camera images and spoken descriptions of objects in the images. The system was evaluated on the training input in a leave-one-speaker-out fashion with and without the accompanying visual information to study the effect of visual information on the speech recognition performance.

This work will be applied to improve speech recognition and understanding in a conversational robot under construction in our lab. This research may also find application in areas that involve speech recognition / understanding on portable devices such as wearable computers and handhelds that are context-aware through a range of modalities such as location (GPS), time of day, etc. For example, the speech processing performance of a context-aware travel assistant may be improved if it applies user context and other modalities to "second guess" what the user is likely to say.

The remainder of this paper proceeds by first describing the visual analysis system. Section 3 describes the visually grounded language model generation module. This module learns to map from visual scenes to descriptive word sequences. This visual-to-language mapping serves as a basis for our model of linguistically-steered visual attention, and our model of visually-steered language models. Section 4 and 5 present the dynamic integration of visual information into the speech recognition, followed by evaluation and concluding remarks.

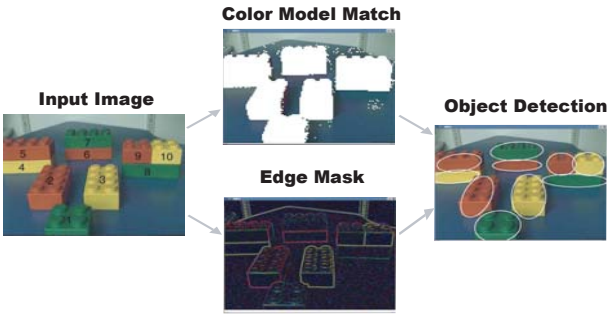

Figure 2: *The object detection procedure.*

2. Visual Scene Analysis System

The visual analysis system, similar to the system reported in [7], supports tracking of multiple partially occluded solid-colored objects placed on a table top in real-time. The system detects objects in a scene and extracts object properties and inter-object spatial relationships that are passed to the language model generation module as well as the visual attention module.

Figure 2 describes object detection combining local edge finding with color model based foreground/background separation. Once objects are detected, intra-object features including object color, shape, height, width, area, center, height-to-width ratio and ratio of maximum dimension to minimum dimension are extracted. Inter-object spatial relations including center-of-mass distances between objects and others similar to [8] are also extracted from the images.

3. Visually-Grounded Language Model Generation

The language model generation system performs automatic acquisition of visual semantics of individual words and grammar to generate dynamic class conditional language models representing a given visual scene. The training data consists of visual feature vectors extracted from the scene analysis system paired with transcriptions of expressions referring to target objects.

3.1. Learning

3.1.1. Word Class Formation

In order to generate dynamic class conditional language models depicting objects in a scene, word classes that integrate semantic structure of visually relevant words must be learned. A weighted combination of two methods of clustering words into syntactically equivalent classes was investigated. The first relies on distributional analysis of word bigram occurrence patterns [9]. The second method clusters words that co-occur in similar visual contexts [4]. This method uses shared visual grounding as a basis for word classification.

3.1.2. Grounding Word Semantics

A subset of visual features is automatically selected and associated with each word class. This is done by a search algorithm [6] based on Kullback-Leibler divergence that finds the subset of visual features for which the distribution of feature values conditioned on the presence of the word is maximally divergent from the unconditioned feature distribution. For each word (token type), a multidimensional Gaussian model of feature distributions is computed using all observations that co-occur with that word using the feature subset associated with the corresponding word class.

3.1.3. Learning Word Order

A class-based bigram statistical language model [4] is estimated (based on frequency) to model the syntax of referring expressions. Visually "ungrounded" words form singleton word classes (classes with only one member). The bigram statistical language model representing the word class syntax is learned from the training data in a leave-one-speaker-out form. The model is static and does not change with the visual scene.

3.1.4. Spatial Terms Acquisition based on Focused Training

To learn spatial relations between a target and a landmark object, a user interface was created to enable the user to load an image on the screen and type in spatial phrases such as "above", "below" and "left_of". Participants were instructed to select two objects from the scene that he or she found suitable to serve as the target and the landmark objects for the spatial phrase. Spatial relations between the target and the landmark objects are extracted by the visual analysis system and a multidimensional Gaussian model is computed for each of the spatial lexical items using all observations that co-occur with the lexical item.

3.2. Generation of Dynamic Language Models Using Visual Features

3.2.1. Generation of Object based Descriptions

The generation problem is treated as a two-step constrained search problem. The first considers syntactic constraints that determine the sequence of classes in a T length description. The second constraint is semantic. The semantic constraints select individual words from the word classes that best fit the visual features of the target object.

Syntactic Constraints: The acquisition algorithm described in the previous section generates a static class-based bigram language model that incorporates the entire vocabulary of the task. A subset language model of word classes representing visually relevant object properties is chosen. The subset language model serves as a class based finite state search space. Every possible word class sequence is generated for descriptions of increasing length T for each of the objects in the scene.

Semantic Constraints: Each word class C_i in an utterance may be mapped to a word by choosing the word $C_i(j)$ from the class C_i which maximizes the probability of the target object x. Equal priors are assigned to words in a word class, i.e., $P(C_i(j)|C_i)$ is same for every $C_i(j)$ in C_i.

3.2.2. Scoring based on Contextual Constraints

The utterances generated by the method described above can be ambiguous. Non-target objects in the scene might accidentally match a generated description. Thus, the descriptions are rescored based on a measure of ambiguity [6] in the context of the set of remaining objects. This score finds the closest matching non-target and computes the likelihood ratio of the non-target to target semantic match.

3.2.3. Generation of Class Conditionals of Intra-Object Lexical Items

The lexical items in the rescored descriptions are mapped back to their word classes. For a given word class occurrence in a

given object description, a word occurrence is estimated by the weighted average over the probability scores over all the words in the word class based on the object features. The word class counts as well as the word counts are also weighted by the score of the description they belong to. The procedure estimates class conditionals in the form $P(word|word\_class, object)$ for all objects in the scene.

3.2.4. Generation of Spatial Language Model

To deal with complex descriptions that comprise more than one object and a relative spatial term, we generate a spatial language model. For every possible combination of target and landmark objects in the scene, we use the Gaussian models to score spatial terms given the combination. We generate a language model of the form of $P(target, spatial\_term, landmark)$ where $target$ and $landmark$ are the target object id and the landmark object id.

4. Speech Recognition Using Visual Context

Our continuous speech recognition system [10] aims to search for the word string W' such that

$$W' = \arg\max P(A|W)P(W)$$

The effect of visual contextual information does not play any role to infer on $P(A|W)$ since they are estimated from the acoustic signal alone. The visually steered linguistic models generated by the methods described in section 3 are used to evaluate $P(W)$.

Suppose W is a word string such that $W = w_1, w_2, ..., w_n$, then a class-based bigram language model [4] approximates $P(W)$ as

$$P(W) = P(w_1|c_1)P(C_2|c_1)...P(c_n|c_{n-1})P(w_n|c_n)$$

The factor $P(c_i|c_{i-1})$ is estimated from the static class based leave-one-speaker-out bigram language model . The dynamic class conditional language models generated by the methods described in section 3 to evaluate $P(w_i|c_i)$. We consider three cases

- w_i is a visually grounded word depicting an intra-object property: We linearly interpolate the generated class conditionals over the number of objects in the scene to evaluate $P(w_i|c_i)$

  $$P(w_i|c_i) = \Sigma_{j=1}^{n}\lambda_j P(w_i|c_i, object_j), \quad \Sigma_{j=1}^{n}\lambda_j = 1$$

  where λ_j denotes the prior of the class conditional of the j_{th} object and n is the number of objects in the scene.

- w_i is a visually grounded word depicting a spatial relation: We perform a linear interpolation over the spatial language model. The interpolation is in the form of

  $$P(w_i|c_i) = \Sigma_{j=1}^{n}\lambda_j P(w_i|c_i, target\_object_j), \\ \Sigma_{j=1}^{n}\lambda_j = 1$$

  The spatial language model generated by methods described in Section 3 is used to derive the probability mass functions in the form of

  $$P(spatial\_term|wordclass, target\_object)$$

- w_i is a visually ungrounded or visually irrelevant word: $P(w_i|c_i)$ is estimated from the static leave-one-speaker-out corpus.

The priors provide a measure of visual attention in context of the current scene. Their values are determined dynamically from active partial utterance hypotheses during the Viterbi search.

5. Using Incremental Speech Processing to Drive Visual Attention

The Visual Attention model enables the early integration of visual context to provide dynamic re-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. Given a partial utterance hypothesis, the model rank-orders and scores each object in the current scene based on the visual semantic fit over the partially decoded utterance. These scores are used as the interpolation weights to calculate the class conditional in the form of $P(w_i|c_i)$.

The priors $lambda_j, \quad j = 1, ..., n$ gets dynamically updated when the decoder search algorithm leaves a state that marks the end of a word w_m. The partial hypothesis is of length m at this point. From the nature of the collected utterances, there exist three cases that are described below:

- w_m is a visually relevant word depicting intra-object property, for example "large", "vertical", etc. Here, the update rule is as follows

  $$P'(object_j|w_m) = P(x_j|w_m)P(object_j|w_{m-1}), \\ j = 1, 2, ..., n$$

  where x_j is the visual feature subset of $object_j$. Therefore

  $$\lambda_j = \frac{P'(object_j|w_m)}{\Sigma_{i=1}^{n}P'(object_i|w_m)}, \quad j = 1, 2, ..., n$$

- w_m is a visually relevant word depicting a spatial relation, for example "above", "beneath" and so on. Here, the update rule is in the form

  $$P'(object_j|w_m) = \\ \Sigma_{i=1, i\neq j}^{n}P(object_j|w_m, object_i)P(object_i|w_{m-1}) \\ j = 1, 2, ..., n$$

  where $P(object_j|w_m, object_i)$ is derived from the dynamic spatial language model. Again,

  $$\lambda_j = \frac{P'(object_j|w_m)}{\Sigma_{i=1}^{n}P'(object_i|w_m)}, \quad j = 1, 2, ..., n$$

- w_m is a visually irrelevant or ungrounded word such as "the", "by", etc. In this case, we have the following update rule:

  $$P'(object_j|w_m) = \gamma P(object_j|w_{m-1}), \\ j = 1, 2, ..., n$$

  where γ is a constant score given to the likelihood of ungrounded words. The priors are updated by the same rules described above.

A detailed example of the visual attention procedure is presented through figure 3. Each plot shows the spread of attention across the ten objects in the scene after integrating the words shown in the left of the plot. This occurs for all active partially decoded hypotheses but only the hypothesis containing the most likely word sequence is shown in the plot. Words that have no visually-grounded models have no effect on visual attention and are not shown in the diagram.

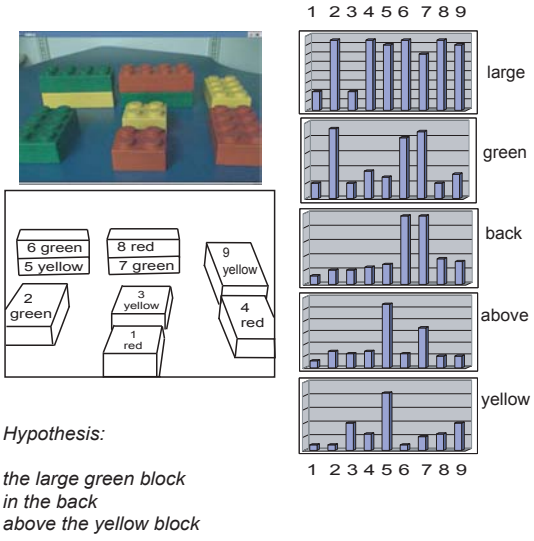

Figure 3: *Dynamic update of visual attention.*

6. Evaluation

As a preliminary evaluation, a dataset of 990 utterances paired with corresponding visual camera images was collected from eight speakers. Each utterance describes one object in a scene of ten objects. The system acquired a visually-grounded vocabulary from the entire dataset. A leave-one-speaker-out class based bigram language model was trained from the dataset.

To study the effect of visual context on speech processing, the speech recognition performance accuracy of the system was evaluated with and without the accompanying visual information. During the evaluation without visual context, the class conditionals were distributed equally among words occurring in the same word class for all visually relevant word classes. The introduction of visual context led to a 31.3% reduction in word error rate, a significant improvement over the baseline system.

Table 1 compares the speech recognition word error rates averaged across all eight speakers with and without the visual context.

Table 1: Speech recognition word error rates (%). Averaged across all eight speakers, the introduction of visual context reduced the word error rate by 31.3%.

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

7. Conclusion and Future Directions

We have presented a complete multimodal system which performs early integration of visual context into speech processing using visually-steered language models to recognize spoken language utterances. The semantics of referring expressions are grounded in visual primitives from the physical environment provided by a real-time visual system. The system uses color, geometry, and spatial relations to anticipate words and phrases in spoken language utterances.

We are expanding this work in two ways. First, we are performing experiments in which human listeners replace our system and find objects in scenes based on spoken descriptions. While they perform the listening task, we will record their eye-movements using a head-worn eye tracker. We will compare the evolution of visual attention in our system to that of human participants. Potential outcomes of this work include cognitive models of how people perform situated language comprehension, and new insights into how to design our systems. Second, we are investigating integration of other non-visual sources of context for other application domains including assistive communication aids.

8. References

[1] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. E., ”Integration of visual and linguistic information in spoken language comprehension”, Science, 268, 1632-1634, 1995

[2] Spivey, M.J., Tyler, M. J., Eberhard, K. M., Tanenhaus, M. K., ”Linguistically mediated visual search”, Psychological Science, 12 (4), 282-286, 2000

[3] Tanenhaus, M. K., Magnuson, J. S., Dahan, D., Chambers, C., ”Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing”, Journal of Psycholinguistic Research, 29, 557-580, 2000

[4] Brown, P., F., Della Pietra, V., J., deSouza, P., V., Lai, J., C., Mercer, R., L., ”Class-based n-gram models of natural language”, Computational Linguistics, 18 (4), 467479, December, 1992.

[5] Rosenfeld, R., ”Two decades of statistical language modeling: Where do we go from here?”, Proc. of the IEEE, 88, 1270-1278, August, 2000.

[6] Roy, D., ”Learning visually-grounded words and syntax for a scene description task”, Computer Speech and Language, 16 (3), 2002.

[7] Roy, D., Gorniak, P. J., Mukherjee, N., Juster, J., ”A trainable spoken language understanding system for visual object selection”, Proc. of the International Conference on Spoken Language Processing, 2002.

[8] Regier, T., Carlson, L., ”Grounding spatial language in perception: An empirical and computational investigation”, Journal of Experimental Psychology: General, 130 (2), 273 - 298, 2001.

[9] Kneser, R., Ney, H., ”Improved clustering techniques for class-based statistical language modeling”, Proc. of the European Conference on Speech Communication and Technology, 1993.

[10] Benjamin Yoder, ”Spontaneous speech recognition using hidden markov models”, M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001