# Improving Automatic Speech Recognition Through Head Pose Driven Visual Grounding

**Soroush Vosoughi**
MIT Media Lab
75 Amherst St, E14-574K
Cambridge, MA 02139
soroush@mit.edu

## ABSTRACT

In this paper, we present a multimodal speech recognition system for real world scene description tasks. Given a visual scene, the system dynamically biases its language model based on the content of the visual scene and visual attention of the speaker. Visual attention is used to focus on likely objects within the scene. Given a spoken description the system then uses the visually biased language model to process the speech. The system uses head pose as a proxy for the visual attention of the speaker. Readily available standard computer vision algorithms are used to recognize the objects in the scene and automatic real-time head pose estimation is done using depth data captured via a Microsoft Kinect. The system was evaluated on multiple participants. Overall, incorporating visual information into the speech recognizer greatly improved speech recognition accuracy. The rapidly decreasing cost of 3D sensing technologies such as the Kinect allows systems with similar underlying principles to be used for many speech recognition tasks where there is visual information.

### Author Keywords

visual grounding; language models; automatic speech recognition; head pose estimation; visual attention

### ACM Classification Keywords

I.2.7 Natural Language Processing: language models; H.5.1 Multimedia Information Systems: miscellaneous

## INTRODUCTION

A significant number of psycholinguistic studies have shown a very strong connection between a person's eye gaze and what the person says [4, 6, 5]. More recently Coco and Keller [2] have shown that where people look is a good indicator of what they will say. One study suggests that gaze direction is tightly connected with the focus of a person's attention [6]. Given this apparent tight link between human gaze and speech, we wanted to utilize human gaze information to improve real-time speech recognition accuracy. Integration of visual information into speech recognition has been done before [9, 12, 8, 7, 10, 11], however these studies either ignored gaze direction, were done in virtual worlds or used an eye-tracker to get gaze information. We wanted a system that can be used in the real world but wearing an eye-tracker is uncomfortable and impractical in many real life situations, so we decided to build a system that integrated gaze information into a real-time speech recognizer without requiring the humans to wear an eye-tracker.

We accomplished this by using the Microsoft Kinect to estimate the head pose of a speaker. We then use the head pose information as a proxy for eye gaze and integrate it into a real-time speech recognizer. The advantage of using a Kinect is that it does not require the user to wear tracking devices. Moreover, the rapidly decreasing cost of 3D sensing technologies such as the Kinect allows systems with similar underlying principles to be used for many speech recognition tasks where there is visual context. Even though gaze direction has been used in speech recognition systems in prior work [9, 8] our approach (replacing the eye-tracker with the Kinect) allows the technique to be moved off the desktop and into the real world.

## OVERVIEW

We implemented our system around a scene description task. A scene description task was chosen so that we could safely assume that people in our experiments would only talk about objects that they could view. The scenes consisted of various geometric objects of difference sizes and colors laid out on a white table which was $200cm$ long and $60cm$ wide. The size of the objects varied anywhere between $190cm^2$ and $40cm^2$. There were a total of 6 different shapes and 8 different colors. Overall there were 50 unique objects. There was a camera in the ceiling ($165cm$ from the surface of the table) that captured the whole scene. You can see a sample scene laid out on the table and captured by the ceiling cam in Figure1.

There was a Kinect stationed in front of the table (visible in Figure1) facing the user who was asked to stand no further than $40cm$ from the table. The user was then asked to describe the objects in various scenes. The participants were not given any instructions on what to say and were asked to describe the scene as best as they could. Below you can see a few sample utterances that the participants used to describe various scenes:

*CHI'14*, April 26–May 01, 2014, Toronto, ON, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2473-1/14/04...$15.00.
http://dx.doi.org/10.1145/2556288.2556957

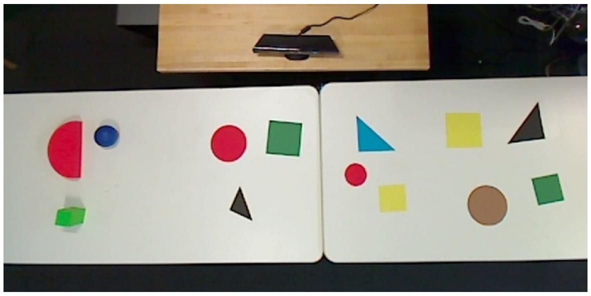

**Figure 1. The view of a sample scene from the camera on the ceiling. You can see the Kinect and the objects on the table. The user is not visible from this view. The camera is approximately 165 cm from the surface of the table.**

- "There is a large red circle to the left of a green triangle and above a small black triangle."
- "There is a blue ball to the left of everything else."
- "There is a small black triangle, a large brown circle and yellow rectangle."

## VISUAL PROCESSING

This section describes the visual processing algorithms used by the system.

### Object recognition

The ceiling cam's only purpose is to identify all objects laid out on the table. In order to detect the objects on the table we used readily available standard color segmentation and edge detection algorithms available in opencv [1]. Objects are segmented based on color. Simple off-the-shelf algorithms are sufficient for detecting simple geometric, uniformly colored objects such as the ones used in our experiments. The shape of an object is represented by the width, height, and bounding box area. The location of the objects on the table is also extracted. Finally, simple spatial relations (such as "left of", "below", etc) between the objects on the table is also encoded. The color, size, shape and spatial relations of all the subjects on the table are extracted and encoded using the ceiling cam.

### Head pose estimation

Real-time head pose estimation was done using the *Random Regression Forests* algorithm proposed by Fanelli et al [3]. The algorithm is trained on a large corpus of different head poses and uses the depth data generated by the Kinect (or any other 3D sensor) to detect specific face parts like the nose. The algorithm is fast enough to be able to run in real-time and can estimate head direction with high accuracy [3]. Figure 2 shows an example of real-time head pose estimation on one of our participants. As you can see in the figure, not only does the algorithm calculate head pose vector, it also provides the location of the speaker (and the speaker's head and nose) which we use when estimating the speaker's visual focus. We should note that in order for the estimations to be accurate, the participant can not be more than $1.5m$ away from the Kinect.

### Visual attention estimation

We use the head pose vector to estimate the region on the table that the speaker's visual attention is focused on. Since the exact location, size and height of the table on which the objects are laid out is known (and since the table is stationary), we can define the table as a plane in 3D space. The head pose vector and the origin point of the vector (the speaker's head) are also known. The origin point and the vector can be used to define a line in 3D space. We now have a 3D plane and a line, the intersection of which is where the speaker's visual attention is most likely focused on (Figure 3). The line-plane intersection can be found using standard algebraic methods.

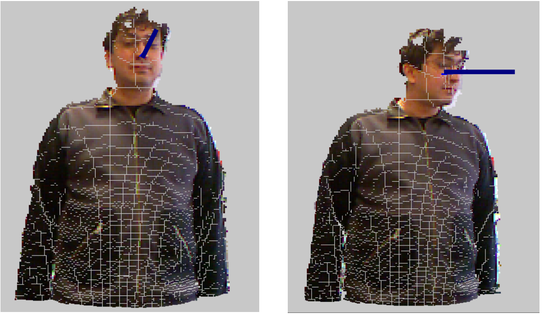

**Figure 2. Real-time head pose tracking with a Microsoft Kinect. Here we are showing the participant looking straight ahead and to his right. You can see the participant being clearly distinguished from the background. The blue line is the real-time estimate of the participant's head pose.**

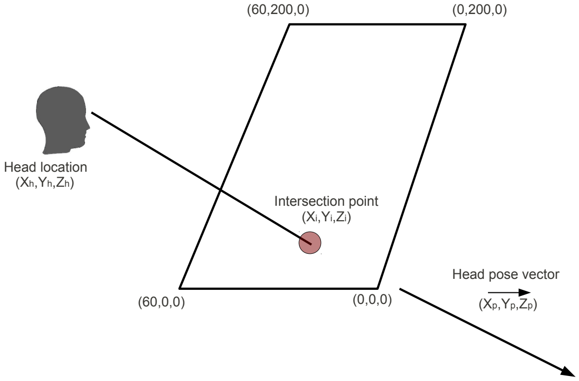

**Figure 3. The speaker's region of visual attention can be estimated by finding the intersection between a line and a plane. The plane is the table which is stationary with known coordinates. The line is constructed by combining the location of the speaker's head with the speaker's head pose vector; both of which are estimated using the Kinect.**

Visual attention is treated as a probability mass function over all objects present at the scene. Most of the probability mass is assigned to objects in the area of the speaker's gaze. The probability mass assigned to other objects drops exponentially the further the objects are from the center of the gaze.

## INCORPORATING VISUAL CONTEXT INTO SPEECH RECOGNITION

Since the focus of this work was on the integration of visual information into speech recognition systems, we decided to use an open source speech recognizer in the core of our system. We settled on the HTK speech recognizer [15]. A speech recognizer combines stochastic acoustic and language models to infer the words that are being uttered by a speaker. As the name suggests, the acoustic model deals with the acoustic aspects of speech such as the sounds that make up the phonemes and the words. The language model deals with the probability of the words being uttered based on previous examples of speech.

In our system, we used the visual context of where the speaker was visually engaged to bias our language model in real-time. The acoustic model was not updated and remained the same at all times. For our acoustic model, we used an off-the-shelf speaker independent acoustic model trained on the Wall Street Journal, using the 40 phones set from the CMU pronunciation dictionary [14].

### Visually biased language modelling

We decided to use a bigram model as the basis of our language model. A bigram language model calculates the probability of a word given the preceding word, as shown in the equation below. A bigram model has the advantage of being simple and powerful enough for our task.

$$P(w_n|w_1, w_2, w_3...w_{n-1}) = P(w_n|w_{n-1})$$

It is $P(W)$ that our system dynamically updates based on the visual context. For each object detected in the scene, the system extracts a set of visual features such as color, size and shape. The system uses these features to produces a set of simple possible referring expressions for all the objects on the table. For example for a blue ball on the table the system could generate the expression “the big blue ball”, “blue ball”, “big ball”, etc. As mentioned, this is done for all the objects on the table. The likelihoods of these expressions are then set based on the probability assigned to each object by the visual attention probability mass function described earlier. So for example if the aforementioned blue ball is far away from the speaker’s gaze then the likelihoods assigned to the expressions generated for that object will be relatively low compared to the expressions that refer to objects closer to the speaker’s gaze. These expressions and their probabilities are then used to create a bigram language model.

Many expressions could be used to refer to the same object, that’s why the system generates many possible expressions for the same object. These expressions are generated using a method similar to the *Describer* expression generation system [13]. It is possible (but very unlikely) that none of words the speaker uses to describe the objects are anticipated by the system which might make it hard for the system to decode the speaker’s speech. This is the source of most of the errors in the system.

It should be noted that this process is done in real-time, so as the speaker moves their head around as they scan the table, the visual attention’s probability mass is redistributed to the objects on the table and a new bigram model is created using the new probabilities. Simply put, the system is anticipating what the speaker will be saying based on where they are looking. Figure 5 shows an overview of the system architecture including the visual and speech pipelines.

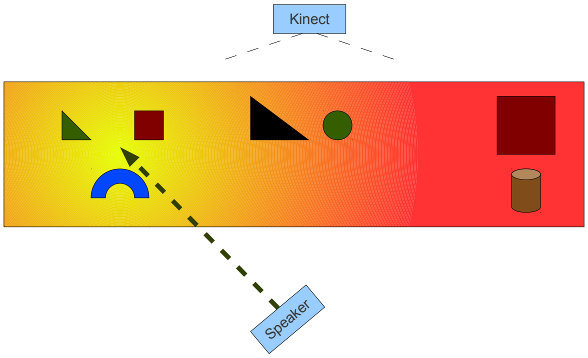

**Figure 4. The system can use head pose information to estimate the location of the speaker’s gaze. The color of the table represents the probability that the speaker is looking at that area of the table. Bright yellow is the highest probability while dark red is the lowest. The probability decreases as the objects get further away from the speaker’s estimated gaze location.**

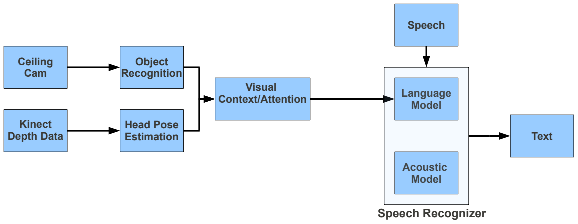

**Figure 5. Overview of the system architecture. The language model is updated dynamically in real-time based on the visual context.**

## EVALUATION

We evaluated the system on 5 participants. Each participant was presented with 20 novel scenes. The participants described each scene using somewhere between 1 to 5 utterances. Across all participants there were a total of 320 spoken utterances describing 100 scenes. Audio was captured using a headset that participants were asked to wear. We measured speech recognition error rate for each of the participants.

The speech recognition error rate was measured for three versions of our system. The first version just used the HTK speech recognizer with a bigram trained on all possible expressions that can be used to describe all of our objects (not just the objects on the table but all the objects in our repertoire) without including any visual information. The second version used the speech recognizer in conjunction with visual information about every object on the table, not using head pose information (In this case the visual attention probability mass assigns equal probability to all objects on the table). The third version was our complete system, which used visual and gaze information. The speech recognition error rate for each speaker are shown in Table 1.

When just using an off-the-shelf speaker independent speech recognizer, the accuracy of the system was fairly high (average error rate of 14.3%), which was expected given the limited nature of our description task. When incorporating visual information about every object on the table, the accuracy increased by about 34.3% (average error rate of 8.4%). This means that even without information about the visual

| Speaker | NVC | VC/NVA | VC/VA |
|---|---|---|---|
| 1 | 15.5 | 9.6 | 4.8 |
| 2 | 13.8 | 7.5 | 3.6 |
| 3 | 14.1 | 8.2 | 3.7 |
| 4 | 13.4 | 7.1 | 4.0 |
| 5 | 14.7 | 9.4 | 4.1 |
| Average | 14.3 | 8.4 | 4.0 |

**Table 1. Speech recognition error rate (%) without visual context (NVC), with visual context but no visual attention (VC/NVA) and with visual context and visual attention (VC/VA). On average incorporating visual context and attention improved the speech recognizer accuracy by 72.0%.**

attention of the speaker, just taking into account visual information about the objects in the immediate environment of the speakers helped improve speech recognition accuracy. Finally, when incorporating visual attention into the speech recognizer, the accuracy increased by about 72.0% when compared to the visual context free speech recognizer (average error rate of 4.0%), which is a rather remarkable improvement on accuracy.

## CONCLUSION

In this paper, we have shown an online, real-time, multimodal automatic speech recognition system that vastly outperforms a traditional speech recognition system. Motivated by studies that show a very strong connection between a person's eye gaze and speech, we have shown how we can use cheap 3D sensing technologies such as the Microsoft Kinect to estimate head pose direction in real-time. We have also shown how to use the estimated head direction as a proxy for human gaze and how to use that information to dynamically update the language model of a speech recognition system in real-time, improving its accuracy for certain tasks.

The system as it stands right now was created as a proof of concept, to show that incorporating gaze information into speech recognition systems is a fruitful endeavor and worth exploring. For instance, an assumption that we make in this paper is that people always talk about the objects in "here and now." This is obviously not true as human speech could consist of events and objects in different times and locations and might include abstract topics that have no visual grounding. This was the main reason why we chose a scene description task, to force humans to talk about objects in the "here and now." However, even with this assumption, the system as it stands right now could be used in games and virtual and augmented reality environments in which all of the objects and their locations are known. An example of such domain would be automotive repair or a factory assembly line.

## REFERENCES

1. Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).
2. Coco, M. I., and Keller, F. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science* (2012).
3. Fanelli, G., Gall, J., and Van Gool, L. Real time head pose estimation with random regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE (2011), 617–624.
4. Griffin, Z. M., and Bock, K. What the eyes say about speaking. *Psychological science 11*, 4 (2000), 274–279.
5. Henderson, J. M. Human gaze control during real-world scene perception. *Trends in cognitive sciences 7*, 11 (2003), 498–504.
6. Just, M. A., and Carpenter, P. A. Eye fixations and cognitive processes. *Cognitive Psychology 8*, 4 (1976), 441–480.
7. Kaur, M., Tremaine, M., Huang, N., Wilder, J., Gacovski, Z., Flippo, F., and Mantravadi, C. S. Where is it? event synchronization in gaze-speech input systems. In *Proceedings of the 5th international conference on Multimodal interfaces*, ACM (2003), 151–158.
8. Prasov, Z., and Chai, J. Y. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, ACM (2008), 20–29.
9. Prasov, Z., and Chai, J. Y. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics (2010), 471–481.
10. Qvarfordt, P., Beymer, D., and Zhai, S. Realtourist–a study of augmenting human-human and human-computer dialogue with eye-gaze overlay. *Human-Computer Interaction-INTERACT 2005* (2005), 767–780.
11. Qvarfordt, P., and Zhai, S. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2005), 221–230.
12. Roy, D., and Mukherjee, N. Visual context driven semantic priming of speech recognition and understanding. *Computer Speech and Language* (2003).
13. Roy, D. K. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language 16*, 3 (2002), 353–385.
14. Weide., H. The CMU Pronunciation Dictionary, release 0.6. Carnegie Mellon University, 1998.
15. Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.