The AAAI-17 Workshop on
Human-Aware Artificial Intelligence
WS-17-10

Using Co-Captured Face, Gaze and Verbal Reactions to Images of Varying Emotional Content for Analysis and Semantic Alignment

Aliya Gangji
Muhlenberg College
ag249083@muhlenberg.edu

Trevor Walden
Rochester Institute of Technology
tjw5866@rit.edu

Preethi Vaidyanathan, Emily Prud'hommeaux, Reynold Bailey, and Cecilia O. Alm
Rochester Institute of Technology
{pxv1621, emilypx, rjbvcs, coagla}@rit.edu

Abstract

Analyzing different modalities of expression can provide insights into the ways that humans interpret, label, and react to images. Such insights have the potential not only to advance our understanding of how humans coordinate these expressive modalities but also to enhance existing methodologies for common AI tasks such as image annotation and classification. We conducted an experiment that co-captured the facial expressions, eye movements, and spoken language data that observers produce while examining images of varying emotional content and responding to description-oriented vs. affect-oriented questions about those images. We analyzed the facial expressions produced by the observers in order to determine the connection between those expressions and an image's emotional content. We also explored the relationship between the valence of an image and the verbal responses to that image, and how that relationship relates to the nature of the prompt, using low-level lexical features and more complex affective features extracted from the observers' verbal responses. Finally, in order to integrate this multimodal data, we extended an existing bitext alignment framework to create meaningful pairings between narrated observations about images and the image regions indicated by eye movement data. The resulting annotations of image regions with words from observers' responses demonstrate the potential of bitext alignment for multimodal data integration and, from an application perspective, for annotation of open-domain images. In addition, we found that while responses to affect-oriented questions appear useful for image understanding, their holistic nature seems less helpful for image region annotation.

Introduction

Despite advances in machine image analysis, humans continue to outperform machines in describing and labeling images. One significant obstacle to achieving human-level performance is the sizable gap between the low-level recognition a computer can achieve and the high-level concepts humans apply when analyzing images (Zhang, Islam, and Lu 2012).

One way to gain insight into the complex processes underlying human image understanding is to capture data from multiple expressive modalities, including eye movements, facial expression, and spoken language, from human observers. Despite the ease of collecting such data in noninvasive ways, approaches for analyzing and integrating such data streams for the purposes of improving machine image analysis remains a research challenge and continues to be relatively underexplored.

Recent work has applied a bitext alignment framework, using algorithms originally developed for machine translation, to meaningfully integrate the language and eye movements of dermatologists as they reason over medical images in order to automatically label image regions with the words the dermatologists used when describing the images (Vaidyanathan et al. 2016). We extend this framework to analyze the responses of non-expert observers of non-medical, open-domain images with varying degrees of neutral and positive emotional content, for which verbal and visual responses can be expected to vary further.

We additionally prompt the observers to provide not only image descriptions but also to comment on their own feelings toward the images and to speculate about the photographer's intent. We gather data from a third modality that helps convey human emotion – facial expressions – to supplement and complement the language and gaze data streams. We find that the facial expressions used by our subjects vary considerably according to the emotional valence of the image being observed. It also seems that observers' linguistic expression is influenced by the emotional content of an image and the questions they are asked to consider when examining an image. Finally, our results indicate that the bitext alignment method, whose utility was demonstrated in previous work on medical images, can be used to annotate regions in non-medical images of varying emotional valence.

By including not only the results of these more subjective image analysis tasks but also the modality of facial expression, we can gather new and interesting data to increase understanding of how humans reason about images. From an application perspective, such insights may be useful in retrieval tasks in which a user seeks images not according to specific objects or events depicted but instead according to the image's perceived or expressed emotional content.

Related Work

In the last two decades, computer vision and automatic image annotation techniques have had success in capturing information about natural scenes, identifying faces, and semantic labeling (Viola and Jones 2004; Zhang, Islam, and Lu 2012; Kong et al. 2014; Yatskar, Zettlemoyer, and Farhadi 2016). Other recent work has focused on generating image descriptions and image captions (Karpathy and Fei-Fei 2014; Vinyals et al. 2014). Computer vision applications continue, however, to have limitations in deciphering high-level semantics. There is a large gap between the restrictive, low-level algorithms created to detect objects in an image and the richness and complexities of human understanding (Müller et al. 2004; Vinyals et al. 2014). This work aims to annotate image semantics by integrating co-captured eye movements and verbal descriptions.

Eye movements have been shown to be closely tied to how humans process language (Meyer, Sleiderink, and Levelt 1998; Griffin and Bock 2000; van der Meulen 2003; Griffin 2004; Vaidyanathan et al. 2012). One finding is that speakers tend to fixate objects before naming them. Griffin and Bock (2000) observed that for simple sentences such as 'pass me the salt', the lag was about 1 sec. Eye movements have also been used in the past to help with object recognition and for image region annotation (Torralba et al. 2006; Clarke, Coco, and Keller 2013; Yun et al. 2013; Vaidyanathan et al. 2016; 2015b).

Several researchers have investigated multimodal integration and proposed methods such as mutual information and machine translation to integrate the two modalities (Roy 2000; Yu and Ballard 2004). Duygulu et al. (2002) used machine translation for the purpose of object recognition.

While facial expressions have been used in the past as a tool for better understanding interactions between computers and humans (Busso et al. 2004; Jaimes and Sebe 2007), we apply insights from facial expression in another context; for analyzing emotional vs. neutral image content.

Data Collection

Two of the authors jointly categorized images as having positive or neutral valence, with 15 images per category (30 images total). For instance, images portraying love, joy, or excitement, such as the images in the top row of Figure 2 depicting two lion cubs hugging and a family playing in a park, were regarded positive. The content in neutral-valence images tended to have little or no expressive emotional context, exemplified in the bottom row of Figure 2 by photos of the exterior of a Scandinavian hotel and of a tidy living room. We selected the images from public domain sources, and pilot tests verified that there was emotional and visual diversity among them.

We collected eye movement data, audio recordings of spoken descriptions, and video recordings of facial expressions for 20 (13 male, 7 female) college-aged participants as they viewed and described images. The images were shown to each subject on a 22-inch LCD monitor (1680x1050 pixels) approximately 65 cm from the subject, as shown in Figure 1. Eye movement data was collected using a SensoMotoric Instruments (SMI) RED250 eye-tracking device running at 250 Hz. SMI's BeGaze software V3.0.0X with default parameters and a velocity-based (I-VT) algorithm was used to detect eye-tracking events, including fixations. A Logitech QuickCam Pro 9000 recorded both audio of the subjects' descriptions and video of their faces.

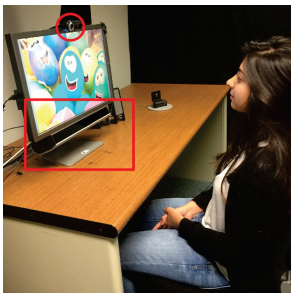


Figure 1: Experimental setup: Observer sits in front of the LCD screen viewing images while responding aloud to description-oriented and affect-oriented questions about the images. Eye movements are captured using a video-based eye-tracker (indicated by a red rectangle) and speech and facial expressions are captured using a microphone-equipped webcam (indicated by a red circle).

(a) Image P1 (b) Image P2

(c) Image N2 (d) Image N7

Figure 2: Examples of images presented to observers. Images (a-b) categorized as positive (P); (c-d) as neutral (N).

The data collection followed a prompt and response format. Participants were each shown a randomized series of 90 prompt-image pairs from 30 images with three prompts per image. Participants were given the opportunity for short breaks after every 20 prompt-image pairs, with a mandatory two minute break after 40 pairs.

Participants were asked to respond to all three prompts for each image. Prompts were presented as text on a gray slide before the image they pertained to. For each image, subjects were prompted to (1) describe what was shown in the image (DESCRIBE); (2) discuss how the image made them feel (FEEL); and (3) speculate about the feeling that the photographer hoped to capture (PHOTOGRAPHER). After data col-

lection, the observers were given a survey asking about their experience during the experiment.

Eye-tracking data from 5 participants was removed from further analysis due to excessive eye-tracking loss. One subject's verbal response to an image-pair prompt was also excluded due to a technical issue that resulted in audio loss. Facial expression data is valid only when the subject is facing the camera. Periods of facial track loss, for example if a subject looks away from the camera, were not included in the subsequent analysis.

As a first step, the following reports on computational analysis considering 11 images. In automated preprocessing, observers audio recordings were machine-transcribed using the IBM Speech-to-Text API, an Automatic Speech Recognition (ASR) system. A subset of the ASR output (specifically, the participants' responses to the DESCRIBE and PHOTOGRAPHER prompts for three images (P1, P2, N7) were manually corrected using the speech annotation and analysis tool Praat (Boersma 2001). These corrected transcriptions enabled us to directly evaluate the accuracy of the ASR output and to explore the utility of substituting automatically generated transcriptions for careful manual transcriptions. Word error rates (WER) ranged from 22% to 40%, depending on the image and prompt. We discuss the interaction between transcription accuracy and alignment performance in the results section.

The video recordings of subjects' faces were cut into segments corresponding to each prompt and run through Affectiva's Affdex SDK, which uses facial landmarks to report on emotions conveyed. Affdex SDK outputs, for each time interval, values representing the degree to which 7 different emotions are detected from the facial actions. In our analysis, we considered dominant emotion per time interval for aggregate analysis. We include five emotions (fear, sadness, joy, anger, and surprise) and exclude disgust and contempt, whose recognition appeared less reliable and often linked to resting faces.

Bitext Alignment

People's eye movements when looking at an image are a good indicator of the prominent regions in that image. Verbal descriptions of images captured in tandem with eye movement data can provide labels to assign to those prominent regions. Associating the words with regions, however, is not a trivial problem, as research has shown that speech and eye movements cannot be assumed to have a one-to-one or fixed-delay temporal correspondence (Meyer, Sleiderink, and Levelt 1998). Here, we follow the bitext alignment approach first proposed by Vaidyanathan et al. (2015a; 2016) for finding the alignment between eye movements over an image and the words in a spoken description of that image.

Bitext alignment was originally developed in the statistical machine translation (SMT) research community to align words in one language with their corresponding translations in another. Languages show great variation in word order (e.g., English *the most handsome man* vs. Spanish *el hombre más guapo*), making simple temporal or linear alignments insufficient for this task. The bitext alignment methods generally used in SMT rely on unsupervised expectation maximization algorithms, in which word co-occurrence statistics are gathered from large parallel corpora, comparing sentences in one language to their translations in another language. As more sentence pairs are processed, evidence accumulates for translation equivalences between pairs of words that tend to occur together in parallel sentences. The multimodal scenario recasts the alignment problem as finding associations between the regions of an image (*visual units* or VUs) and the words used to describe that image (*linguistic units* or LUs).

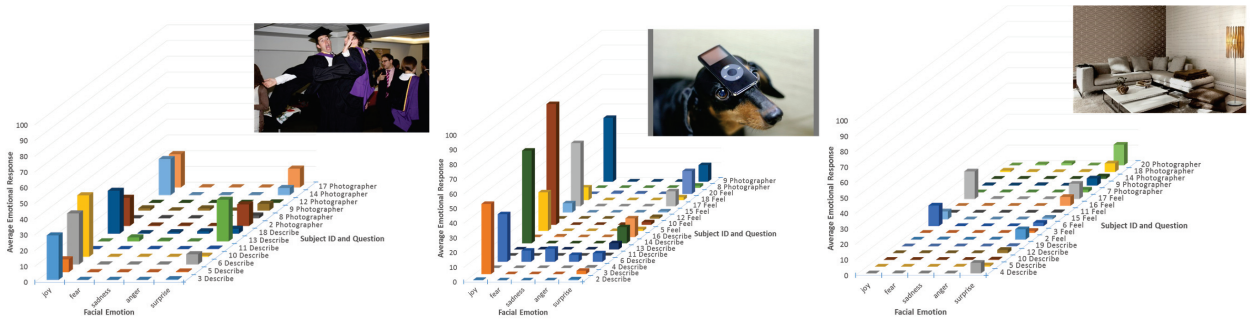
Linguistic Units Following Vaidyanathan et al. (2016), we focus on nouns and adjectives in observers' narratives. We used the Berkeley Parser (Petrov and Klein 2007) to identify these and explored four different types of LUs to use as input to the aligner:

1. Uncorrected DESCRIBE LUs: ASR output of responses to the DESCRIBE prompt containing nouns and adjectives produced by the user.
2. Shortened DESCRIBE LUs: ASR output of responses to the DESCRIBE prompt filtered to contain only the top 15 most frequently uttered potential LUs by all observers for an image, occurring at least 5 times.
3. Corrected DESCRIBE LUs: ASR output of responses to the DESCRIBE prompt, with ASR errors manually corrected. Available for images N7, P1, and P2.

Visual Units Each fixation identified by the eye-tracker is coded as a pair of *x,y* coordinates and a duration. As in Vaidyanathan et al. (2016), we used mean-shift clustering (Santella and DeCarlo 2004) to assign all of the fixations on an image produced by all speakers into clusters in order to identify a set of "regions of interest" for use as VUs. Each fixation in a participant's series of fixations can then be assigned to a cluster based on its spatial position. Any clusters outside the image region are discarded in the process. We used this cluster information to obtain a linearly ordered sequence of VUs for input to the aligner.

Reference Alignments In order to evaluate the performance of the bitext alignment system trained on these inputs of VUs and LUs, we manually created *reference alignments* associating the words (LUs) used by the observers with regions of the image produced by the mean-shift clustering algorithm (VUs). Because we are interested in extracting labels for the most salient parts in an image, we filtered our reference alignments to include only the most frequently used nouns and adjectives produced by participants for a particular image, equal to or exceeding a threshold mc. From that list, we then selected up to the k most frequent. Also, LUs that could not be human aligned were ignored. For DESCRIPTION responses $k = 15$ and $mc = 5$, while for PHOTOGRAPHER responses, $k = 15$ and $mc = 3$. These values were determined via trial and error; determining the optimal values for these parameters is left for future work.

Performing Alignment The expectation maximization (EM) algorithm used to generate alignments between LUs



(a) P14 (b) P15 (c) N1

Figure 3: Average facial response for five emotions, per speaker/prompt, differ for positive (a-b) and neutral (c) images.

and VUs requires a relatively large parallel corpus. We have only a small set of descriptions for each image, which is insufficient to train an alignment model using EM. Following Vaidyanathan et al. (2016), we used a 5-second sliding window over the training corpus to create additional sentences to include in the parallel corpus. Another problem we encountered was that the sequences of VUs were substantially longer than the sequences of LUs. To balance the number of VUs and LUs, we merged contiguous identical units for both sequences, and then selected only the VUs with the longest fixations. For alignment, we then used the Berkeley Aligner, a widely used word alignment package for machine translation (Liang, Taskar, and Klein 2006), with default parameter settings.

Results

Facial Expression Analysis Not surprisingly, we found that subjects tended to respond more strongly with facial expressions associated with joy when presented with the images identified as containing positive emotional content, as shown in Figure 3 (panels a-b). When viewing an image with neutral content (panel c), the participants showed little detectable emotion in their facial expressions, with a slight preference for joy and surprise. Some variation is expected as there is a fuzzy distinction between neutral and emotional image content. Individualized variation by prompt type also occurred. In combination, these results indicate that observers' facial actions do change according to the emotional content of the images they are viewing and the prompt they are asked to consider when responding.

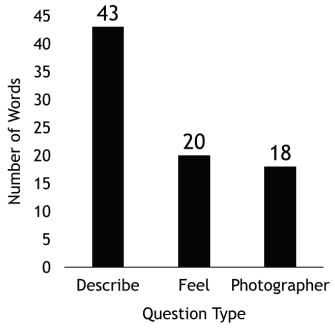
Linguistic Analysis We also observed an interaction between the prompt under consideration and the participants' linguistic choices, using TextBlob for analysis. Observers evoked different linguistic patterns when responding to DESCRIBE, PHOTOGRAPHER, and FEEL prompts. The average word count for DESCRIPTION responses roughly doubled compared to the average word count for FEEL and PHOTOGRAPHER responses; see Figure 4 (panel a). The ratio of adjectives (adjectives/total words) was higher for FEEL and PHOTOGRAPHER responses than DESCRIPTION responses (panel b). Lexical richness (unique word/total words) was higher for PHOTOGRAPHER and FEEL responses than DESCRIBE responses (panel c). The linguistic metrics indicate that responses to FEEL and PHOTOGRAPHER prompts differ in character, as compared to DESCRIBE responses. Particularly, responses to FEEL and PHOTOGRAPHER, which are affect-oriented and more interpretative, result in smaller output but more diverse vocabulary, and lexical choices that differ in nature. For instance, the most common filtered and manually corrected words for DESCRIBE responses for the lion image (P1) were *cubs, grass, hugging, lion*, and *lions*, but for PHOTOGRAPHER, the responses were *animals, cubs, cute, family, happiness, lion, love*, and *moment*.

Prompts that are affect-oriented vs. description-oriented appear to be complementary, drawing out/eliciting different kinds of information valuable for image analysis and understanding. Also, since alignment for image region annotation depends on repetition of LUs (with VUs), the responses to the DESCRIBE prompt are more suitable for it.

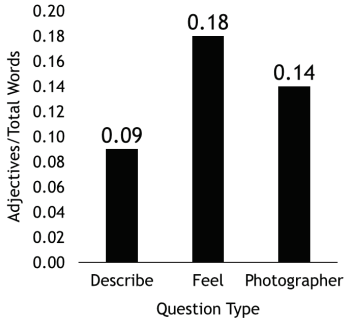
Positive and neutral valence of images appeared mirrored in observers' linguistic choices. On average, when responding to FEEL and PHOTOGRAPHER responses, observers polarity (most positive 1 to most negative -1) was higher for positive images as expected, see Figure 5 (panel a), although for the DESCRIBE prompt, observers polarity was close to 0 for positive images. For images with neutral valence, the latter rather held for all three prompts. Similarly, subjectivity, which captures degree of expressed attitude or viewpoint (0-1 range), was higher for FEEL responses and then highest for positive image valence (panel b). These results reveal a connection between image valence, prompt type, and affective lexical usage in responses.

Alignment The temporal baseline is an alignment that assumes that an observer utters the word corresponding to a region at the moment his/her eyes fixate on it. For alignment, the number of VUs was set to be equal to the number of available LUs. Since the LUs involved constraining input to the aligner, VUs whose fixation length was insufficient for inclusion were removed. This process occurred before metrics were calculated, and non-selected VUs were excluded when computing performance.

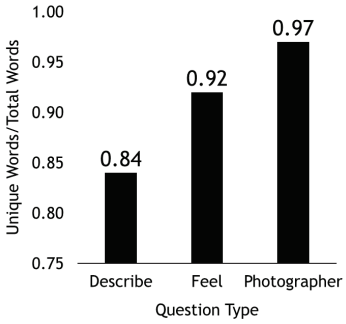
Alignment with uncorrected ASR input exhibits marked



(a) Mean word count.

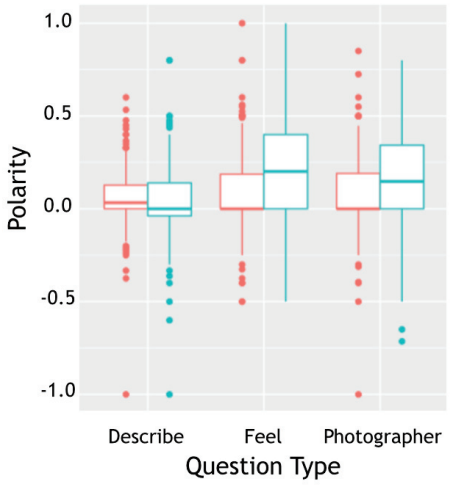


(b) Mean adjective ratio.

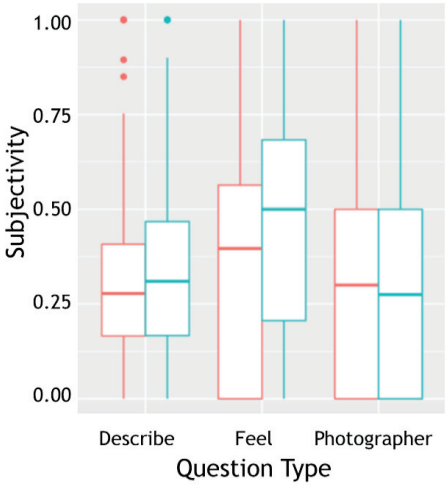


(c) Mean lexical richness.

Figure 4: DESCRIBE responses (left) elicit most words but less lexical richness and fewer adjectives than other prompts.



(a) Mean polarity.



(b) Mean subjectivity.

Figure 5: Image valence (red = N, turquoise = P) and prompt type yield differences in lexical polarity and subjective tone.

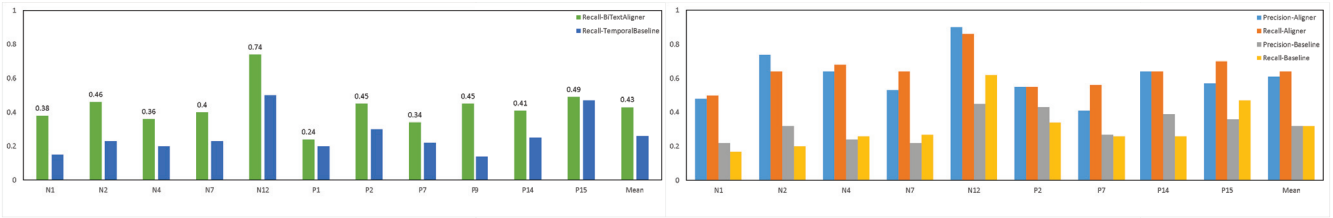
improvement over baseline temporal alignment, as in prior application of this method (Vaidyanathan et al. 2016). Figure 6 (panel a) shows that recall improved for most images when aligning with *Uncorrected* DESCRIBE *LUs*. Improvement was more pronounced for some images, such as N1 and P2, than for others, most notably P1 and P15. Due to thresholding LUs in reference alignments, precision does not apply. When aligning *Shortened* DESCRIBE *LUs* (panel b) we report recall and precision, similarly showing mostly clear improvement over the baseline.

In addition, comparison of alignment with vs. without manual correction of ASR input suggests varying impact. For example, the percent of improved recall over baseline rose marginally from 17% to 19% for image N7, and from 15% to 19% for P2. In contrast for image P1, this improvement was large, from 4% to 35%. Tentatively, the variability relates to the word error rate (WER) of the ASR. P1 had higher WER than N7 and P2.

Lastly, an example of resulting image region annotations (with uncorrected vs. manually corrected ASR) for P1 is shown in Figure 7; as noted such correction could improve the result.

Conclusions and Future Work

Our analysis of observers' facial expressions reactions suggests that faces can provide an indication of images' emo-



(a) Uncorrected DESCRIBE LUs (recall comparison)

(b) Shortened DESCRIBE LUs (precision and recall comparison)

Figure 6: Left: Aligner outperformed temporal baseline for DESCRIBE prompt. Right: Improved also for shortened LUs.

(a) Aligned annotation with automated ASR

(b) Aligned annotation with corrected ASR

Figure 7: Each box of words is located at the center of its fixation cluster. Magenta: Words aligned to that cluster using our method. Yellow: Words in our reference alignment, not aligned by our method. Results are better for P1 after manual correction of ASR (ASR word error rate for P1 was high).

tional valence. Also, the linguistic analysis of responses indicate both that images' valence and the prompt type matter. The FEEL and PHOTOGRAPHER prompts elicit answers of distinct linguistic character. Thus, they are less suitable for image region annotation than the DESCRIBE prompt, but nonetheless valuable for holistic image understanding; this may be useful for captioning or assigning labels to images as a whole (vs. annotating regions). The impact of prompts on response output deserves further study to tease out their respective usefulness for image analysis and understanding.

Limiting the LUs and VUs should be explored further. Determining the best method of constraining the LU input is left to future work. We also noticed that some VUs had very brief duration; a form of popularity-based short-listing process could also be applied to VUs. Investigation is needed to find optimal values for k and mc; mc may be calculated according to the proportion of observers who mentioned a word rather than by overall frequency of occurrence.

This work used ASR to automate the transcription process. The quality of the ASR output varied depending on the content of the image, the prompt, and the participants' individual speech characteristics. While using manually corrected LUs resulted in some improvement in alignment, the degree of improvement depended on the initial word error rate of the ASR output. Our qualitative comparison of alignment results tentatively suggests that, at least for some images, they may not always justify the human labor required for correcting ASR transcripts. Incorporating ASR into the alignment framework has the potential to improve the practicality and utility of our system for automated labeling and annotation of open-domain images; further study is needed.

This work provides insight into the complexities of human understanding with implications for a range of challenging AI problems including image retrieval, image annotation, and scene classification. A natural extension of this study would be to consider multimodal reactions to emotional content and semantic alignment in videos.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10):341–345.

Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C. M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; and Narayanan, S. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, 205–211. New York, NY, USA: ACM.

Clarke, A. D.; Coco, M. I.; and Keller, F. 2013. The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology* 4.

Duygulu, P.; Barnard, K.; de Freitas, J. F.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision–ECCV*. Springer. 97–112.

Griffin, Z. M., and Bock, K. 2000. What the eyes say about speaking. *Psychological science* 11(4):274–279.

Griffin, Z. M. 2004. Why look? Reasons for eye movements related to language production. *The interface of language, vision, and action: Eye movements and the visual world* 213–247.

Jaimes, A., and Sebe, N. 2007. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding* 108(12):116 – 134. Special Issue on Vision for Human-Computer Interaction.

Karpathy, A., and Fei-Fei, L. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Kong, C.; Lin, D.; Bansal, M.; Urtasun, R.; and Fidler, S. 2014. What are you talking about? Text-to-image coreference. In *CVPR 2014*.

Liang, P.; Taskar, B.; and Klein, D. 2006. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, 104–111. Stroudsburg, PA, USA: Association for Computational Linguistics.

Meyer, A. S.; Sleiderink, A. M.; and Levelt, W. J. 1998. Viewing and naming objects: Eye movements during noun phrase production. *Cognition* 66(2):B25–B33.

Müller, H.; Michoux, N.; Bandon, D.; and Geissbuhler, A. 2004. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *Intl. Journal of Medical Informatics* 73(1):1–23.

Petrov, S., and Klein, D. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*, volume 7, 404–411.

Roy, D. 2000. Integration of speech and vision using mutual information. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 6, 2369–2372. IEEE.

Santella, A., and DeCarlo, D. 2004. Robust clustering of eye movement recordings for quantification of visual interest. In *Proc. of Symposium on Eye Tracking Research & Applications*, ETRA '04, 27–34. New York, NY, USA: ACM.

Torralba, A.; Oliva, A.; Castelhano, M. S.; and Henderson, J. M. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review* 113(4):766.

Vaidyanathan, P.; Pelz, J. B.; McCoy, W.; Calvelli, C.; Alm, C. O.; Shi, P.; and Haake, A. R. 2012. Visualinguistic approach to medical image understanding. In *AMIA*.

Vaidyanathan, P.; Prudhommeaux, E.; Alm, C. O.; Pelz, J. B.; and Haake, A. R. 2015a. Alignment of eye movements and spoken language for semantic image understanding. *IWCS 2015* 76.

Vaidyanathan, P.; Prudhommeaux, E.; Alm, C.; and Pelz, J. B. 2015b. Computational integration of human vision and natural language through bitext alignment. In *Workshop on Vision and Language, Empirical Methods for Natural Language Processing*.

Vaidyanathan, P.; Prud'hommeaux, E.; Pelz, J. B.; Alm, C. O.; and Haake, A. R. 2016. Fusing eye movements and observer narratives for expert-driven image-region annotations. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16, 27–34. New York, NY, USA: ACM.

van der Meulen, F. F. 2003. Coordination of eye gaze and speech in sentence production. *Trends in Linguistics Studies and Monographs* 152:39–64.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

Viola, P., and Jones, M. J. 2004. Robust real-time face detection. *Intl. Journal of Computer Vision* 57(2):137–154.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR 2016*.

Yu, C., and Ballard, D. H. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)* 1(1):57–80.

Yun, K.; Peng, Y.; Adeli, H.; Berg, T.; Samaras, D.; and Zelinsky, G. 2013. Specifying the relationships between objects, gaze, and descriptions for scene understanding. *Journal of Vision* 13(9):1309–1309.

Zhang, D.; Islam, M. M.; and Lu, G. 2012. A review on automatic image annotation techniques. *Pattern Recogn.* 45(1):346–362.