Alignment of Eye Movements and Spoken Language for Semantic Image Understanding

Preethi Vaidyanathan, Emily Prud'hommeaux, Cecilia O. Alm,
Jeff B. Pelz, and Anne R. Haake
Rochester Institute of Technology
(pxv1621|emilypx|coagla|jbppph|arhics)@rit.edu

Abstract

Extracting meaning from images is a challenging task that has generated much interest in recent years. In domains such as medicine, image understanding requires special expertise. Experts' eye movements can act as pointers to important image regions, while their accompanying spoken language descriptions, informed by their knowledge and experience, call attention to the concepts and features associated with those regions. In this paper, we apply an unsupervised alignment technique, widely used in machine translation to align parallel corpora, to align observers' eye movements with the verbal narrations they produce while examining an image. The resulting alignments can then be used to create a database of low-level image features and high-level semantic annotations corresponding to perceptually important image regions. Such a database can in turn be used to automatically annotate new images. Initial results demonstrate the feasibility of a framework that draws on recognized bitext alignment algorithms for performing unsupervised automatic semantic annotation of image regions. Planned enhancements to the methods are also discussed.

1 Introduction

The ability to identify and describe the important regions of an image is useful for a range of vision-based reasoning tasks. When expert observers communicate the outcome of vision-based reasoning, spoken language is the most natural and convenient instrument for conveying their understanding. This paper reports on a novel approach for semantically annotating important regions of an image with natural language descriptors. The proposed method builds on prior work in machine translation for bitext alignment (Vogel et al., 1996; Liang et al., 2006) but differs in the insight that such methods can be applied to align multimodal visual-linguistic data. Using these methods, we report on initial steps to integrate eye movements and transcribed spoken narratives elicited from expert observers inspecting medical images.

Our work relies on the fact that observers' eye movements over an image reveal what they consider to be the important image regions, their relation to one another, and their relation to the image inspection objectives. Observers' co-captured narrations about the image naturally express relevant meaning and, especially in expert domains, special knowledge and experience that guide vision-based problem-solving and decision-making. Despite being co-captured, precise time synchronization between eye movement and narrations cannot be assumed (Vaidyanathan et al., 2013). Therefore, techniques are needed to integrate the visual data with the linguistic data. When meaningfully aligned, such visual-linguistic data can be used to annotate important image regions with appropriate lexical concepts.

We treat the problem of integrating the two data streams as analogous to the alignment of a parallel corpus, or bitext, in machine translation, in which the words of a sentence in one language are aligned to their corresponding translations in another language. For our problem, eye movements on images are considered to be the visual language comprising visual units of analysis, while the transcribed narratives contain the linguistic units of analysis. Previous work has investigated the association of words with pictures, words with videos, and words with objects and image regions (Forsyth et al., 2009; Kuznetsova et al., 2013; Kong et al., 2014). But the combination of perceptual information (via eye movements) and

Proceedings of the 11th International Conference on Computational Semantics, pages 76–81,
London, UK, April 15-17 2015.

more naturally obtained conceptual information (via narratives) will greatly improve the understanding of the semantics of an image, allowing image regions that are relevant for an image inspection task to be annotated with meaningful linguistic descriptors. In this ongoing work, we use dermatological images as a test case to learn the alignments between the two types of units of analysis. We primarily seek to determine whether a bitext alignment approach can be used to align multimodal data consisting of eye movements and transcribed narratives.

2 Prior Work

Automatic semantic annotation of images and image regions is an important but highly challenging problem for researchers in computer vision and image-based applications. Algorithms employing low-level image features have succeeded somewhat in capturing the statistics of natural scenes, identifying faces, and recognizing objects (Zhang et al., 2012). Some researchers have proposed sophisticated models for semantic annotation through object recognition that while successful on certain types of images either failed to capture the semantics of the image and the relations between regions or fared poorly in expert domains such as medicine (Müller et al., 2004; Vinyals et al., 2014). Multiple approaches have been developed for generating image captions (Kuznetsova et al., 2013; Kong et al., 2014). The method we propose here differs from these earlier approaches in its use of spoken language descriptions and eye movement data. In addition to being more natural than approaches such as labeling of images by drawing or eliciting image descriptions on Mechanical Turk, our approach has the potential to provide more information (Vaidyanathan et al., 2013).

Empirical experiments have shown that eye movements are closely time-locked with human language processing (Ferreira and Tanenhaus, 2007). Roy (2000) proposed a technique for integrating vision and language elicited from infants using a mutual information model. Although useful in infant-directed interactions it is unlikely that this would translate successfully to complex scenarios containing multiple objects/concepts such as viewing medical images. Researchers have used techniques such as Latent Dirichlet Allocation to address the multimodal integration problem (Li and Wang, 2003). Machine translation approaches have been used with image features to recognize and annotate objects in scenes, to automatically match words to the corresponding pictures, and to describe scenes and generate linguistic descriptions of image regions (Duygulu et al., 2002; Berg et al., 2004; Yu and Ballard, 2004). While prior work supports the feasibility of integrating the visual and linguistic modalities, it also leaves open the question of whether multimodal integration is feasible in more complex scenarios such as medical image annotation. Moreover, eye movements in complex scenarios are usually not considered in this work. We make a key contribution by focusing on integrating experts' gaze data, known to encode their perceptual knowledge for vision-based problem solving in their domain (Li et al., 2013).

3 Data Collection

Twenty-nine dermatologists were eye tracked and audio recorded while they inspected and described 29 images of dermatological conditions. A Sensomotoric Instruments (SMI) eye tracking apparatus and a TASCAM audio recording equipment were used. The participants were asked to describe the image to the experimenter, using a modified version of the Master-Apprentice method (Beyer and Holtzblatt, 1997), a data elicitation methodology from human-computer interaction for eliciting rich spoken descriptions (Womack et al., 2012). Certain data were missing or excluded for quality reasons, leaving a multimodal corpus consisting of 26 observers and 29 images.

4 Aligning eye movements and transcribed narratives

The goal of this research is to develop a method for aligning an observer's eye movements over an image (the visual units) with his spoken description of that image (the linguistic units). Our first step is

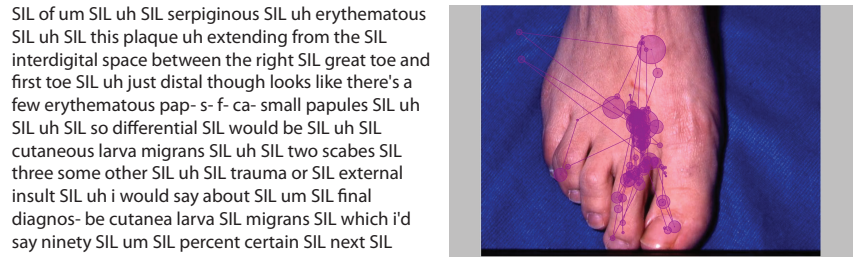

Figure 1: **A multimodal data example**. *Left*: Transcribed narrative with speech disfluencies, silent (SIL) pauses, and no utterance boundaries. *Right*: Eye movements overlaid on the corresponding image. The circles represent locations where the observer gazed, with the size of the circle reflecting gaze duration. The lines connecting the circles indicate changes of gaze location. Image courtesy:Logical Images

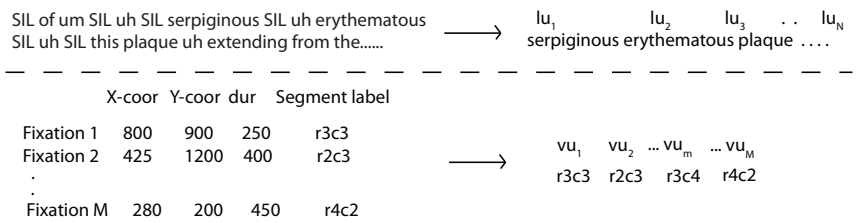

Figure 2: The top panel shows an excerpt of a transcribed narrative and the resulting linguistic units after filtering based on parsing evidence. The bottom panel shows eye movement data and the resulting gaze-filtered image regions or visual units. The linguistic and visual units jointly act as input to the Berkeley aligner. Linear order of the units is maintained and reflected in the parallel data windows.

therefore to extract these visual and linguistic units from the eye-tracking data and audio recordings. The audio recordings of the observers were transcribed verbatim, as shown in Figure 1. Most dermatological concepts present in an image tend to involve either noun phrases or adjectives. Accordingly, the linguistic units for alignment are nouns and adjectives, which we identify using the following process. Utterance boundaries are added and silent and filled pauses are removed before parsing the data with the Berkeley parser (Petrov and Klein, 2007). From the parsed output, tokens in noun phrases and adjective phrases are extracted. The linear order of the data is maintained. Following the extraction process, we filter out regular stopwords along with a set of tokens reflecting the data collection scenario (e.g., *differential, certainty, diagnosis*).[1] Names of diagnoses (*psoriasis, basal cell carcinoma*, etc.) are also removed since such words do not correspond to any particular image region but instead reflect the disease that an image depicts as a whole. The resulting filtered and linearly ordered sequences of linguistic units serve as one of the inputs to the bitext alignment algorithm. An example is shown in Figure 2.

The eye movement data consists of fixation locations indicating where in the image observers gazed, as shown in Figure 1. We use fixation locations in conjunction with image regions to obtain the visual units. The images are of size 1680 x 1050 pixels. We divide each image into a grid of 5 rows (r) and 5 columns (c). Each cell in the grid is associated with a label that encodes the row and column number for that cell, for example *r3c9, r4c12*. The fixations of an observer are overlaid on this grid, and each fixation is labeled according to the grid cell it falls within. In this way, we obtain a linearly ordered sequence of visual units consisting of fixated image regions, encoded using the grid labels, for the other input to the alignment algorithm. An example is shown in Figure 2.

In machine translation, an alignment model is trained on a parallel corpus of sentences rendered in two different languages. In our case, each observers' entire narrative and entire set of fixations is one training sentence pair. For each image, this would yield a corpus of only 26 parallel sentences, which would be insufficient to generate a reliable alignment model. To increase the amount of training data, we

[1] Observers were supposed to provide a differential, a final diagnosis, and to indicate diagnostic certainty.

Image	Precision	Recall	F-measure
1	0.71	0.65	0.68
2	0.65	0.44	0.52
3	0.28	0.32	0.30
4	0.44	0.36	0.40
5	0.24	0.32	0.28
Overall	0.42	0.40	0.41

Table 1: Alignment precision, recall, and F-measure for 5 images. Higher values indicate better alignment.

use a moving window of 5 seconds, extracting linguistic and visual units within each sliding timeframe and adding that pair of unit sequences as a "sentence" pair in the corpus. The time window is applied incrementally over a narrative, resulting in a much larger parallel corpus. In order to ensure that the two sequences in a given sentence pair are of roughly equal length, we merge contiguous identical visual units. (For example *r2c3, r3c3, r3c3* is converted into *r2c3, r3c3*.) We then randomly select visual units, still maintaining the linear order, such that the number of visual units is equal to number of linguistic units within that time window. We leave optimization of the parameters relating to grid segmentation, time window size, and visual unit selection for future work.

The pairs of sequences of linguistic and visual units, organized as described above, serve as a parallel corpus for training the aligner. We use the Berkeley aligner, recognized for its accuracy and flexibility in testing and training alignment models (Liang et al., 2006). Following standard approaches to word alignment for machine translation, the Berkeley aligner uses expectation maximization to learn an alignment model in an unsupervised manner. To generate gold standard reference alignments for evaluating the aligner, a researcher with reasonable knowledge of the regions in the image and the vocabulary used by the observers to describe the images manually produced alignments for 5 of the 29 images. Future work will involve more images.

5 Results

Word alignment output for building machine translation models is typically formatted as a set of paired indices for each input sentence pair being aligned. Each index pair represents an alignment between the words at the locations indicated by the indices. The standard metric used to evaluate word alignment for machine translation is the word alignment error rate (AER), which compares the index pairs output by the aligner to the manually derived reference index pairs. Knowing the locations of the words being aligned is necessary for subsequent steps in the MT pipeline, such as building the phrase table. In contrast, our end goal is to learn which words are being used to describe the various regions in the image, independently of when the words were uttered and when the fixations occured. Thus, rather than evaluate the accuracy of the index pairs output by the aligner, we instead evaluate our aligner in terms of how accurately it identifies the correct correspondences between words and regions indicated by those index pairs.

In Table 1, we report the precision, recall, and F-measure of our aligner output, calculated as described above, for the 5 images for which we produced manual reference alignments. We see that although the performance of the aligner varies depending on the image, we achieve strong performance values in some cases and reasonable performance overall. Figure 3 shows one of the images in our evaluation set overlaid with the 5x5 grid. In addition, the figure includes a randomly selected subset of words from the observers' narratives overlaid on the regions with which they were associated by the alignment model. Many of the words were correctly aligned with the regions they depict.

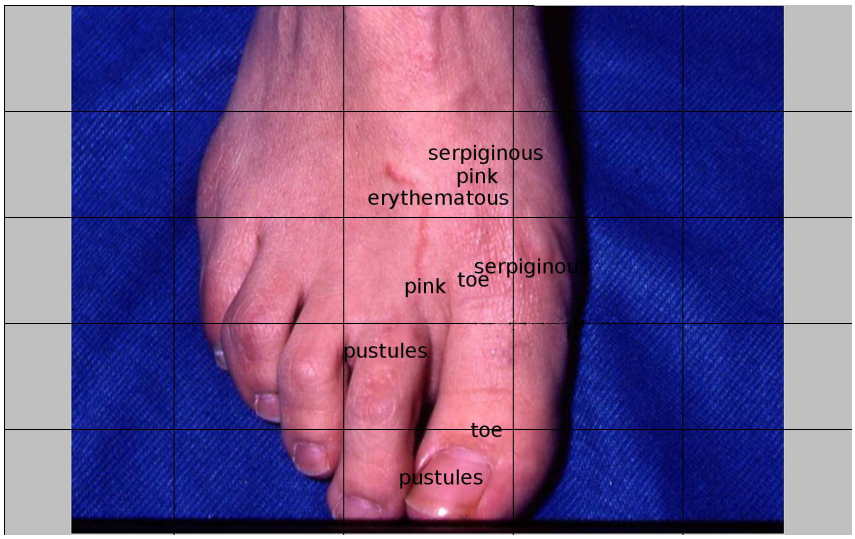

Figure 3: Some occurrences of a set of randomly selected lingusitic units (word tokens) overlaid on the visual units (image regions) with which they were aligned by the alignment model.

6 Discussion and Conclusion

The reasonably high precision, recall, and F-measure alignment results for images 1 and 2 indicate that a bitext alignment model can be used to align eye movements and transcribed narratives. Visual inspection of the aligner output might explain why the results are not as high for images 3, 4, and 5. In particular, some abstract linguistic units that are not physically present in the image (e.g., *unusual, thought*) but tagged as noun and adjective tokens by the parser, are included in the input to the aligner. These abstract units are not, however, included in the manually derived reference alignments, thereby lowering the precision of the alignment output for these images. We also note that the reference alignments generated by the researcher could be different from those generated by a dermatology expert.

The next phase in the development of our system for automated semantic annotation of images will be to use these alignments to map the low-level image features of these images and the image regions to the lexical items and semantic concepts with which they are associated via alignment. A model built on these mapping could be used to generate semantic annotations of previously unseen images. In addition, further collection of visual-linguistic information could be made more efficient using automatic speech recognition adapted to the medical or dermatological domain.

There are many improvements that can be made to the existing system, particularly in the way that the various parameter values were selected. The size of the time window used to expand the parallel corpus, the image segmentation approach, and the selection of the visual units all can be tuned in order to optimize alignment performance. In addition, our method for extracting the linguistic units relies on parsing output, which could be improved by training the parser on spoken language data from the biomedical domain. In future work, we also intend to use a more sophisticated image segmentation algorithm together with a medical ontology such as the Unified Medical Language System (UMLS) to learn even more meaningful relations between important image regions and lexical concepts.

In summary, the work presented here introduces a new approach for obtaining semantic annotations for images by creatively integrating visual and linguistic information. These preliminary results highlight the potential of adapting existing NLP methods to problems involving multimodal data.

References

Berg, T. L., A. C. Berg, J. Edwards, and D. Forsyth (2004). Who's in the picture? *Advances in neural information processing systems 17*, 137–144.

Beyer, H. and K. Holtzblatt (1997). *Contextual design: Defining customer-centered systems*. San Diego: Elsevier.

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

Ferreira, F. and M. K. Tanenhaus (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language 57*(4), 455–459.

Forsyth, D. A., T. Berg, C. O. Alm, A. Farhadi, J. Hockenmaier, N. Loeff, and G. Wang (2009). Words and pictures: Categories, modifiers, depiction, and iconography. In S. J. Dickinson (Ed.), *Object Categorization: Computer and Human Vision Perspectives*. Cambridge: Cambridge University Press.

Kong, C., D. Lin, M. Bansal, R. Urtasun, and S. Fidler (2014). What are you talking about? Text-to-image coreference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3558–3565.

Kuznetsova, P., V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi (2013). Generalizing image captions for image-text parallel corpus. In *Proceedings of ACL*, pp. 790–796.

Li, J. and J. Z. Wang (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*(9), 1075–1088.

Li, R., P. Shi, and A. R. Haake (2013). Image understanding from experts' eyes by modeling perceptual skill of diagnostic reasoning processes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2187–2194.

Liang, P., B. Taskar, and D. Klein (2006). Alignment by agreement. In *Proceedings of NAACL-HLT*, pp. 104–111.

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

Petrov, S. and D. Klein (2007). Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*, pp. 404–411.

Roy, D. (2000). Integration of speech and vision using mutual information. In *Proceedings of ICASSP*, pp. 2369–2372.

Vaidyanathan, P., J. B. Pelz, C. O. Alm, C. Calvelli, P. Shi, and A. R. Haake (2013). Integration of eye movements and spoken description for medical image understanding. In *Proceedings of the 17th European Conference on Eye Movements*.

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

Vogel, S., H. Ney, and C. Tillmann (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational Linguistics-Volume 2*, pp. 836–841.

Womack, K., W. McCoy, C. O. Alm, C. Calvelli, J. B. Pelz, P. Shi, and A. R. Haake (2012). Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pp. 1–9.

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

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