Computational Integration of Human Vision and Natural Language through Bitext Alignment

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

Abstract

Multimodal integration of visual and linguistic data is a longstanding but crucial challenge for modeling human understanding. We propose a framework that uses an unsupervised bitext alignment method to integrate visual and linguistic data. We present an empirical study of the various parameters of the framework. Our results exceed baselines using both exact and delayed temporal correspondence. The resulting alignments can be used for image classification and retrieval.

1 Introduction

Modeling and characterizing human expertise is a major bottleneck in advancing image-based application systems. We propose a framework for integrating experts' eye movements and verbal narrations as they examine and describe images in order to understand images semantically. Eye movements can act as pointers to important image regions, while the co-captured descriptions provide conceptual labels associated with those regions.

Although successful when applied to scenic images in controlled experiments, many multimodal integration techniques do not transfer directly to scenarios requiring domain-specific expertise. Our approach is inspired by Yu and Ballard (2004), who combine NLP methods with eye movements to generate linguistic descriptions of videos, and Forsyth et al. (2009), who use image features to match words to the corresponding pictures. We expand here on earlier work (Vaidyanathan et al., 2015) exploring multimodal integration in medical image annotation.

Because an exact temporal match between the visual and verbal modalities cannot be assumed (Griffin, 2013), our framework integrates the two modalities without enforcing strict temporal correspondence. We use a bitext word alignment algorithm, originally developed for word alignment in machine translation, to align an expert's fixations on an image with the words in that expert's description of that image. The resulting alignments are then used to annotate image regions with corresponding conceptual labels, which in turn may aid image labeling and captioning applications. In this paper we discuss the parameters of our framework and their effects on alignment accuracy.

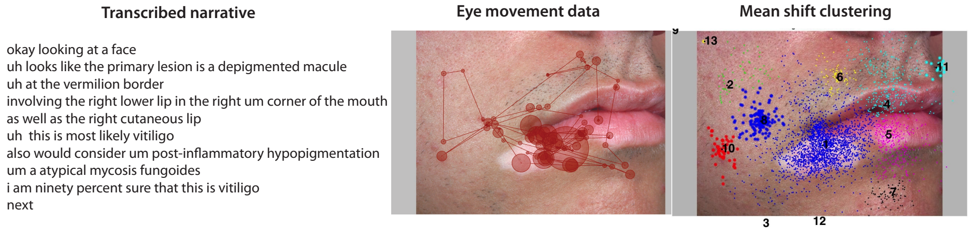
2 Data and Method

We eye tracked and voice recorded 26 dermatologists as they examined and described 29 dermatological images. From the narrations, we extract nouns and adjectives to create a temporally ordered set of linguistic units. To obtain the visual units, we cluster the fixations for all observers using mean shift clustering with a bandwidth (72 pixels) approximating the foveal size (Santella and DeCarlo, 2004). For each observer, we use these clusters to produce a temporally ordered sequence of visual units. Figure 1 shows a manually transcribed narrative, a scanpath for an observer, and clusters of fixations from all observers.

Prior research has established that there is a temporal lag between fixations and concept mentions (Griffin, 2013). Our method aligns visual and linguistic units without explicit assumptions about their temporal relationships. This is analogous to translating one language into another where the structural characteristics and word order of the two languages may be different. In our multimodal scenario, the observer's narrative description and fixations on an image represent a training pair. To create a sufficiently large parallel corpus, we use a 5-second sliding window over the pairs and add the linguistic and visual units within each window as a "sentence" to the corpus.

The sequences of visual units are substantially longer than the sequences of linguistic units. In order to balance the sequence lengths, we select

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 4–5,
Lisbon, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.



okay looking at a face
uh looks like the primary lesion is a depigmented macule
uh at the vermilion border
involving the right lower lip in the right um corner of the mouth
as well as the right cutaneous lip
uh this is most likely vitiligo
also would consider um post-inflammatory hypopigmentation
um a atypical mycosis fungoides
i am ninety percent sure that this is vitiligo
next

Figure 1: **Example of a multimodal data pair**. *Center*: Circle and circle size represent observer gaze location and duration, respectively. *Right*: Clusters shown with colors and/or shape and numerical labels.

	P (SD)	R (SD)	F1 (SD)
1-sec. delay	0.38 (0.1)	0.44 (0.17)	0.39 (0.1)
bitext alignment	0.45 (0.1)	0.56 (0.16)	0.49 (0.1)

Table 1: Comparison of performance for the 1-second delay baseline and our alignment method.

visual units in two ways, both preserving temporal order. In one method, the fixations are selected at random. In the other, the fixations are ranked and selected according to their duration.

We use the Berkeley aligner (Liang et al., 2006), an EM-based word aligner known for high accuracy and adaptability. The aligner is run on each visual-linguistic parallel corpus (one for each image), with the posterior threshold for decoding set to 0.1, a value empirically determined using a data subset. The resulting alignments for each corpus are evaluated against a set of reference alignments produced manually by an investigator experienced in analyzing dermatological images.

3 Results and Conclusions

We test the model on pairs of full narratives and fixation sequences. The alignment results are compared with two temporal baselines. One baseline assumes that an observer utters the word corresponding to a region at the moment the eyes fixate on that region. The second baseline assumes that there is a one-second delay (Griffin, 2013) between a fixation and the utterance of the word corresponding to that region.

Our alignment method yields strong performance in comparison to both baselines. As shown in Table 1, we achieve 7%, 10%, and 12% absolute improvement over the baselines in precision, F-measure, and recall, respectively. The results hold on a per-image basis as well, with the alignment approach yielding higher recall in all 29 images, higher F-measure in 28 images, and higher precision in 24 images. Using fixation length to select the visual units substantially improves the performance in comparison to the random selection process. Neither the size of the sliding window nor the ratio of visual to linguistic units affected alignment performance.

Both methods perform well on images with solitary lesions, and performance generally decreases as the number of lesions increases. Interestingly, the largest improvement of our aligner over the baseline occurs in images with multiple lesions, suggesting that a fixed temporal correspondence is particularly unlikely in more complex images.

In future work, we plan to use image segmentation algorithms to extract image features and a medical ontology to discover more complex relationships between image regions and semantic concepts. In addition, we will explore methods of alignment with soft temporal constraints to better model the relationship the two modalities.

References

P. Liang et al. 2006. Alignment by agreement. In *Proceedings of NAACL-HLT*, pages 104–111.

D. Forsyth et al. 2009. Words and pictures: Categories, modifiers, depiction, and iconography. In S. Dickinson, editor, *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press, Cambridge.

P. Vaidyanathan et al. 2015. Alignment of eye movements and spoken language for semantic image understanding. *IWCS 2015*, page 76.

Z. Griffin. 2013. Why look? Reasons for eye movements related to language production. In J. Henderson and F. Ferreira, editors, *The interface of language, vision, and action: Eye movements and the visual world*. Psychology Press, New York.

A. Santella and D. DeCarlo. 2004. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of ETRA*, pages 27–34.

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