The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Multimodal Alignment for Affective Content

Nikita Haduong
Indiana University
nhaduong@indiana.edu

David Nester
Eastern Mennonite University
david.nester@emu.edu

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

Abstract

Humans routinely extract important information from images and videos, relying on their gaze. In contrast, computational systems still have difficulty annotating important visual information in a human-like manner, in part because human gaze is often not included in the modeling process. Human input is also particularly relevant for processing and interpreting affective visual information. To address this challenge, we captured human gaze, spoken language, and facial expressions simultaneously in an experiment with visual stimuli characterized by subjective and affective content. Observers described the content of complex emotional images and videos depicting positive and negative scenarios and also their feelings about the imagery being viewed. We explore patterns of these modalities, for example by comparing the affective nature of participant-elicited linguistic tokens with image valence. Additionally, we expand a framework for generating automatic alignments between the gaze and spoken language modalities for visual annotation of images. Multimodal alignment is challenging due to their varying temporal offset. We explore alignment robustness when images have affective content and whether image valence influences alignment results. We also study if word frequency-based filtering impacts results, with both the unfiltered and filtered scenarios performing better than baseline comparisons, and with filtering resulting in a substantial decrease in alignment error rate. We provide visualizations of the resulting annotations from multimodal alignment. This work has implications for areas such as image understanding, media accessibility, and multimodal data fusion.

Introduction

Human eye movements have rarely been leveraged in the understanding and annotation of visual content, although gaze is a core component for modeling human-aware computer vision. A particular challenge for image understanding methods is to ensure that they equally apply when visual data have an affective character, the interpretation of which is highly subjective. To bridge the performance gap, there is a need to understand how human observers process visual imagery based on their patterns of eye movements, verbal descriptions, and potentially also their facial expressions. We analyze these modalities and, for static images, also address the challenge of meaningfully fusing image observers' verbal and gaze-based human reactions. We apply this visual-linguistic integration to the useful task of automatically labeling regions of affective images with naturally elicited lexical items. Specifically, a multimodal bitext alignment approach addresses the challenges presented by the temporal disconnect between an observer's gaze and his or her spoken language description in this mapping process. This method does not require hand-labeled data and instead uses gaze to map words from co-collected spontaneous spoken descriptions to viewed image regions (see Figure 11).

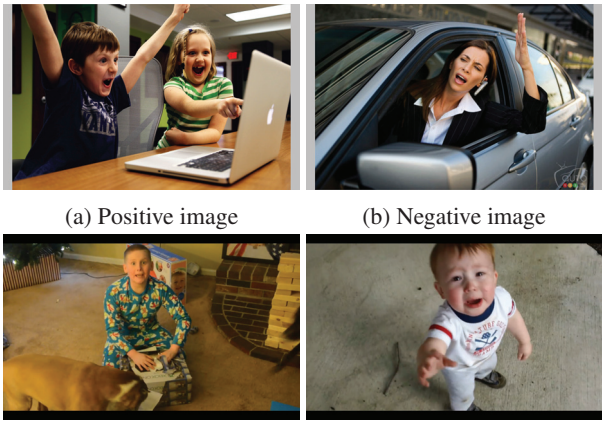
This is a continuation and expansion of our previous work (Vaidyanathan et al. 2016) on aligning dermatologists' gaze data with their spoken descriptions of medical images. The framework, which leverages multimodal human data, was also used by Gangji et al. (2017) to annotate regions of static images depicting either neutral or positive content and to make preliminary observations about observers' language use and facial expressions.

Our contributions include: (1) Analyzing observers' language use and facial expressions as they react to complex visual stimuli that are either negative or positive. (2) Extending and verifying the effectiveness of the multimodal alignment framework for automated image region annotation with complex emotional static images. (3) Assessing the impact of image valence and low frequency word filtering in static image region annotation.

Related Work

The explored framework attempts to annotate regions in images based on observers' multimodal reactions to them. A number of previous researchers have investigated how gaze and language interact. Meyer et al. (1998) examined the movement of eyes during image naming and recognition tasks for obscured images with recognizable objects. Although humans can still successfully recognize obscured objects, computer vision systems continue to struggle with this task. Griffin et al. (2000; 2004) observed that there is around a one-second lag between when we look at objects and when we name them, demonstrating that the relationship between gaze and object naming is not perfectly temporally coordinated and hence is more complex than expected. Kienzle et al. (2007) developed a spatiotemporal interest point detector by training a neural net on observers' eye movements over

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Positive image (b) Negative image

(c) Positive video (d) Negative video

Figure 1: Examples of imagery presented to subjects in our experiment. Stimuli consisted of 20 images and 20 videos, half depicting positive content and half negative content.

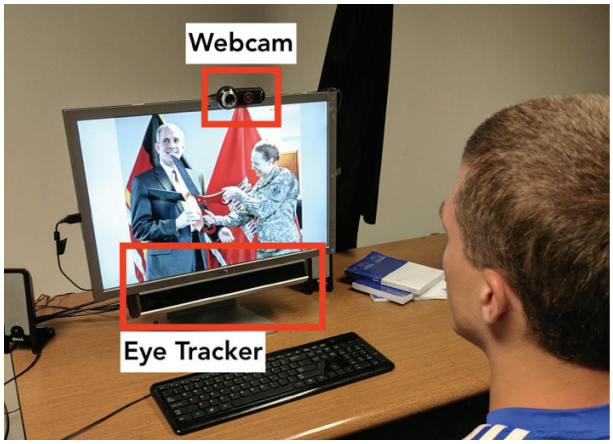


Figure 2: Experiment setup. An observer sits in front of a monitor displaying stimuli and is recorded with a lapel microphone while answering questions about the content and affective impact of the image. A webcam records facial expressions, and an eye tracker captures eye movements.

short black-and-white video clips to predict where humans are most likely to direct their gaze. Mahapatra et al. (2008) looked at how motion in videos attracts the gaze of an observer to the moving region. These works point to interesting questions about the similarities and differences between images and videos, and the value of considering each type of visual stimulus.

Several studies have investigated how people associate objects in visual stimuli with their corresponding names using gaze data (Vaidyanathan et al. 2016; Clarke, Coco, and Keller 2013). Yu and Ballard (2004a; 2004b) combined gaze and head motion with object names to annotate objects and categorize scenes in video stimuli involving simple events such as printing. Qu and Chai (2008) used word-gaze alignment for a single synthetic visual scene of a room with furniture, providing insights into whether gaze can aid in the acquisition of new words. Bojanowski (2015) and Naim et al. (2014) aligned videos with parallel natural language descriptions to annotate the video as events occurred. Bojanowski et al. (2015) learned the beginning and end of events based on the narratives. Fang et al. (2009) investigated how attention indicated by spoken language corresponds to the intensity of gaze fixations. Their preliminary results suggest interesting alignments between language and gaze intensity for measuring attention. Overall, these studies convey the potential for mapping eye movements to verbalized concepts, which our framework leverages in addressing the task of image region annotation.

Data Collection

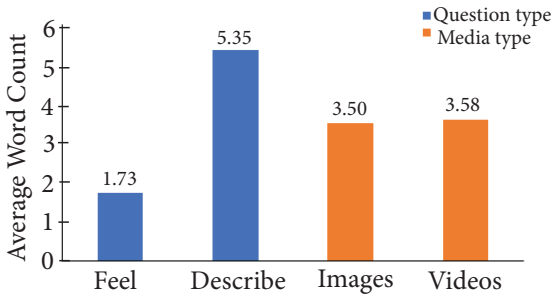
Two authors jointly collected 20 images and 20 videos (8-10 seconds in length) containing complex content to serve as stimuli. Of the 20 stimuli for each modality, half contained positive emotional content and half contained negative emotional content. These images and videos were acquired from the MSCOCO database (Lin et al. 2014), a content-annotated database for computer vision; LIRIS-ACCEDE (Baveye et al. 2015), a valence annotated database of videos for affective computing research; and Google image and video search with a Creative Commons filter. All were available for use under the Creative Commons license. Examples of some of these images and representative video frames are shown in Figure 1. All authors reviewed and unanimously agreed upon the positive or negative valence of each stimulus. Sound was removed from the videos to limit the number of sensory inputs to the observer. Videos and images primarily contained expressive humans or animals in order to elicit emotional reactions in observers. Two questions accompanied the stimuli, reflecting two levels of subjectivity:

1. How does the following image/video make you feel? (FEEL)
2. Describe what is shown in the following image/video. (DESCRIBE)

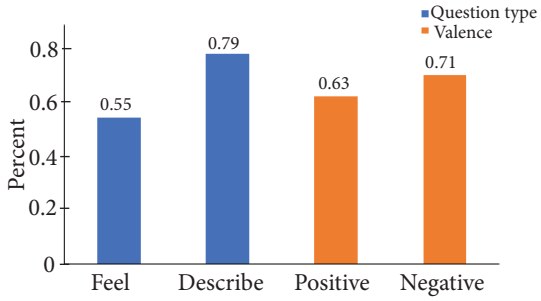
Observers completed a pre-survey for gathering demographic data and a post-survey for rating the design of the IRB-approved experiment. The experiment consisted of four parts, with all stimuli randomized per observer: 1) 20 images accompanied by FEEL, 2) 20 videos accompanied by FEEL, 3) 20 images accompanied by DESCRIBE, and 4) 20 videos accompanied by DESCRIBE. Observers received 20 USD in compensation. Native English speakers were recruited to ensure quality automated transcription using cost-effective automatic speech recognition (ASR).

Eye movements, facial expressions, and spoken description data were collected using a Sensomotoric Instruments RED250 desktop eye-tracker, a Logitech QuickCam Pro 9000, and a TASCAM DR-100MKIII saving in WAV audio format with a 96kHz sample rate, respectively. Audio was transcribed using IBM's Bluemix cloud-based ASR technology[1]. The experimental setup with data collection instru-

[1]https://www.ibm.com/watson/services/speech-to-text/



(a) Average number of words by question type and media type



(b) Percent nouns in spoken descriptions by question type and valence

Figure 3: The average number of considered words differs for FEEL and DESCRIBE questions (blue) as shown in 3a, but not by video or image stimuli (orange). The percentage of nouns in spoken descriptions is shown in 3b by question type (blue) and image valence (orange), and there is a close to 10% increase from positive to negative stimuli.

ments is shown in Figure 2. One subject was excluded due to data loss. The multimodal data for 21 subjects was used in the subsequent experiment.

Data Analysis

We report on the analysis of linguistic data, facial expressions, and the integration of linguistic data and lexical semantic scoring.

Linguistic Analysis

The linguistic analysis considers unique nouns and adjectives. Figure 3a shows the average number of nouns and adjectives by question type and visual stimulus type. Many participants uttered a single word in response to the FEEL question and said more for the DESCRIBE question, whereas the two different stimulus types (image vs. video) behaved similarly. Figure 3b shows that more nouns are used for the DESCRIBE question and also occur more often when responding to negative stimuli. This could reflect greater lexical diversity for negative stimuli, as suggested by Figure 4; while the word *happy* dominates the positive cloud, the negative cloud spreads out over more words.

In order to explore the semantic relationships between these words, we used gensim's[2] implementation of word2vec to extract 300-dimensional word embedding vectors for a subset of these words from the pre-built Google News word2vec model. We then projected these vectors down to two dimensions with the scikit-learn[3] library's implementation of t-SNE in order to generate plots illustrating these semantic relationships (see Figure 5). We examine only those words occurring at least 5 times overall for the FEEL question (red circle) and their corresponding DESCRIBE responses (green circle). Stimuli associated with plotted words are embedded in the subfigure panels. Figure 5a and Figure 5b (top row) show that the affectively similar FEEL words *angry* and *annoyed* from the DESCRIBE responses are located near each other in the semantic space, capturing the negative depiction of being in a traffic jam. In contrast, the antonyms *happy* and *sad* (bottom row) are used with multiple images and associate with distinct situational denotations, reflecting positive and negative valence, respectively.

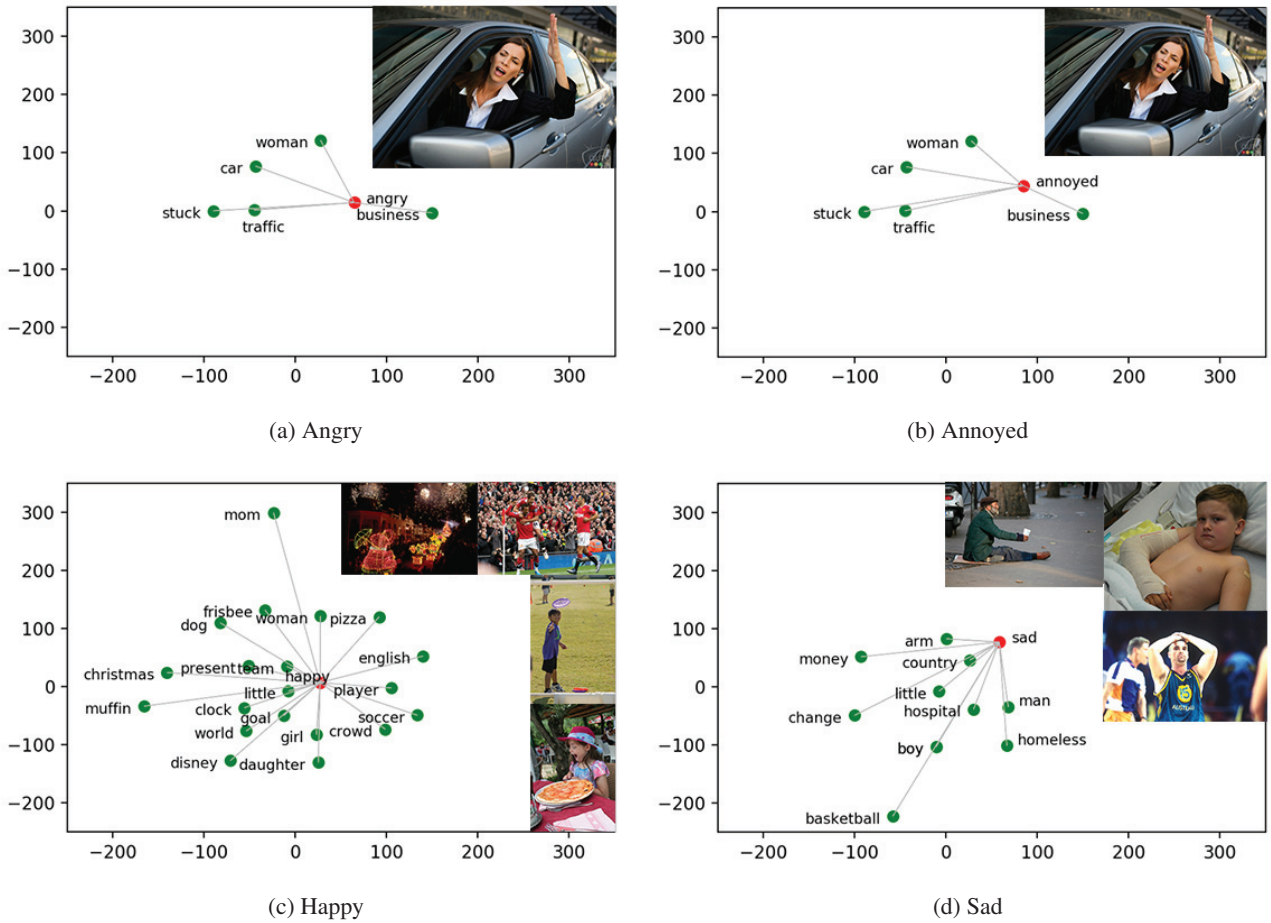
(a) Positive visual stimuli

(b) Negative visual stimuli

Figure 4: Word clouds for 125 most frequent words in positive and negative stimuli. The vocabulary diversity in negative stimuli is greater than in positive stimuli, where there is a predominant usage of the term *happy*.

[2]https://radimrehurek.com/gensim/

[3]http://scikit-learn.org/



(a) Angry (b) Annoyed

(c) Happy (d) Sad

Figure 5: Visualization of word embeddings with frequency of 5 or more across stimuli, including the images with which these words were used. While antonymic emotion words *happy* and *sad* occur with distinct images and denotative situational associations, near synonyms *angry* and *annoyed* are associated with the same image and share a lexical semantic space.

Facial Expression Analysis

We use the Affectiva SDK[4] to analyze the facial expression data. Affectiva returns a confidence rating of the presence of 7 emotions (contempt, surprise, anger, sadness, disgust, fear, joy) in the 0-100 range and valence in the -100 to 100 range. Due to common false positives for DISGUST, and low incidence for most other emotions, we focus on two positive emotions that are clearly analyzed from facial expressions: JOY and SURPRISE. As expected, there is more presence of facial expressions conveying JOY or SURPRISE for positive stimuli, with longer spans in videos as compared to images. There was a higher detection rate and longer duration rate for SURPRISE compared to JOY. Facial expressions are not used for the multimodal alignment process.

Emotional Linguistic Analysis

We also scored the emotional nature of the two most frequent words per stimulus using two affective lexicons (Mohammad and Kiritchenko 2015; Mohammad and Turney 2013; Mohammad 2012; Mohammad and Turney 2010). As seen in Figure 6, positive visual stimuli elicited lexical items scoring high for JOY and SURPRISE, whereas negative visual stimuli elicited lexical items scoring higher for SADNESS and ANGER. FEAR and DISGUST had similar scores for either stimulus valence. Interesting examples are in Table 1 and include *nervous* elicited for both positive and negative stimuli but scoring high for FEAR. The word *soldier*, also soring high for FEAR, was used to describe a positive video stimulus showing the event of a soldier returning home.

In Figure 7, valence was determined using the ANEW lexicon (Bradley and Lang 1999), with 70% coverage, and the extended ANEW lexicon WARRINER (Warriner, Kuperman, and Brysbert 2013), with 95% coverage. ANEW and WARRINER resulted in similar scores, with positive stimuli eliciting more positive words and vice versa.

Multimodal Alignment with Affective Images

The Berkeley Aligner (Liang, Taskar, and Klein 2006) is used in machine translation to align parallel sets of sen-

[4]https://developer.affectiva.com/

	Disg.	Fear
daughter	0.48	
dog	0.64	
nervous		**1.66**
soldier		0.83

(a) Positive stimuli

	Disg.	Fear
woman	0.51	
wood	0.57	
nervous		**1.66**
anxious		1.57

(b) Negative stimuli

Table 1: The top-2 DISGUST and FEAR words, calculated with the NRC Hashtag Emotion Lexicon scores for positive and negative visual stimuli. The presence of *nervous* with positive stimuli made the score for FEAR in positive stimuli comparable with that of negative stimuli as seen in Figure 6.

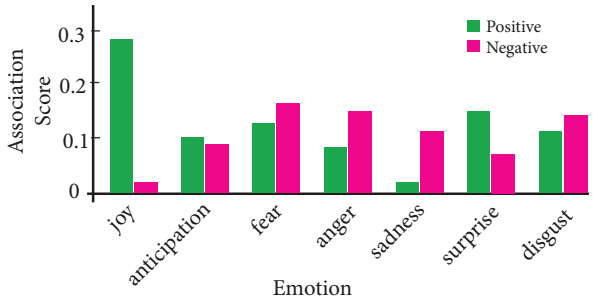


Figure 6: NRC Hashtag Lexicon scores for the 2 most frequent words in positive and negative stimuli. JOY and SADNESS have similar differences between positive and negative visual stimuli, as do ANGER and SADNESS.

tences in two languages. The aligner identifies words in each language that frequently occur together in order to map meaning correspondences across the languages. For our multimodal alignment framework, processed linguistic units (LUs) and gaze-based visual units (VUs) represent our parallel bitext. The process identifies words that correspond to fixated regions of interest in images.

Processing Data for Multimodal Alignment

For the multimodal alignment facial expressions and FEEL questions were not included. Before the aligner can be used, the multimodal gaze and language data must be processed into parallel "sentences." Fixations are clustered using the mean shift clustering algorithm (Santella and DeCarlo 2004). This estimates the regions of interest based on the fixations and not information from the image itself. An example is shown in Figure 8.

For an image, the VUs for each observer are determined based on the clusters containing their fixations. We remove the fixations that occur before the observer begins their description and after the observer finishes their description. We extract the start time, end time, and duration of each fixation using the Sensomotoric Instruments's analysis software BeGaze. This yields a list of clusters that were fixated upon while the observer was speaking with timestamps to show when they moved to a new cluster.

The spoken narratives generated by the observers were

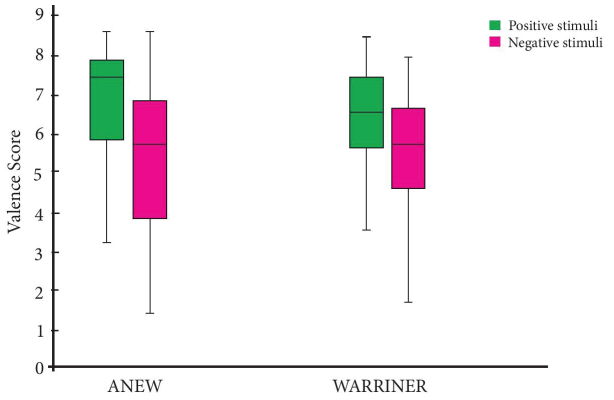


Figure 7: The overall valence of words compared in positive and negative stimuli. Words with positive valence were used more with positive stimuli, and vice versa.

Figure 8: Clustered fixation regions on a negative image.

automatically transcribed using ASR. Words other than nouns and adjectives were removed. The remaining words were the LUs that could be aligned to the VUs.

For the multimodal mapping process, we use the Berkeley Aligner, which requires a large amount of parallel data to perform well. Because we only had 21 narratives per image, we used a sliding window of 5 seconds to create additional training data, following Vaidyanathan et al.'s method (2016).

Alignment after Filtering

For each image, the alignment was performed with two sets of LUs, unfiltered and filtered. The unfiltered set contained all the words spoken by all observers. The filtered set contained only those words with frequency greater than 1, which resulted in removal of words that were uttered only once or were likely transcription errors.

Reference Alignments and Evaluation

The alignment was compared to a baseline alignment created by aligning temporally coinciding VUs and LUs, as done by Vaidyanathan et al. (2013). As aligning gaze with spoken language is a relatively new task, there are no existing benchmarks. This baseline uses an intuitively reasonable

Figure 9: A higher value for AER indicates weaker performance. In all image cases, our framework performed better with the filtered lexicon.

assumption: that fixating on and naming parts of an image occur simultaneously.

In addition, reference "gold standard" alignments, used to evaluate the performance of the framework, were created by manually identifying the LUs associated with each VU. We report the precision, recall, and alignment error rate (AER) of our alignment and the baseline alignment relative to these manual reference alignments using $Precision = \frac{|A \cap S|}{|A|}$, $Recall = \frac{|A \cap S|}{|S|}$, and $AER = 1 - \frac{|A \cap S|}{|A+S|}$, where A is the list of VU-LU pairs produced by the aligner and S is the list of VU-LU pairs in our reference set. Precision and recall are the ratio of pairs correctly aligned in the aligner output set and the reference set respectively. For both metrics, a higher number means better performance. AER is the ratio of incorrect or unmatched pairs that occurred. Lower AER corresponds with better alignment performance.

	Filt.	Unfil.	F-baseline	U-baseline
Precision	**0.53**	0.39	0.32	0.29
Recall	**0.67**	0.59	0.30	0.29
AER	**0.42**	0.53	0.69	0.72

Table 2: Average alignnent results using filtered (F) and unfiltered (U) word lists show that the multimodal alignment framework outperforms the baseline, and that filtering of rare words improves performance.

Alignment Results and Discussion

The average performance of each evaluation metric of the alignment before and after filtering for both our framework and the baseline is shown in Table 2. In all cases, filtering the word lists improves performance. Comparing the metrics before and after filtering, the baseline shows only minor improvement across the three metrics while our alignment framework substantially improves across all three metrics. Additionally, in the filtered case, compared to its baseline, the framework shows a more than 25% decrease in AER and more than 20% increase in precision. The AER for individual images is shown in Figure 9, indicating varying performance by image. Figure 10 compares the AER for positive (green) and negative (pink) images for the filtered and unfiltered scenarios, respectively. AER is stable across valence. In contrast, filtering is an important factor in improving the framework's performance irrespective of the valence. Factors such as complexity, subjectivity, automated transcription errors, and the general-domain nature of images, can increase vocabulary size. Removing rare words ensures more image-relevant verbal data and aids alignment.

Figure 11 demonstrates the annotation of image regions based on alignments, contrasting the use of filtered and unfiltered lists of words. For Image 9 (top row), many missed words (yellow) could be considered holistic words that characterize the entire image rather than a specific region, such as *army, military*, and *soldiers*. The aligner was able to correctly label several regions such as *sergeant, serious*, and *face*. The aligner had the worst performance on Image 6 (bottom row). Many factors could have caused this poor performance including a smaller number of VUs and the proximity of distinct regions. When comparing the filtered images and unfiltered images there is a noticeable difference in the sizes of the word lists and in the number of missed (yellow) words. Some incorrectly aligned words were correctly aligned after the list was reduced, such as *army* on the left hand side of the image. This gives further evidence of the benefits of constraining data for the alignment framework.

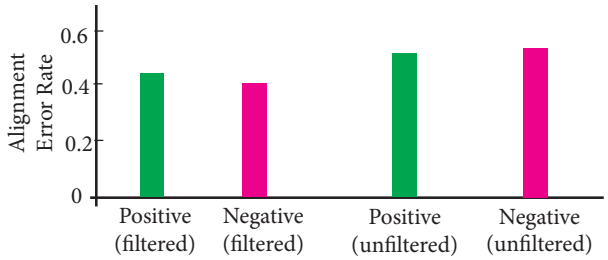


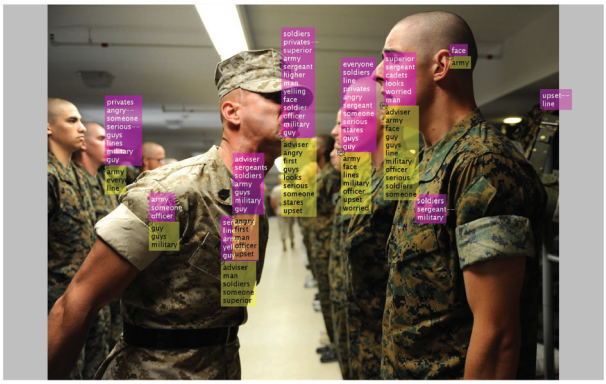
Figure 10: Average AER across positive (green) and negative (pink) images with and without filtering.

Conclusions and Future Work

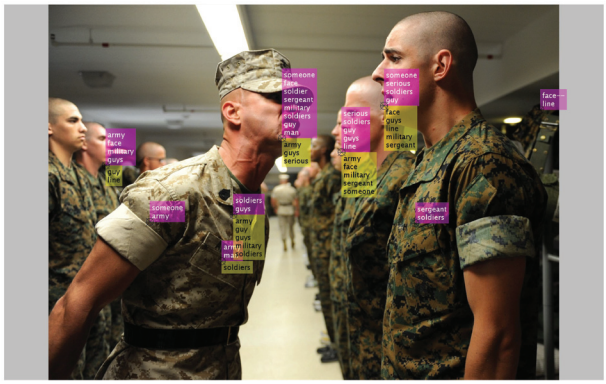
With our data elicitation method, we find that participants produce distinct subjective and affective lexical items for region annotation according to image valence. Mapping these lexical items to a two-dimensional semantic space, we observe that the configuration of the semantic space varies according to the valence of the image. We also see that the level of subjectivity of the question type has an impact on the linguistic data elicited. Both the word clouds and emotional valence lexica show that positive images generally reflect more positive linguistic valence, and negative images reflect more negative valence.

Moreover, this work shows that the multimodal bitext alignment method can be used for image region annotation of complex images that are affective and subjective. More examination is needed to understand what characterizes images that align well. One factor that might influence the multimodal alignment is the number of fixation clusters (VUs) produced by the mean-shift clustering algorithm. Some images had many clusters, while the fixation clusters found for other images were potentially spurious. Some images also had larger clusters, which might more easily have multiple words associated with them. Similarly, some images had more words to align. Systematic investigation could reveal how these factors affect alignment and optimize how data is used.

While our dataset is modest, there are no large-scale datasets of co-acquired eye movements, facial expressions, and spoken language of people examining affective visual



(a) Image 9 - negative: Unfiltered (AER - 0.52)



(b) Image 9 - negative: Filtered (AER - 0.38)

(c) Image 6 - positive: Unfiltered (AER - 0.70)

(d) Image 6 - positive: Filtered (AER - 0.66)

Figure 11: Example of labeled regions for negative (Image 9) and positive (Image 6) static images with both an unfiltered (*left*) and filtered word list (*right*). Words in magenta indicate correctly aligned words. Words in magenta with dashes are incorrectly aligned. Words in yellow should have been aligned for that region but were not returned by the aligner.

imagery. High-fidelity eye-tracking requires equipment and careful calibration, making leveraging crowdsourcing platforms infeasible. It is not yet practical to collect a large dataset to train a deep network. Indeed, a merit of our approach is that it can do well with less data, which still characterizes many crucial multimodal problems.

Our framework can be extended to annotate videos by incorporating temporal information into the clustering process. Further experimentation with facial expression analysis could help annotate the expected dominant emotional impact an image may have on its observers. Increased availability of human-elicited data will aid deeper understanding of the impact of complex, affective image content on gaze and spoken language behaviors.

Acknowledgements

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

Baveye, Y.; Dellandrea, E.; Chamaret, C.; and Chen, L. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6(1):43–55.

Bojanowski, P.; Lagugie, R.; Grave, E.; Bach, F. R.; Laptev, I.; Ponce, J.; and Schmid, C. 2015. Weakly-supervised alignment of video with text. *CoRR* abs/1505.06027.

Bradley, M., and Lang, P. 1999. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Clarke, A. D.; Coco, M. I.; and Keller, F. 2013. The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology* 4:927.

Fang, R.; Chai, J. Y.; and Ferreira, F. 2009. Between linguistic attention and gaze fixations inmultimodal conversational interfaces. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, 143–150. New York, NY, USA: ACM.

Gangji, A.; Walden, T.; Vaidyanathan, P.; Prudhommeaux,

E.; Bailey, R.; and Alm, C. O. 2017. Using co-captured face, gaze and verbal reactions to images of varying emotional content for analysis and semantic alignment. In *Proceedings of the Human-Aware AI Workshop at AAAI*, 621–627.

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. In Henderson, J. M., and Ferreira, F., eds., *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Psychology Press. 213–248.

Kienzle, W.; Schölkopf, B.; Wichmann, F. A.; and Franz, M. O. 2007. *How to Find Interesting Locations in Video: A Spatiotemporal Interest Point Detector Learned from Human Eye Movements*. Berlin, Heidelberg: Springer Berlin Heidelberg. 405–414.

Liang, P.; Taskar, B.; and Klein, D. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*, 104–111. Association for Computational Linguistics.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollar, P. 2014. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*.

Mahapatra, D.; Winkler, S.; and cheng Yen, S. 2008. Motion saliency outweighs other low-level features while watching videos. In *SPIE Human Vison and Electronic Imaging*, volume 6806.

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.

Mohammad, S. M., and Kiritchenko, S. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31(2):301–326.

Mohammad, S. M., and Turney, P. D. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, 26–34. Stroudsburg, PA, USA: Association for Computational Linguistics.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Mohammad, S. M. 2012. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, 246–255. Stroudsburg, PA, USA: Association for Computational Linguistics.

Naim, I.; Song, Y. C.; Liu, Q.; Kautz, H.; Luo, J.; and Gildea, D. 2014. Unsupervised alignment of natural language instructions with video segments. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Qu, S., and Chai, J. Y. 2008. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 244–253. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

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

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.

Warriner, A.; Kuperman, V.; and Brysbert, M. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods* 45:1191–1207.

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

Yu, C., and Ballard, D. H. 2004b. On the integration of grounding language and learning objects. In *Proceedings of Nineteenth AAAI Conference on Artificial Intelligence*, 488–493.