2017

Visual-Linguistic Semantic Alignment: Fusing Human Gaze and Spoken Narratives for Image Region Annotation

Preethi Vaidyanathan
pxv1621@rit.edu

Visual-Linguistic Semantic Alignment: Fusing Human Gaze and Spoken Narratives for Image Region Annotation

by

Preethi Vaidyanathan

M.S. Electrical Engg., Rochester Institute of Technology, 2009
B.Tech (equivalent to B.S.) Electronics Engg., KNMIET, India, 2007

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

2017

Signature of the Author _____

Accepted by _____
Coordinator, Ph.D. Degree Program Date

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE

COLLEGE OF SCIENCE

ROCHESTER INSTITUTE OF TECHNOLOGY

ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

Ph.D. DEGREE DISSERTATION

The Ph.D. Degree Dissertation of Preethi Vaidyanathan
has been examined and approved by the
dissertation committee as satisfactory for the
dissertation required for the
Ph.D. degree in Imaging Science

Dr. Jeff B. Pelz, Co-advisor

Dr. Rajendra Raj, External Chair

Dr. Cecilia O. Alm, Co-advisor

Dr. Emily T. Prud'hommeaux, Co-advisor

Dr. Anne R. Haake

Dr. Christopher Kanan

Date

Visual-Linguistic Semantic Alignment: Fusing Human Gaze and Spoken Narratives for Image Region Annotation

by

Preethi Vaidyanathan

Submitted to the
Chester F. Carlson Center for Imaging Science
in partial fulfillment of the requirements
for the Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

Advanced image-based application systems such as image retrieval and visual question answering depend heavily on semantic image region annotation. However, improvements in image region annotation are limited because of our inability to understand how humans, the end users, process these images and image regions. In this work, we expand a framework for capturing image region annotations where interpreting an image is influenced by the end user's visual perception skills, conceptual knowledge, and task-oriented goals. Human image understanding is reflected by individuals' visual and linguistic behaviors, but the meaningful computational integration and interpretation of their multimodal representations (e.g. gaze, text) remain a challenge. Our work explores the hypothesis that eye movements can help us understand experts' perceptual processes and that spoken language descriptions can reveal conceptual elements of image inspection tasks. We propose that there exists a meaningful relation between gaze, spoken narratives, and image content. Using unsupervised bitext alignment, we create meaningful mappings between participants' eye movements (which reveal key areas of images) and spoken descriptions of those images. The resulting alignments are then used to annotate image regions with concept labels. Our alignment accuracy exceeds baseline alignments that are obtained using both simultaneous and a fixed-delay temporal correspondence. Additionally, comparison of alignment accuracy between a method that identifies clusters in the images based on eye movements and a method that identifies

clusters using image features shows that the two approaches perform well on different types of images and concept labels. This suggests that an image annotation framework could integrate information from more than one technique to handle heterogeneous images. The resulting alignments can be used to create a database of low-level image features and high-level semantic annotations corresponding to perceptually important image regions. We demonstrate the applicability of the proposed framework with two datasets: one consisting of general-domain images and another with images from the domain of medicine. This work is an important contribution toward the highly challenging problem of fusing human-elicited multimodal data sources, a problem that will become increasingly important as low-resource scenarios become more common.

Acknowledgements

I would like to gratefully acknowledge the support, guidance, and encouragement of my doctoral advisor Dr. Jeff Pelz who believed in me at every step, even at times when I did not. I wish to express my indebtedness to my mentors Dr. Cecilia O. Alm who pushed me beyond my limits and shaped me into an independent researcher and Dr. Emily Prud'hommeaux, who provided valuable guidance on word alignment, data augmentation, and spoken language transcription. My gratitude extends to Dr. Anne Haake who laid the foundation of the research, and Dr. Christopher Kanan for sharing his dissertation experience and motivating me. Special thanks to Dr. Rajendra Raj for stepping in as my external chair and Drs. Pengcheng Shi and Cara Calvelli for their expert suggestions.

This dissertation would not have been possible without funding from the National Institutes of Health and National Science Foundation.

I sincerely thank Drs. Stefi Baum and David Messinger for giving me the opportunity to pursue my dreams. Special thanks to Sue Chan, Joyce French, Elizabeth Lockwood, Melanie Warren, and late Cindy Schultz, who were always eager to help and would cheer me up at difficult times. Also, the endless cookies and cakes provided by the MVRL and Carlson team contributed in keeping the energy going.

I also thank Mr. Dixon Cleveland and the family of LC Technologies for giving me the opportunity to do an internship. I was fortunate to gain more knowledge about eye tracking as well as witness the difference it can make in people's lives.

My gratitude extends out to my colleagues Kathryn Womack, Dr. Rui Li, Sai Krishna Mulpuru, Dong Wang, Xuan Guo, Rakshit Kothari, Kamran Binaee, and Wilson McCoy with whom I have collaborated on various occasions and enjoyed working with. I also thank my friends Laura Sesma, Sravani Vaddi, Nikita Moharir, Ashima Chhabra, Renu Singh, Tusharika, and Varun Maurya.

These acknowledgements would not be complete if I did not mention my good friend Kate Walders who made sure I paid attention to physical fitness along with mental fitness and Ramin Djawadi for his music that kept me motivated during my Ph.D. pursuit. I thank my colleagues and friends for their encouragement, support, and most of all for keeping me smiling.

I must acknowledge with tremendous and deep thanks my mother Buvaneswari Vaidyanathan and father K. Vaidyanathan for giving me a good foundation with which to meet life, sister Priya Vaidyanathan who is more than a best friend, and brother-in-law

Mahesh Vencata who would constantly ask me “Are you done yet?” but supported me when in need. Last, but not the least, I thank my beautiful niece Meera and mischievous nephew Siddharth who were all the distraction one could have during a Ph.D. lifetime.

To Chechappa, Bamu, Bapu, Parvathi, and Chitra.

Contents

List of Figures

List of Tables

1

Introduction

Digital imaging has seen an exponential growth in the past decade with usage ranging from personal photos and social media to more complex applications in education and medicine. With advanced cameras, photographs (images) are not only used for capturing memory or evidence, but for facilitating decision making as well. For example, doctors use medical images to help diagnose and determine the treatment of diseases. Ideally, for computers to be able to assist humans in their reasoning and decision making process, they need to process these images as well as humans do. Intelligent computers should be capable of making inferences about what people look at and what they say about what they look at. Therefore, computers should be able to acquire and use learned associations. This is known as semantic image annotation, and when performed on images to identify objects or regions it is called semantic image region annotation. With this knowledge and learning, computers would be able to provide useful and detailed information about an object. For instance, when a user gazes at a *painting* in a museum, the computer can highlight areas of the painting where an expert artist looked at and provide more conceptual information about that area. This work integrates gaze and linguistic information indicating 'what people look at' and 'what they say', to identify the objects and their corresponding names or labels in images.

Automatic semantic image region annotation is the task of computationally identifying image regions that are perceptually meaningful for humans and associating them with appropriate natural language concept labels. It plays a key role in developing sophisticated image-based information systems but it is a difficult and long-standing problem [Smeulders et al., 2000, Zhang et al., 2012, Karpathy and Fei-Fei, 2015]. An example of

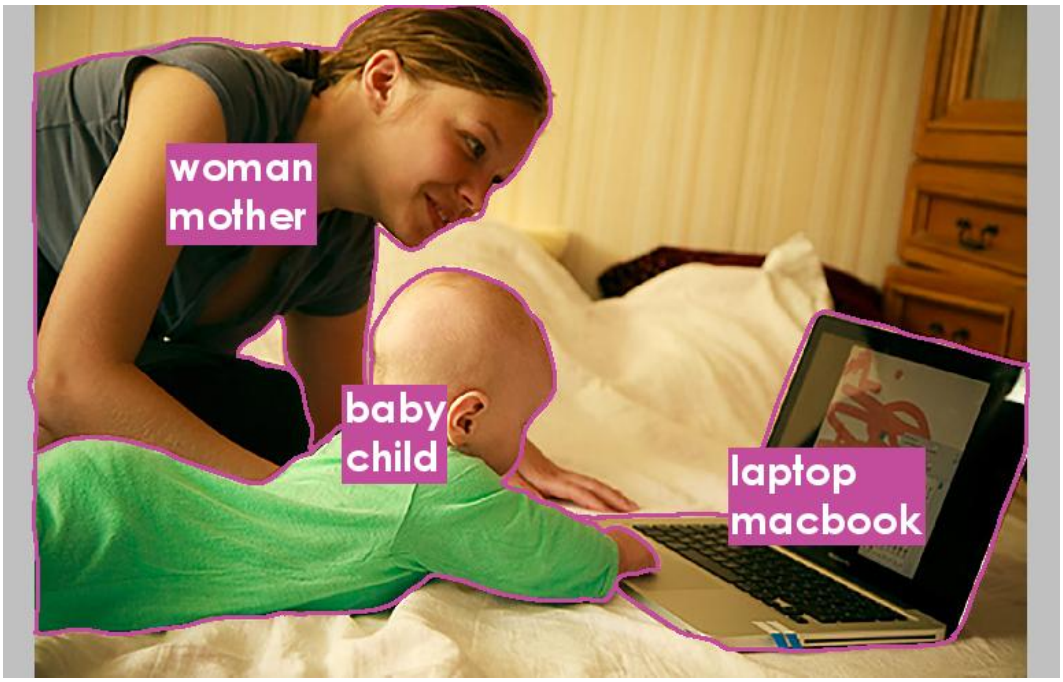

Figure 1.1: Example illustrating the concept of image region annotation. The two-fold process involves identifying and segmenting correct regions in an image and labeling them correctly.

semantic image region annotation where regions in an image are well segmented and labeled with corresponding appropriate words is shown in Figure 1.1. Although the entire image in Figure 1.1 could be annotated as, for example *mother playing with baby*, it is intuitive to annotate regions or objects such as *woman* and *baby* that constitute the image. These detailed annotations for image regions can assist in important applications such as image retrieval where the user could be searching for images of babies or visual question-answering where the user could be asking what the baby is playing with. Further, relations between annotated regions could also be inferred on their basis. High-level cognitive processing and experience enable humans to process images at a semantic level that remains difficult for a computer [Shanteau, 1992, Goldstone, 1998, Zhu et al., 2016, Zitnick et al., 2016, Tavakoli et al., 2017]. This work proposes a novel framework to fuse multimodal visual and linguistic data elicited from humans to achieve semantic annotations of image regions. Gaze locations over an image can act as pointers and reveal perceptually important regions and their relation to one another from the perspective of multiple observers. Also, when humans communicate their understanding of and reasoning about

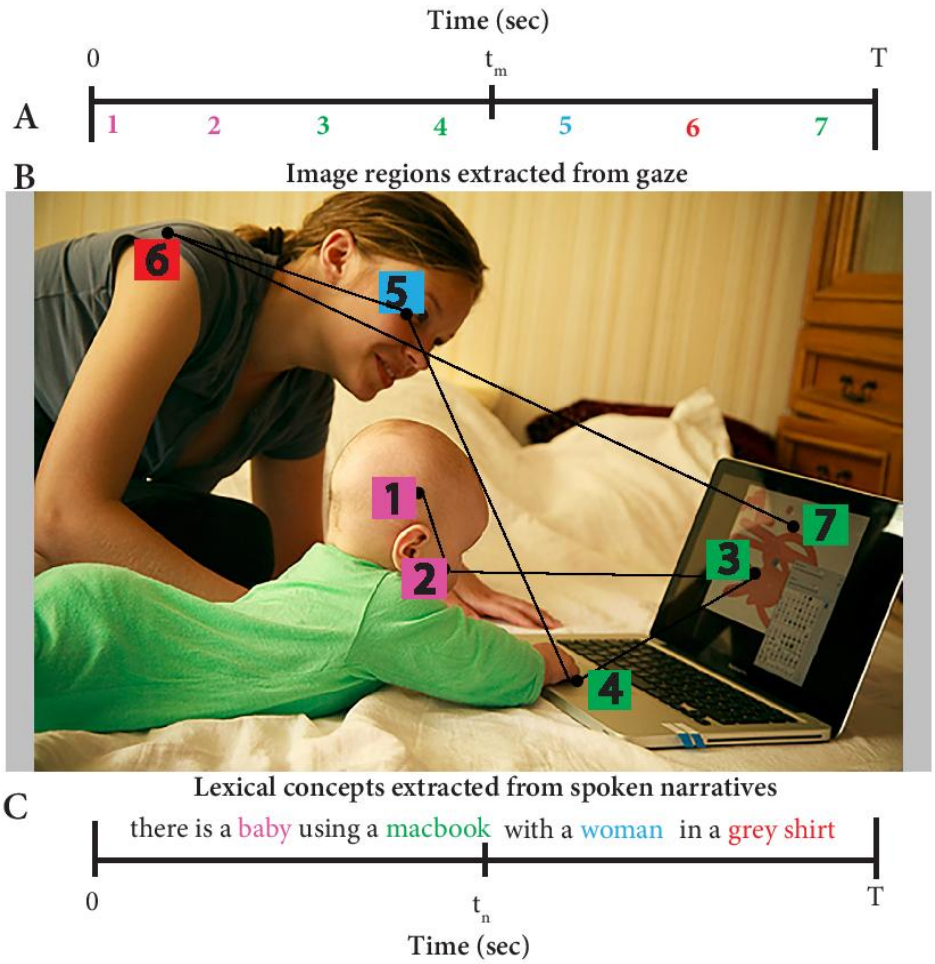

Figure 1.2: Panels A and C show the eye fixation locations extracted from eye movements and lexical concepts (labels) obtained from spoken narratives, respectively, over a common time scale. This hypothetical example shows that the data collection session for this image took T seconds. Panel B shows the seven image regions that were looked at by the participant in the original image. The proposed algorithm will align words such as *baby* and *woman* with corresponding regions, using the bitext alignment technique.

images, spoken language is the most natural and convenient instrument of expression. In this case co-captured image descriptions convey relevant meaning, particularly special knowledge and experience that the human observers possess. An important novelty of this work lies in the integration of human observers' perceptual and conceptual knowledge using natural language processing (NLP) methods to annotate images.

People often have the intuition that when they look at an object and mention its name, they do so simultaneously. However, research in sentence production has shown that there is a variable amount of time between when a person looks at an object and when they name it aloud [Meyer et al., 1998, van der Meulen, 2003, Griffin, 2004]. Therefore, even when visual and linguistic information is co-captured we cannot assume that a fixation on a region will occur simultaneously with the verbal naming of the region. This lag, which can vary in length, demands more sophisticated methods.

The bitext word alignment method [Brown et al., 1993], widely used in machine translation, aligns the words of a sentence in one language with the word or words in another language that are likely to be translations. For our problem, the location of eye fixations on images are analyzed as *visual units* that encode visual regions while the spoken descriptions about the images contain the *linguistic units*. Prior work confirms the utility of associating words and sentences with images, objects and image regions, and videos [Forsyth et al., 2009, Kuznetsova et al., 2013, Kong et al., 2014, Socher et al., 2014, Thomason et al., 2014]. Many of these works rely on written description of general-domain images, making the framework difficult to translate to domain-specific images. This work, in contrast, focuses on building a framework that can be applied to any image dataset. Perceptual and conceptual information is combined via the integration of gaze and narratives to advance annotation of image regions.

This study aims to understand and encode important image information by semantically annotating important regions of an image with natural language descriptors as shown Figure 1.1. As shown in the Figure 1.2, the framework uses gaze locations on images together with words uttered by observers to learn perceptually important image regions and the corresponding linguistic descriptors. The study also asserts that the combination of perceptual information (via eye movements) and more naturally obtained conceptual information (via spoken narratives) contributes to the understanding of an image.

1.1 Contributions

The four main contributions of this work are as follows:

1. Show that human-elicited gaze and narratives jointly provide information that if considered separately would be insufficient to understand how humans perform image inspection and description tasks.

2. Exemplify the broad applicability of the visual-linguistic alignment framework by comprehensively using and evaluating it with both domain-specific and scaled-up general-domain image datasets.

3. Compare the performance of various image region segmentation techniques used to identify the visual units for the two datasets to illustrate the strengths and weaknesses both for the described framework and respective segmentation techniques.

4. Provide the research community with a large multimodal dataset comprised of co-captured gaze and spoken descriptions data collected during an image inspection task involving general-domain images.

Chapter 2 discusses prior works in image region annotation as well as multimodal data integration. Chapters 3 and 4 provide details about the collected multimodal data and findings from preliminary analysis. This is followed by the discussion of the proposed visual-linguistic alignment framework in Chapter 5. Results of the framework are discussed in Chapter 6 followed by future work and conclusions in Chapter 7. Chapter 8 lists all publications stemming from this work with some mathematical details about technical concepts used in this work explained in the appendices.

2

Previous Work

2.1 Challenges in image annotation

The goal of this work is to automatically annotate images through the integration of end users' perceptual and conceptual information with the information in the images. Research efforts in automatic image annotation can be categorized into three types of approaches [Zhang et al., 2012, Li et al., 2015]. The first approach involves manual annotations by humans using text [Tamura and Yokoya, 1984, Chang and Hsu, 1992]. This approach is brittle since as the number of images increases, manual annotation becomes impractical. The second approach annotates images using low-level features such as color, shape, and texture to index images [Saber et al., 1996, Jain and Vailaya, 1996, Sivic and Zisserman, 2003]. The third approach is more recent and attempts to bridge the semantic gap. It involves understanding images and learning the semantic concept models that can be used to label new images [Duygulu et al., 2002, Qu and Chai, 2008, Ballerini et al., 2009, Karpathy and Fei-Fei, 2015, Johnson et al., 2015].

Treisman and Gelade (1980) were the first to introduce the concept of semantic understanding of images. In their feature integration theory, they proposed that processing of image information is a dynamic interaction between bottom-up and top-down directed processes. The bottom-up process corresponds to the stimulus-driven discovery of low-level image information pieces whereas the top-down process is the user-driven processing of the discovered information pieces. The user-driven processing involves linking these disjoint information pieces into perceptually meaningful image concepts and objects.

In spite of the proposed integration theory, for a long time image annotation

algorithms were built solely on low-level features such as color and texture to perform segmentation and retrieval [Saber et al., 1996, Shi and Malik, 2000]. Algorithms employing these low-level features succeeded in capturing basic statistics of natural scenes [Fei-Fei and Perona, 2005], identifying faces [Viola and Jones, 2004], or segmenting single objects in a scene [Kumar et al., 2010, Jaber and Saber, 2010] but were unable to deal with multiple objects in the scene, statistics of domain-related images, and other high-level processing tasks. For example, while the bottom-up methods helped in automatic detection and segmentation of objects in a scene, they did not provide the relationship between these objects or the contextual meaning of the scene [Li et al., 2009]. Recent researchers have had some success with generating image descriptions and semantic labeling [Kong et al., 2014, Karpathy and Fei-Fei, 2015, Yatskar et al., 2016]. However, their techniques cannot be easily translated to complex domains such as medicine.

To bridge the semantic gap, Duygulu et al. (2002) proposed the use of machine translation to combine image content with the accompanying text for object recognition [Duygulu et al., 2002]. Following this, other researchers proposed several integrating techniques using different mathematical approaches such as Bayesian methods, Latent Dirichlet Allocation and Latent Semantic Analysis methods [Barnard et al., 2003, Li and Wang, 2003, Berg et al., 2004b, Berg et al., 2004a]. Similarly, researchers proposed the use of deep learning to combine text and images for image annotation [Karpathy and Fei-Fei, 2015, Vinyals et al., 2014], as well as unsupervised alignment to align text instructions with video segments [Naim et al.,]. In their recent work, Johnson et al. (2015) suggested the use of neighboring test images and their annotations to disambiguate and annotate otherwise ambiguous images. These approaches bridge the semantic gap to a certain extent by bringing in multimodal information through images and text. However they do not involve speech or gaze data and are only successful on certain types of images failing to capture the semantics of images in complex domains. Qu and Chai extended the idea of using multimodal data by using speech and eye gaze that are more natural to elicit than traditional methods. However their application scenario is a 3D simulated scene with no real-life challenges such as occlusion to deal with [Qu and Chai, 2008].

All the above approaches are interesting and successful to some extent, for example on scenic images and single object images, but they cannot be applied to complex tasks requiring more than identifying simple objects. This is because they do not incorporate the end users' goal or experts' knowledge during the modeling/learning

stage. For crucial applications such as clinical decision making or pilot training these methods are unreliable and demand approaches that incorporate more human intelligence [Stark and Privitera, 1997, Scheirer et al., 2014]. The works of Scheirer et al. (2014) emphasized that one needs to involve the human early on in the modeling process as opposed to using the human performance solely for validation of machine performance. In their work they used visual psychophysics to draw out information reflecting human capacity which they call *perceptual annotation* and combine it with image features to build a better face detector [Scheirer et al., 2014]. Motivated by the body of prior research, our work proposes to fuse naturally obtained multimodal visual-linguistic data from experts and build a semantic image region annotation framework over it.

2.2 Importance of capturing perceptual and conceptual expertise

An integral component of this study is the use of eye movements and spoken narratives to elicit human expertise or knowledge. Eye movements can be considered pointers to the perceptually important regions of an image while spoken narratives can reveal conceptual elements associated with those regions. Capturing perceptual and conceptual information relevant to the image processing system's end user's goal is of paramount importance to improve the annotation of images. Image-information systems must be reliable enough to assist in goal-oriented performance [Müller et al., 2004]. End users may not merely seek images or regions that have similar low-level features such as color or texture but they may want to locate, classify, or segment an image based on high-level reasoning features. Moreover, in domain-specific images, such as medical images, low-level features do not sufficiently capture the subtle but key attributes that are crucial for decision making in the visual domain of interest [Tang et al., 1999].

Studies have found that perceptual and conceptual expertise help a user formulate more specific and comprehensive descriptions of images and these correlate with the user's ability to express their information needs [Goldstone, 1998, Vakkari, 2002]. Williams and Elliott (1999) examined the effect of anxiety and perceptual skill on the visual search strategy of karate experts while viewing taped karate offensive sequences. They observed that karate experts showed an increased awareness and superior anticipation under all levels of anxiety as opposed to novices [Williams and Elliott, 1999]. In another study, researchers investigated the anticipatory skills of rugby players using a video-based

test and observed faster responses and higher accuracy in highly skilled players [Gabbett and Abernethy, 2013]. In the field of radiology as well, through expert-novice comparison it is evident that novices tend to categorize objects first at the general level whereas experts show a preference to identify objects at a more specific level [Tanaka et al., 2005, Hoffman and Fiore, 2007]. Moreover, in their study with radiologists, Hoffman and Fiore (2007) reported that experts can perceive certain aspects that are literally invisible to the novice. Similarly, Krupinski (2000) showed that perceptual skills exhibited by radiologists when searching medical images do not necessarily transfer to a more general task such as "Where's Waldo" where one has to search for a character called Waldo among other similar looking characters. Therefore the same expert of one field could be a novice in another area, implying that investigation into expertise-related differences must be done in a domain or task-specific manner.

Researchers use various knowledge elicitation methods to capture human users' expertise. One of the most common methods is interviewing and asking participants to describe the decision making process through the think-aloud protocol. One problem with this method is that it will only produce what an expert can verbalize as an answer to the particular question [Shadbolt and Smart, 2015]. It also requires the expert to perform a secondary task in parallel with the primary task. Any non-verbalizable information is lost such as where these experts look in the image. Another widely used technique is to ask the experts to manually mark important regions in images, etc., [Shyu et al., 1999, Wang et al., 2012b]. The drawback with this technique is the loss of any information pertaining to how the expert arrived at that decision, i.e. information in the image that the expert used to decide where to mark. This work uses eye movements and spoken language as they are non-invasive and more natural tools that enable us to draw out the tacit perceptual and conceptual information of humans.

2.3 Need to integrate eye movements and spoken narratives

This work emphasizes a multimodal approach to achieve image region annotation in order to improve our understanding of images. Psychologists have previously used eye movements to find answers to various perceptually-related image understanding queries. Researchers have found a strong connection between visual attention strategies and cognitively driven perception [Oliva et al., 2003, Borji, 2009]. Although the inherent low-level structure (such as bright color or high-contrast edges) of images drive the

initial stages of visual attention, meaningful content of the image soon comes into play [Stark and Privitera, 1997]. As a result eye movement behavior can differ with the change in task even if the stimuli remain the same [Yarbus, 1965, Yarbus et al., 1967]. Recent studies suggest that top-down processes influence visual perception more than bottom-up processes in real tasks [Castelhano et al., 2009]. Although eye movements cannot completely reveal complex cognitive processes, empirical studies have established relationships between visual perception and recognition. Walther et al. (2005) used visual attention to learn to recognize objects in cluttered indoor and outdoor scenes. Likewise, Mishra et al. (2009) used eye movements to aid their segmentation algorithm. Eye movements have also been used to annotate video frames of paper printing and stapling tasks [Yu and Ballard, 2004a].

Analogously, researchers in psycholinguistics have used language to understand certain aspects of human psychology. Natural language is a fundamental knowledge representation system, and spoken narratives can indicate viewers' focus of attention [Ji and Ploux, 2003]. Researchers have previously used verbal narratives to investigate the process of language production in simple day-to-day tasks and storytelling [Meyer et al., 1998, Holsanova, 2006]. Language is also used by researchers in computer vision to caption images and video frames [Karpathy and Fei-Fei, 2015, Naim et al., , Naim et al.,]. Through simultaneous multimodal gaze-verbal capture higher-level conceptual knowledge of the expert can be added to the eye movements for analysis.

Empirical experiments have shown that eye movements are closely time-locked with human language processing [Just and Carpenter, 1976, Ferreira and Tanenhaus, 2007, Griffin, 2004]. Linguists, for instance scholars active in psycholinguistics, have used eye movements as a tool to understand language. Similarly, eye movement researchers have incorporated linguistic input into their studies. Just and Carpenter (1980) described how measures like fixation duration changed depending on the linguistic characteristics of the text being read. Soon Frazier and Rayner (1982) pioneered the use of eye movements to understand written language and syntactic processing. During the following two decades numerous contributions were made by researchers who used eye movements as a tool to reveal the way written language is processed [Heller, 1988, Pollatsek et al., 1993, Rayner, 1998].

Some researchers took an interest in investigating language comprehension through the use of eye movements [Tanenhaus et al., 1995, Dahan et al., 2001, Spivey et al., 2002, Richardson and Dale, 2005]. They revealed that it was possible to investigate how

people understand spoken language by measuring people's eye movements while listening to verbal commands and executing them. Richardson and Dale (2005) conducted a study to understand the coupling between speakers and listeners, reporting that the interlocutors' eye movements were closely time-locked. Another study showed that eye movements can be used to understand the stages of language comprehension such as hearing a command, interpreting it, and engaging in resolving and executing commands [Kaiser and Trueswell, 2008]. Such prior works revealed that a relation between cognition, vision, and language exists and that by integrating eye movements and spoken narratives, an understanding of cognitively complex tasks can be obtained.

Inspired by language comprehension studies, Cooper (1974) used eye movements to investigate language production. He observed that participants' fixations were generated before the end of words they used in narration. Similarly, Meyer et al. (1998) investigated sentence generation and fixation duration during simple noun phrases and found that people fixated the next object only after lexically encoding, but before executing the prior word. Authors also observed that mean viewing time for speakers was significantly longer for objects with low frequency names (names that were not used very often) than with high frequency names (names used very often). This is particularly interesting because our work focuses on modeling the visual-linguistic relation and prior research has revealed that factors such as frequency of a name can play an important role. In another study, van der Meulen (2003) observed that participants fixated the objects to be named in the order of mention and once just before naming. This indicates that speech is performed in an incremental fashion, i.e. speakers tend to look at the objects they are about to find words for in the same order in which the object names were mentioned in the utterance.

The growing interest in this multimodal field motivated Griffin and Bock (2000) to study the temporal relation between event apprehension, sentence formulation, and speech execution. They compared the timing and trajectories of selective fixations to agents (objects performing the action) and patients (objects undergoing the action) across different tasks and found that speakers' eye movements were guided by an overall understanding of the event/scene rather than by the salience of the individual objects in it. The distribution of fixation times anticipated the order of mention regardless of sentence structure, partly confirming van der Meulen's findings. They also found that when speaking extemporaneously, speakers began fixating elements less than a second before naming them, suggesting that people spend some time looking at objects prior to naming them [Griffin and Bock, 2000, Griffin, 2004]. Recently, a study was conducted to

understand how complex noun phrases are produced and if the production process was similar to that of simple noun phrases [Shao et al., 2013].

The above findings indicate that vision and language are tightly integrated. In 1964, Kirsch published a paper that attempted to combine the two cognitive modalities to understand semantic processing. His work combined lexical and visual data from newspaper photographs and briefly laid the ground for studying the two together. Although there were other researchers who performed studies along the same lines [Badler, 1975, Waltz, 1980, Herzog and Wazinski, 1994], the focus shifted away from understanding the interactions of the two modalities until 1995 when Srihari investigated the correspondence problem and visual semantics [Srihari, 1995]. In the following years there was an increased interest in developing methods to integrate language and vision and understand how human cognition works, including a proposed technique to integrate the two modalities using the mutual information model [Roy, 2000, Roy and Pentland, 2002]. Several researchers investigated the multimodal integration problem in relation to sentence prediction and object naming in scenic images [Coco and Keller, 2012, Clarke et al., 2013, Yun et al., 2013a, Yun et al., 2013b]. While these works were successful in infant-directed interactions or on scenic images it is not clear that they would translate successfully to complex scenarios such as clinical decision making.

Although there is some relationship between the timing of eye movements and spoken narratives, an exact or fixed-delay temporal match indicating that a fixation on a region will occur simultaneously or after a fixed time interval with the verbal naming of the region cannot be assumed. Holsanova (2006) studied the interaction of vision and language over time by investigating the dynamics of picture viewing and picture description. Her research revealed that correspondence between the spoken words and the objects in the scene could be of different types, e.g. one-to-one or many-to-one [Holsanova, 2008]. These findings partly confirm hypotheses such as the existence of a temporal relationship between when objects are gazed upon and when their names are uttered but lack any quantitative validity or technical modeling that could be used in automated systems. An important concern that arises from prior research is how feasible it is to use temporal correlation to model the temporal relation between eye movements and language given that there are various factors that affect this relation. Therefore, we need to employ other techniques such as bitext alignment, which is widely used in machine translation to align words in one language to their corresponding translations in another language.

Duygulu et al. investigated a method to automatically recognize and annotate objects

in scenes [Duygulu et al., 2002]. They segmented images into regions and clustered them into region types that they referred to as *blobs*. Further, expectation-maximization was used to learn the mapping between the blobs and the keywords for a given image. However, the image regions or blobs and keywords were obtained using image segmentation methods and a large vocabulary from captions without any human-elicited eye movements and spoken narratives. A similar technique was used by other scholars to automatically match words to the corresponding pictures [Barnard et al., 2003], faces in pictures to names [Berg et al., 2004a, Berg et al., 2004b], and natural language instructions to video frames for a particular task [Naim et al.,]. Jamieson et al. (2006) addressed the problem of grouping image features, namely SIFT (scale-invariant feature transform) features, by associating them with the names of objects appearing in cluttered scenes obtained through captioning. Qu and Chai (2008) proposed that a modified IBM translation model II [Brown et al., 1993] together with perceptual information and observer's domain semantic information expressed using spoken language could be helpful in interpreting unexpected user language inputs in conversational systems. They applied this idea to computer-generated videos involving 3D objects in a room scene where the participants were asked various questions about the decoration of the 3D simulated room. Their work provides an interesting factor of including domain knowledge into the translation model. However, the use of a 3D simulated room scene with objects simplifies many challenges faced when dealing with complex real-life scenarios.

Yu and Ballard (2004a) appear to be the first to have a pioneering paper in exploring how word alignment methods could be extended to a challenging task of grounding spoken language in sensory perceptions. Similar to our work, they transcribed the audio and extracted nouns as object names [Yu and Ballard, 2004b]. For the perceptual representation of objects, Yu and Ballard segmented the objects in the video using gaze data. Further, these objects were represented using multidimensional color and shape features. The multimodal data consisting of words and objects was then integrated using IBM Model 2, a non-HMM based word alignment method commonly used in machine translation, to learn correspondences. In their extended work, they combined scene video, participant's gaze, head motion, and object names obtained from verbal narratives while performing everyday tasks such as stapling printed papers, to annotate objects and categorize action scenes in video [Yu and Ballard, 2004a]. Their work provides a good understanding of how multimodal data can be combined for a video annotation task. However, their work involves only nine [Yu and Ballard, 2004b] and

six [Yu and Ballard, 2004a] participants and three trivial video stimuli. Primarily, Yu and Ballard explored object annotation with images that had uniform background and consisted of distinct objects that were trivial to segment. It is also unclear whether their work could be easily generalized or extended to other domains such as medical image inspection. Lastly, Yu and Ballard's work does not provide a clearer evaluation and baseline comparison. Motivated by their work, we investigate multimodal image region annotation with images that do not have uniform background and consist of image regions including skin lesions that are difficult to segment. We explore the annotation task using two larger datasets consisting of images from general-domain and specific-domain, respectively, as well as provide baseline comparison.

2.4 Summary

To summarize, it is evident that image-based annotation systems would only benefit from incorporating the end user during the design process. A framework that fuses multimodal information, obtained from the humans in more natural ways, can contribute to a higher-level understanding and modeling of both simple and complex images.

3

Eye Tracking and Spoken Description

This chapter begins by briefly describing the components of the experimental design in section 3.1. Four IRB-approved eye tracking studies were conducted as part of this work. Sections 3.2, 3.3, and 3.4 describe three of the eye tracking studies conducted with a larger research team, with the detailed description of the fourth study, specific to this work, in Chapter 4. Section 3.5 provides insight into the data quality and Section 3.6 discusses some preliminary results obtained using the collected data.

3.1 Components of the experimental design

In this section we briefly introduce eye tracking, spoken description, our approach to elicit natural data from participants.

3.1.1 Eye tracking

Visual perception is an active dynamic process in which the viewer seeks out specific information to support ongoing cognitive and behavioral activity [Malcolm and Henderson, 2010]. Visual perception can be divided into two main phases, low-level vision and high-level vision. Low-level vision incorporates gathering of visual information from the outside world, such as extracting object boundaries or color, which is then transmitted to the visual cortex for further processing. High-level vision is concerned with problems such as object recognition and classification that involves appropriate

interpretation of the information obtained from low-level vision [Ullman, 2000]. Both these phases are intertwined temporally in cognitive processes and are crucial components of perceptual skill. Humans integrate the low-level information gathered through their vision system (e.g. eye movements) with high-level knowledge in their mind to perform the reasoning process. By tracking human experts' eye movements we can investigate where they focus their attention over time and what perceptual strategies they employ during image inspection. Therefore, eye movements of humans can be used to extract useful information about complex cognitive processes.

The eye moves frequently, shifting the gaze to subsequently foveate different portions of the world. There are two key concepts in eye movements that will be referred to frequently in this work: (1) *saccade* - a type of eye movement, and (2) *fixation* - a state of the eye when gaze is relatively stable. Fixations occur when a stationary observer is viewing a static object. Other types of eye movements include smooth pursuit, vergence, and vestibulo-ocular eye movements. *Smooth pursuit* movements occur when a stationary observer smoothly pursues a moving object [Leigh and Zee, 2015]. Leigh and Zee describe *vergence* movements as the movements that allow rotation of the two eyes simultaneously in the same or opposite directions so that gaze can be shifted between different depth planes. *Vestibulo-ocular* movements are reflex movements that come into play when the observer is in motion [Leigh and Zee, 2015]. When an observer fixates at an object and moves their head the eyes rotate in the opposite direction of the head to compensate for the head movement. Since our data collection process involves stationary observers viewing static images, we focus on fixations and saccades.

1. **Saccade:** A saccade is a ballistic eye movement that observers make to shift their point of regard. On average a person makes about 150,000 saccades a day and can execute about 2-4 saccades per second [Phillips and Edelman, 2008]. Although we move our eyes frequently, we do not consciously perceive the image motion resulting from saccades.

2. **Fixation:** The saccades are separated by fixations, periods of retinal image stability when we obtain high resolution information about the visual environment. The duration of a fixation depends on participants' interest in the visual region but typically ranges from 200 to 400 ms and can vary with the underlying task [Pelz and Canosa, 2001, Lipps and Pelz, 2004].

There are various techniques used to track a person's eyes. In this work, we use remote

eye trackers to collect eye movement data. Remote eye trackers are typically placed more than 50 cm from the participant, and hence, are not invasive. The majority of trackers use infrared to illuminate the eyes and can be binocular or monocular. Depending on where the infrared illuminator is placed with respect to the image capturing camera, eye trackers can be categorized as bright pupil or dark pupil. More information about various types of eye trackers can be found in Duchowski's recent textbook [Duchowski, 2017].

A calibration procedure is usually required in eye tracking to collect enough information about the participant's eye to accurately predict the gaze point and to account for the individuality of the participant. It is the process through which the eye tracker measures characteristics such as shape of the participant's eyes and the relative position of the fovea. During calibration a participant is required to look at some known points within the scene while certain eye features are captured (depending on the eye tracking technique) for each point. Some researchers include a validation procedure after calibration to determine the participant's calibration accuracy and if the validation indicates (using a machine or manually defined threshold) unacceptable accuracy, the calibration procedure is repeated.

3.1.2 Spoken descriptions

Conceptual knowledge of a human is not directly observable but capturing observers' spoken descriptions of images during image inspection can give insight into their conceptual reasoning process. Transcribed spoken narratives can be analyzed at various levels depending on the goal of the analysis. Narratives can be segmented into paragraphs, utterances, word tokens, morphemes, and other units of analysis. In this work we record verbal data and transcribe them.

3.1.3 Master-Apprentice approach

Participants may not be aware of the steps they may take while describing images, so they are often unable to explain their steps explicitly [Beyer and Holtzblatt, 1997]. Therefore, the experiment is designed carefully in order to draw out these complex cognitive processes in a natural and efficient way. According to Beyer and Holtzblatt, teaching in the context of work provides an efficient way to bring out the participants' tacit knowledge. This is known as the Master-Apprentice (MA) model. We used an MA model in all data collection setups to draw out the details that participants might miss during a regular widely used think-aloud process. The MA model also avoids the secondary-task problem inherent in

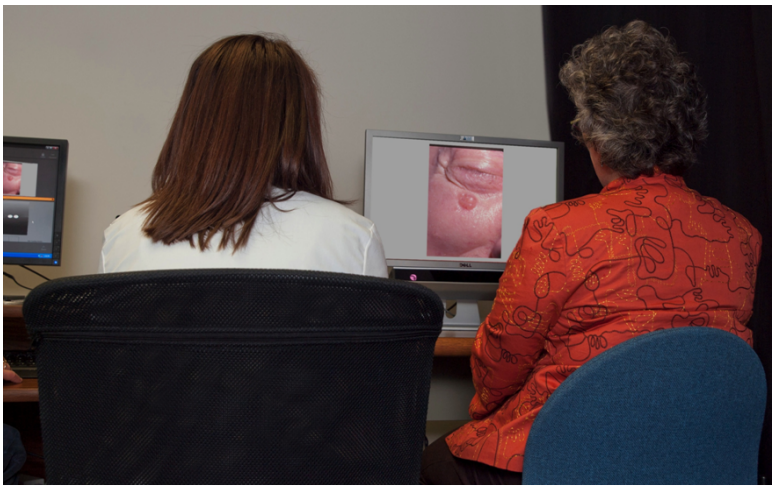

Figure 3.1: Master-Apprentice approach: Here the domain expert (right) is the *Master* sitting in front of the eye-tracker which is located underneath the computer screen displaying the images. In this case, a Physician Assistant student (left) is the *Apprentice* who does not talk during the experiment.

the traditional think-aloud paradigm because 'teaching' in this context feels more natural than 'thinking out loud,' a task that observers often need to be reminded to continue.

3.2 Gaze-verbal data collection for experts (DERM I)

We eye-tracked 16 participants, including 12 board-certified experienced dermatologists ('attendings') and four dermatologists in-training ('residents'). An illustration of the experimental setup is shown in Figure 3.1. In this study all participants were recruited from the Rochester area. Apart from the 16 participants we also had Physician Assistant students who served as the Apprentices in order to motivate the dermatologists in the Master-Apprentice approach (Figure 3.1). A set of 50 dermatological images (provided by Logical Images Inc, Rochester, NY and Dr. Cara F. Calvelli, M.D.) was selected for the study, with each image representing a different diagnosis. These image cases vary in complexity in terms of both dermatological knowledge and clinical attributes. We presented each image to the participants on a 22-inch LCD monitor (1680×1050 pixels) approximately 70cm from the participant. The full display subtended approximately 38×22 degrees of visual angle at that distance, though most images did not fill the field. The aspect ratio of the images varied, but on average, the images subtended approximately

32×22 degrees on the display. We used SensoMotoric Instruments (SMI) RED remote eye-tracker attached to the above display, as shown in Figure 3.1, and running at 50Hz to collect gaze data. The reported accuracy of the RED eye-tracker is 0.5 degree. It monitors the position of a participant's point of regard on the image in a non-intrusive way. We use a double computer set-up wherein one of the computers was used to present the image and the other ran the software iViewX gaze tracking system and Experiment Center 2.3. The dermatologists were instructed to "examine and describe each image verbally as if teaching the trainee sitting next to you to make a diagnosis based on the image." A nine point calibration followed by a four point validation was conducted after every 10 images with a re-calibration done, if necessary. In addition, we recorded verbal narratives using an Olympus VN-6000 Digital Voice handheld recorder. It was small and convenient but not high quality. This dataset is referred to as DERM I dataset in this thesis.

3.3 Gaze data collection for novices (NOV)

The second experiment included 15 undergraduate students with no medical training recruited from Rochester Institute of Technology. Eight of the images from the original DERM I dataset were judged to be too disturbing for a non-medical audience, and were removed, leaving a subset of 42 images for the second study. The eye-tracker set-up and calibration routine were identical to that used in the collection of the DERM I dataset. In order to replicate as closely as possible the conditions of the physician group in the DERM I dataset, the undergraduates in the novice group were instructed to "examine and describe each image as if you are describing it over the phone to a dermatologist who cannot see the image but has to diagnose it." No verbal data were collected in this experiment. This dataset is referred to as the NOV dataset in this thesis indicating it involves novices with respect to dermatology.

3.4 Gaze-verbal data collection for experts (DERM II)

In this study, we eye tracked 9 attending dermatologists and 18 dermatology residents inspecting 30 dermatological images. The images were selected such that they span about three to four primary lesion types with each type consisting of equal number of images as others. Participants were instructed to examine and describe each image aloud to an imaginary trainee. In addition to other descriptions, they were asked to provide a

Table 3.1: Sample raw data as obtained from SMI eye tracker showing from left to right: system timestamp, left-eye horizontal and vertical fixation locations, right-eye horizontal and vertical locations, left-eye and right-eye event, respectively.

Time	L_x[px]	L_y[px]	R_x[px]	R_y[px]	L Event	R Event
7456470899	919.19	504.03	919.19	504.03	Fixation	Fixation

differential diagnosis, a final diagnosis as well the certainty of their final diagnosis expressed as a percentage, while their eye movements and verbal narratives were being recorded. For the calibration routine, we performed a validation after every 5 images and re-calibration was performed only if the participants' validation error was more than one degree. The 50Hz SMI remote eye-tracker was replaced with a 250Hz SMI remote eye-tracker. The 250Hz tracker results in higher number of samples and also uses a velocity-based saccade detection algorithm to find saccades more accurately as compared to the dispersion-based algorithm used in the 50Hz. A blank gray slide and a test slide with a small, visible target with an invisible trigger area of interest were inserted between every two stimuli. Using a gray slide ensured that the gaze on one image did not influence the gaze on the following image. The test slide helped us measure (post-experiment) the drift in eye movement accuracy that takes place over time. Measuring the distance between the participants' fixation at the center target and the actual location of the center target provided us with information regarding drifting (due to participants' movement) that may have occurred. We used a TASCAM DR-100MKII audio recorder with a lapel microphone to collect the audio recordings as opposed to the recorder used in DERM I data collection. The rest of the eye-tracking set-up was similar to DERM I data collection. This dataset is referred to as DERM II dataset in this thesis and is used in the visual-linguistic alignment framework.

3.5 Fixations, narratives, and data quality

Fixations: The collected raw gaze data is processed using the SMI software package BeGaze 3.1.117 to detect fixations and saccades. The fixations are reported as x, y pixel coordinates of the image indicating where the observer gazed at. Table 3.1 shows an example of the output from the BeGaze software. Figure 3.2 (right) shows an observer's eye movements overlaid on the image with the red circles representing the fixation locations.

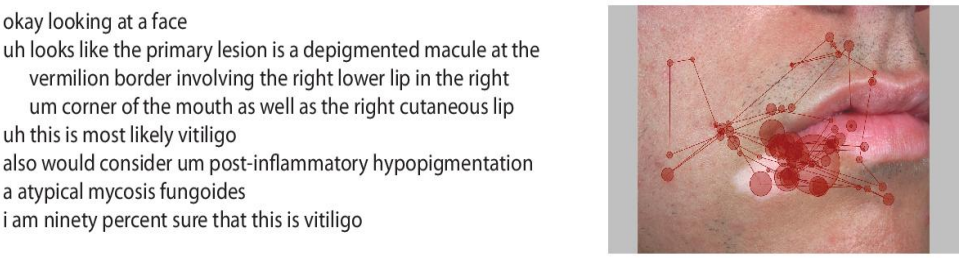

Figure 3.2: Example of multimodal data. On the left is the transcription of the spoken description. On the right is the eye movement data overlaid on the image. The radius of the red circles represent the amount of time spent fixating that location and the red lines represent change of fixation location i.e. a saccade.

Narratives: We manually transcribed and time-aligned the spoken description recordings at the word level for both the DERM I and DERM II datasets using Praat, a software package for speech analysis [Boersma, 2002]. An example of transcribed verbal description from the DERM II dataset is shown in Figure 3.2 (left). As mentioned earlier no verbal data was collected for the NOV dataset.

Data quality: Since the accuracy of eye trackers is not exactly as stated by the manufacturer [Wang et al., 2012a], we analyzed the sets of collected eye movement data for calibration errors. For each study, we calculate the mean calibration accuracy in the horizontal and vertical direction for every participant by averaging over the participants' calibration data. Following this the overall mean and the standard deviation across all the participants is calculated for the two directions respectively. Participants whose means in both directions were within two standard deviations of the overall mean in that direction were included in further analysis. Using the calibration method for data quality, we selected 75%, 100%, and 86% participants of the total in DERM I, NOV, and DERM II dataset, respectively. Some participants and images were not considered for further analysis for reasons such as unacceptable calibration accuracy, accidental loss of eye movement or verbal data, and too much noise. The overall mean and standard deviation in the two directions for the three dataset after removing participants with poor calibration along with final number of images used is shown in Table 3.2.

We performed first-order descriptive analysis of the gaze and spoken description data for the DERM II dataset. Average fixation duration across the 26 observers was 320 milliseconds and average duration of narratives was about one minute. The manually

Table 3.2: Mean calibration accuracy after participants with poor calibration were removed for the three datasets (all values are in degrees) The last two columns of this table show the number of participants and images used in further data analysis.

Dataset	X Mean	X SD	Y Mean	Y SD	Participants	Images
DERM I	0.51	0.13	0.51	0.09	12 (75%)	50
NOV	0.63	0.29	0.70	0.14	12 (100%)	34
DERM II	0.71	0.16	0.81	0.23	26 (86%)	29

Table 3.3: Mean, standard deviation, minimum, and maximum number of word tokens, word types, and type-token-ratio across the 754 narratives (26 observers, 29 images) for the DERM II dataset. The high value of mean type-token ratio with a low value of standard deviation suggests high lexical diversity.

	Mean	SD	Min.	Max.
No. of Tokens	80	39	16	264
No. of Types	56	21	14	128
Type-Token Ratio	0.74	0.09	0.48	1

transcribed narratives were segmented into word tokens using the default NLTK word tokenizer. Various measures for the first-order analysis of the narratives were then calculated. Table 3.3 shows the mean number of word tokens, word types, and type-token ratio along with the standard deviation, minimum and maximum number of tokens and types. The mean number of tokens and the average duration of narratives together indicate that on average observers uttered 1.3 words per second, which indicates that the experts did not rush through the image inspection and description task. The mean type-token ratio of 74% suggests that there is significant lexical diversity across the dataset highlighting the richness of the dataset. Figure 3.3 shows a scatter plot for the mean number of word types against the mean number of word tokens for the 29 images. As expected, the plot is linear since higher number of tokens typically result in higher number of types. Images 10, 5, and 16, highlighted in green, have fewer mean word tokens and types than images 6, 23, and 17, highlighted in magenta. This is possibly because images 10, 5, and 16 have only one primary morphology with some associated secondary morphology whereas images 6, 23, and 17 have more than one primary morphology with their respective secondary morphologies. For example image number 10 has multiple *plaques* whereas image number 6 has *nodule, papules, erythema* thereby increasing the likelihood of higher mean type-token

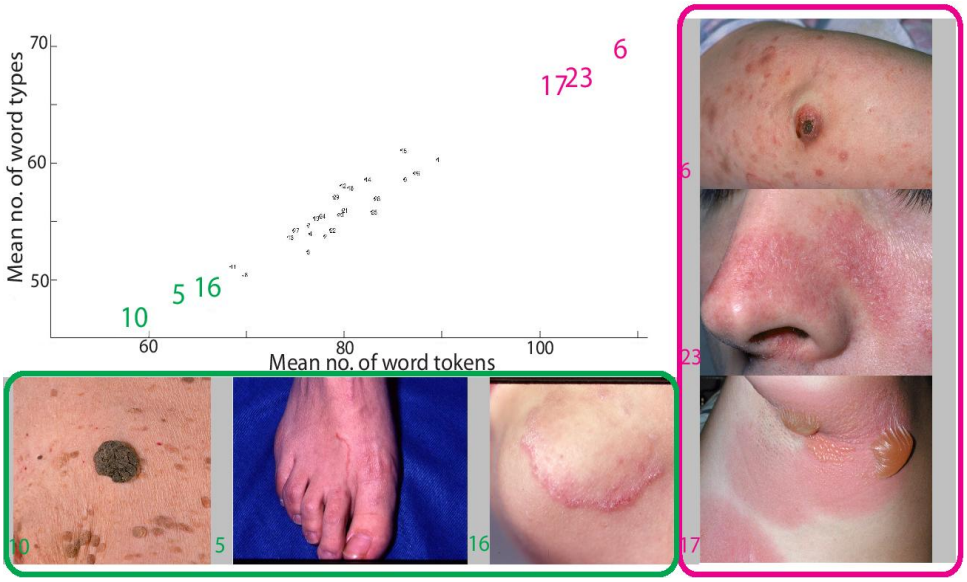

Figure 3.3: Scatter plot showing mean word types vs. mean word tokens for each image across all observers in the DERM II dataset. Each image is a data point. Highlighted images are shown at the bottom (green) and on the right (magenta).

ratio. Additionally, the experts were instructed to describe as if they would diagnose which could have also contributed to the high values of mean type-token ratio. Figure 3.4 shows the mean number of word tokens, word types, and type-token ratio for each observer across all the images. The high values of type-token ratio suggest lexical richness and heterogeneity present in the descriptions provided by the observers.

3.6 Other studies with DERM I, II, and NOV

Before embarking on multimodal alignment, the DERM I and NOV dataset were thoroughly explored.

The verbal data from the DERM I dataset led to some interesting work about uncertainty in physicians' narrative and the diagnostic correctness [McCoy et al., 2012b], understanding medical experts' reasoning processes [McCoy et al., 2012a], investigating disfluencies in descriptions as indicators of context dependency and cognitive reasoning [Womack et al., 2012], and explore units of thoughts in spoken medical narratives [Womack et al., 2013]. The gaze data from the DERM I and NOV datasets were used to model eye movement patterns of medical experts and novices during image inspection tasks

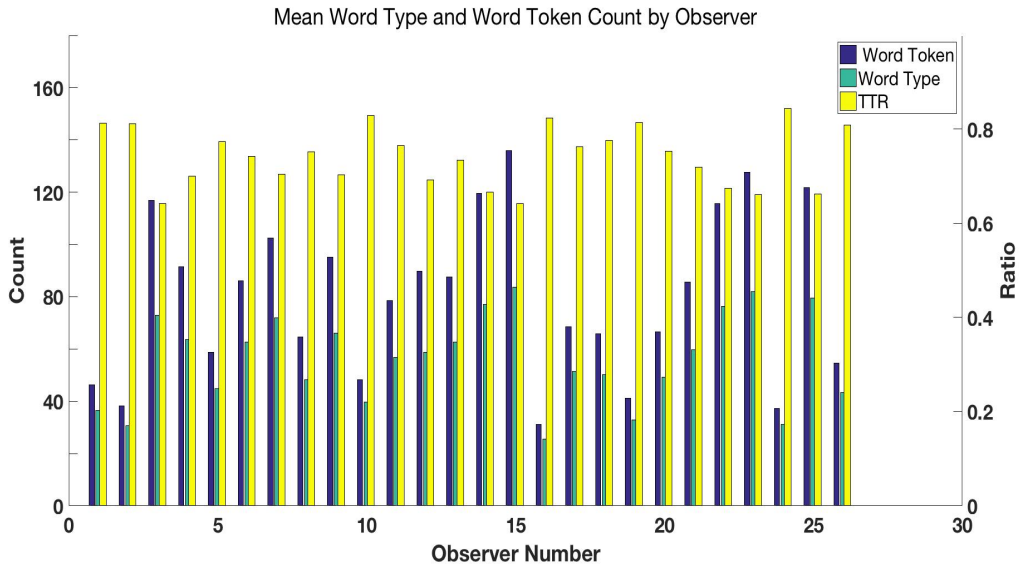

Figure 3.4: Bar plot showing the mean number of word tokens, word types, and type-token ratio (TTR) for each observer across the 29 images in the DERM II dataset. All the observers have a mean type-token ratio greater than 0.6 suggesting stronger lexical diversity.

and explore the differences [Li et al., 2012, Li et al., 2013, Li et al., 2016]. In addition, Guo et al. (2014a) used the gaze and verbal data to design a human-centered image retrieval application [Guo et al., 2014a]. Guo et al. (2014b) also studied the narrative data. The verbal data from the DERM II dataset was also used to shed light on physician decision making [Hochberg et al., 2014a] as well as automate the process of annotation of those styles [Hochberg et al., 2014b]. Bullard et al. (2014) used the DERM II dataset to model physicians' diagnostic confidence and self-awareness.

In the following subsections we discuss preliminary work performed using the two datasets that motivated the alignment-annotation framework. We first introduce some technical concepts that are helpful to understand the preliminary work and related findings.

3.6.1 Technical concepts

Fixation, Union and Intersection Maps: Fixation maps as defined in this work are 3D grayscale maps where two of the dimensions denote the x, y location of a fixation and a foveal region around it and the third dimension represents the duration of the

fixation and the surrounding region. Using BeGaze 3.1.117 software from the SMI package [Sensomotoric Instruments, 2016], *x, y* locations of fixations on the stimuli and their corresponding durations are obtained. These *x, y* locations indicate where on the stimuli the participant gazed at a point in time. To visualize this, a value of 1 is assigned to *x, y* pixel coordinates that were fixated and 0 to the rest of the image coordinates resulting in a binary *fixation plot*; see Figure 3.5(b). Since the fovea is not a single pixel but subtends a larger area, a simple binary plot would not be appropriate to represent the region over which visual information is acquired during the fixation. Therefore, to approximate the fovea, a 2D Gaussian kernel of size $\sigma_{horizontal} = 2, \sigma_{vertical} = 3$ degrees is convolved with the binary plot to yield a grayscale map representing regions of visual information [Wooding, 2002]. The intensity of the darkness of the regions are further weighted by the individual fixation durations and finally normalized by dividing each value by the maximum to range from 0 to 1. The resulting continuous heatmap overlaid on the original image as shown in Figure 3.5(c) is called the *fixation map*. The standard deviations in the two directions for the kernel are not the same because, generally in eye tracking, a participant's eyes tend to drift more in the vertical direction. Union maps are generated by adding every participant's fixations maps and normalizing, per image. These union maps illustrate pixels fixated by one or more participants. Likewise intersection maps are generated by taking the area in the union map shared by at least 80% of the participants. Most of the results discussed in the following sections were obtained using binary fixation maps unless stated otherwise.

Area Under the Curve (AUC): Receiver operating curves (ROC) can be used as a metric to evaluate how well one participant's fixations match another participant or a group of participants [Green and Swets, 1966]. In this method the fixation map of any participant or a group of participants is treated as a binary classifier on every pixel in the image. The fixations of the participant or group to be compared are used as ground truth. By varying the threshold, the ROC curve is drawn as the false positive rate vs. true positive rate and the area under this curve (AUC) indicates how well the fixation map ranks a ground truth fixation with values ranging from 0.5 (chance performance) to 1 (ideal performance).

CIELAB: CIELAB color space is an appropriate color space representation for dermatology images, due to the relation of the L and b components to melanin and of the a component to hemoglobin [Takiwaki, 1998]. In this color space the red, green, and blue color channels from the original RGB image are transformed to three channels: a

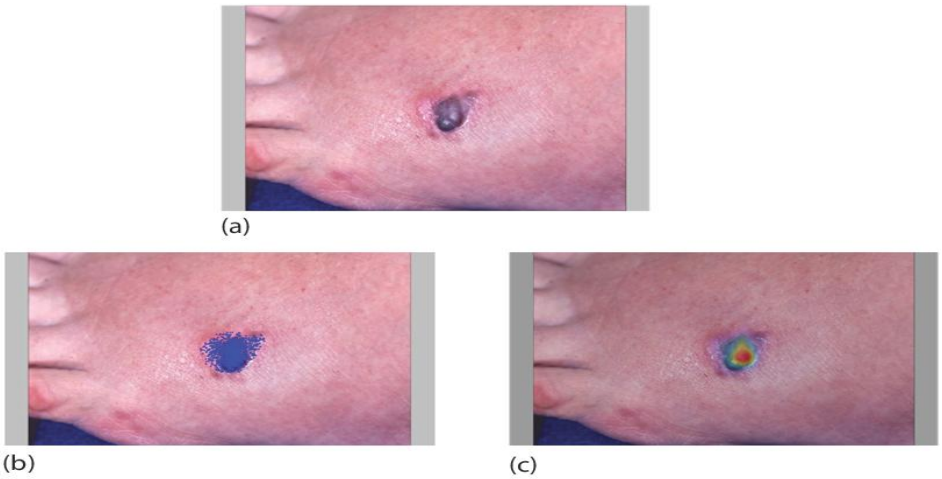

Figure 3.5: Panel (a) the original image showing melanoma on foot; Panel (b) participant's fixations (blue) overlaid on the image: the blue locations are assigned 1 and others 0 creating a binary map; Panel (c) the participant's fixations from Panel (b) convolved with a Gaussian to obtain a heatmap overlaid on the original image.

luminance channel L, a red-green opponent channel a, and a yellow-blue opponent channel b. *Lab* space has been widely accepted and shown to be effective for differentiating between lesioned and normal skin [Shin et al., 2002, Bosman et al., 2010].

Scale Invariant Feature Transform (SIFT): The SIFT algorithm extracts distinctive features in an image or video that are invariant to image scale and rotation [Lowe, 1999]. The image data is transformed into scale-invariant coordinates with respect to local features. Firstly, interest points that are invariant to scale and orientation are identified. In the next step, interest points are tested for stability and keypoints are selected based on the measures of their stability. Each keypoint is then assigned with one or more orientations based on the local image gradient directions. Finally, local image gradients around each keypoint are transformed into a set of descriptors that are invariant to shape distortion and range of illumination.

3.6.2 Importance of eliciting perceptual behavior of experts

To demonstrate why perceptual behavior would be useful, a comparison between dermatology experts and undergraduate novices using different metrics was conducted. For the NOV dataset, out of 42 images, 34 images had good eye movement data fit for further analysis. Therefore, eye movement data for the same 34 images were selected for

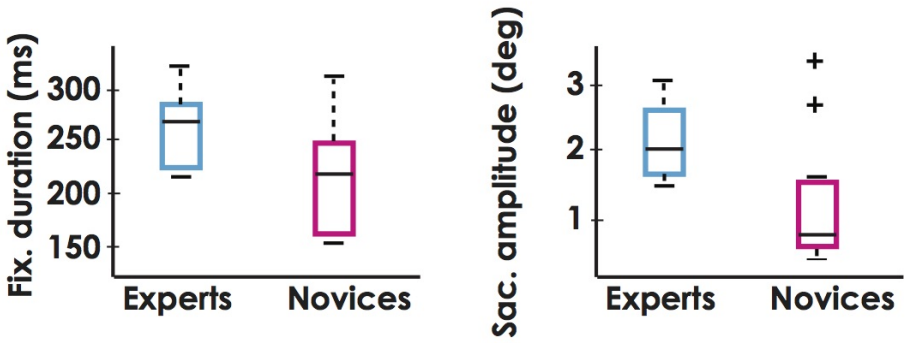

Figure 3.6: Box plots for median fixation duration and saccade amplitude for the experts and novices. Notice that experts tend to have longer fixation durations and saccade amplitudes compared to novices.

the DERM I dataset so that the analysis is performed on the same set of images. Both groups respectively comprised data for 12 participants.

For the experts and novices, the median fixation duration and saccade amplitude over all images and all participants were calculated separately. Figure 3.6 shows that for the novice group these two metrics were lower than for the expert group. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles. The whiskers extend to the most extreme datapoints that are not considered as outliers. The outliers are plotted individually. A two-tailed Student's t-test indicated that these differences were significant ($p < 0.05$).

While these measures indicate a difference between the two groups they are not strong enough to rule out the null hypothesis that the two groups are not different. Also, more than one type of comparison metric is required to compare different aspects of eye movement behavior [Riche et al., 2013]. For this purpose two other measures were used namely ideal area under the curve (iAUC) and recurrence quantification analysis (RQA). They are discussed below.

The ideal area under the curve (iAUC) measures the variability among participants. It is computed by measuring how well the fixations of one participant can be predicted by the fixations of the other *n-1* participants, iterating over all n participants and averaging the results for all the participants and images [Borji, 2009]. For example, if we assume that all participants have the exact same eye movement behavior then every participant will be a perfect match to every other participant and the iAUC will be 1. The concept

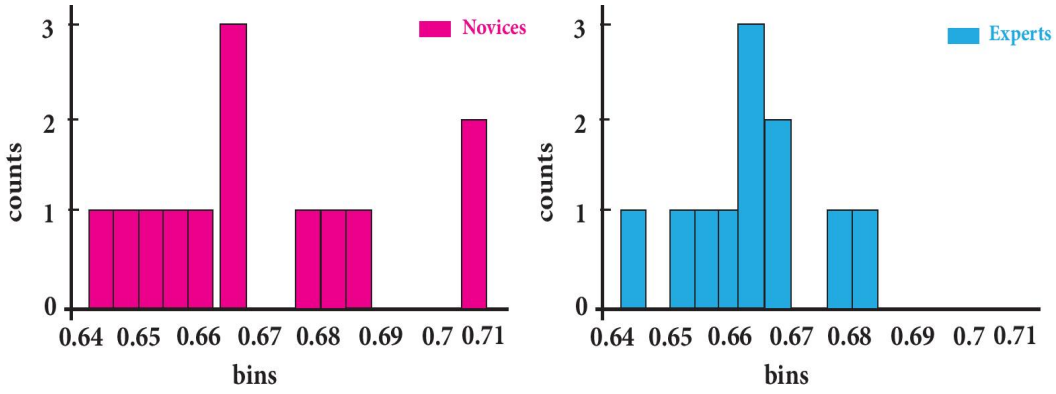

Figure 3.7: Histogram of ideal Area under the Curve (iAUC) values averaged over 34 images.

of iAUC was used in the following way to investigate the existence of differences between dermatology experts group (DERM I) and undergraduate novices group (NOV). Taking one group at a time, for a given image:

1. A grayscale Gaussian fixation map for a participant is generated at a time using the method described above. This is the Test Eye Map.

2. Next, the *x-y* fixation coordinates of the remaining *n-1* participants in the same expertise group are accumulated and a binary map is generated by assigning 1 to the locations specified by the accumulated x-y coordinates and 0 to the rest of the pixels. This is referred to as the Ground Truth Eye Map.

3. With the above two maps the AUC is calculated in the traditional way as described above for the participant. This is done for all the participants and an average over all the participants is calculated.

4. This is done for all the images in the dataset and the average across images is calculated resulting in a single value of iAUC for a group.

The iAUC values obtained for expert and novice group were 0.68 and 0.66 respectively. Figure 3.7 shows histograms of the iAUC values for the experts (right) and novices (left). A two-tailed Students t-test demonstrated that the difference of 2% between the iAUC values of experts and novices was significant ($p < 0.05$), indicating the experts are more likely to have eye movements similar to other experts as opposed to novices and vice-versa.

None of the above measures take into account the crucial temporal order of the eye movement sequences (referred to as fixation patterns). The temporal order differences in global and local temporal fixation patterns between the two groups were explored using recurrence quantification analysis (RQA). Classical RQA is a technique to investigate the time evolution of data series widely used in describing complex dynamic systems [Webber and Zbilut, 1994]. Recently cross-recurrence analysis has been used to investigate the coupling between speakers' and listeners' eye movements [Richardson and Dale, 2005]. The RQA method and measures have been used to investigate the differences in the spatial and temporal characteristics of expert and novice eye movement behavior [Anderson et al., 2013]. A brief description of the method that takes fixation duration into account is provided below.

For a fixation sequence f_i and corresponding durations $t_i, i = 1, \ldots N$, two fixations (i, j) are recurrent if they are within certain distance of each other and a recurrence plot (visualization technique) is created by assigning the sum of the corresponding durations to the position i, j:

$$r_{ij} = \begin{cases} t_i + t_j, & d(f_i, f_j) \leq \rho. \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

where d is the distance metric and ρ is the radius, i.e. the maximum distance between two fixations to be considered recurrent. Distance can be defined in various ways. This study used Euclidean distance with radius $\rho = 64$ pixels, approximately 1.5° visual angle for our experimental setup. The value approximates the size of the fovea and tracker error in the employed eye tracker. For calculations only the upper triangle is taken into account since the recurrence plot is symmetric and the diagonal does not provide additional information.

These plots provide useful visualization of the temporal behavior of a participants' eye movements. The four RQA measures used by Anderson et al. and explored in this work are: *recurrence*, *determinism*, *laminarity* and *center for recurrence mass*. The recurrence and center for recurrence mass measures are rather global temporal fixation sequences whereas local patterns are captured by determinism and laminarity. The sum of recurrences in the upper triangle is defined as $R = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} r_{ij}$, and $T = \sum_{i=1}^{N} t_i$ is the sum of the fixation durations used for normalization purposes. Each RQA measure quantifies a certain aspect of the fixation sequence and is defined as:

Recurrence (REC): This measure can be thought of as representing (in percent) how

often a location is refixated.

$$REC = 100\frac{R}{(N-1)T} \tag{3.2}$$

Determinism (DET): Determinism measures how often participants repeat short subsequences in their overall fixation sequence. Recurrent points in the plot can form diagonal lines (D_L) that indicate repetition of short subsequences. For example if a participant looks back and forth between two locations creating a repeated pattern, those fixations would constitute a diagonal line. The reported results were calculated using L = 2 (other line lengths showed similar results).

$$DET = \frac{100}{R}\sum_{(i,j)\in D_L} r_{ij} \tag{3.3}$$

Laminarity (LAM): Recurrent points can also form vertical (V_L) and horizontal (H_L) lines. Since the plot is symmetrical, vertical and horizontal lines in the upper half of the plot are the same as horizontal and vertical lines in the bottom half, respectively. A vertical line (upper half) indicates detailed rescanning of a location that was previously fixated with a single fixation. On the other hand, a horizontal line (upper half) shows brief refixation to a location that was previously scanned in detail with multiple fixations. Together the horizontal and vertical lines are used to calculate what is called *laminarity* representing revisited locations in the scene.

$$LAM = \frac{100}{2R}\left(\sum_{(i,j)\in H_L} r_{ij} + \sum_{i,j\in V_L} r_{ij}\right) \tag{3.4}$$

Center of recurrence mass (CORM): This measure quantifies the temporal distribution of the recurrent points. A small CORM value would mean that most of the refixations occurred very close in time whereas a large CORM value shows that refixations were widely separated in time.

$$CORM = 100\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(j-i)r_{ij}}{(N-1)^2T} \tag{3.5}$$

The left panel in Figure 3.8 shows an example of a dermatology image overlaid with a participant's fixations and the right panel shows the corresponding recurrence plot. Using equations described above we obtained 12 participants × 34 images recurrence plots for the two groups and the four RQA measures. Wilcoxon rank-sum test with $p = 0.05$ was used for significance testing to deal with the non-normal nature of the data.

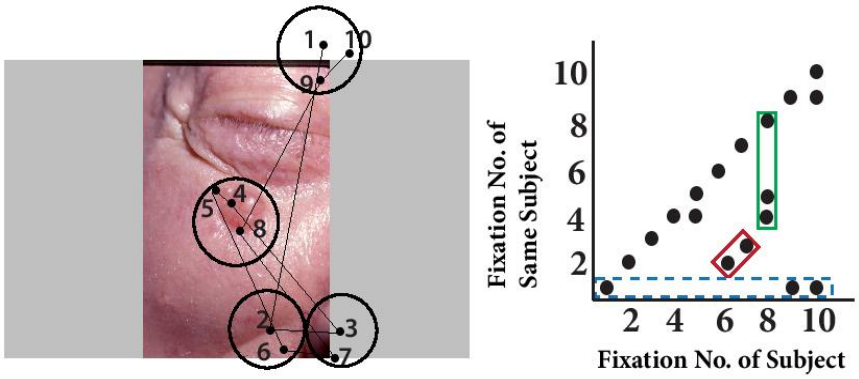

Figure 3.8: Left: Hypothetical fixation sequence overlaid on the image to illustrate the RQA method. Numbers represent fixation order; circles represent a radius of 64 pixels. Right: Recurrence plot for the scanpath shown on the left. The black squares represent recurring fixations which means they were within 64 pixel radius of each other. Examples of diagonal line for *determinism* (solid green box) and of horizontal and vertical lines for *laminarity* (dotted red and dashed blue boxes) are indicated.

Recurrence: Significant difference between experts and novices was observed with recurrence for experts being lower than recurrence for novices as shown in Figure 3.9. This shows that expert dermatologists tend to refixate previously inspected areas less often than novices suggesting that perceptual expertise probably helps experts to quickly obtain the required information pertaining to a region thereby requiring less rescanning.

Determinism: The results suggest that experts repeat short sequences of fixations less often than novices. The rank-sum test indicates that the difference is significant and that determinism is higher among novices.

Laminarity: Experts were observed to be significantly lower in laminarity than were novices, indicating that they had fewer instances of repeated fixations within a relatively small region (defined by ρ).

Center of recurrence mass: CORM values among experts were significantly higher, indicating that experts refixated regions after longer intervals than did the novices. A probable reason is that experts fixate regions at the beginning of the trial and then revisit those regions towards the end when confirming their final diagnosis [Li et al., 2012].

Figure 3.9 shows that experts had lower recurrence, determinism, and laminarity. This suggests that experts are able to weigh a region's importance after a brief fixation, while novices exhibit multiple refixations. This could mean that experts use their perceptual expertise to guide their gaze to maximize information intake. The high recurrence value

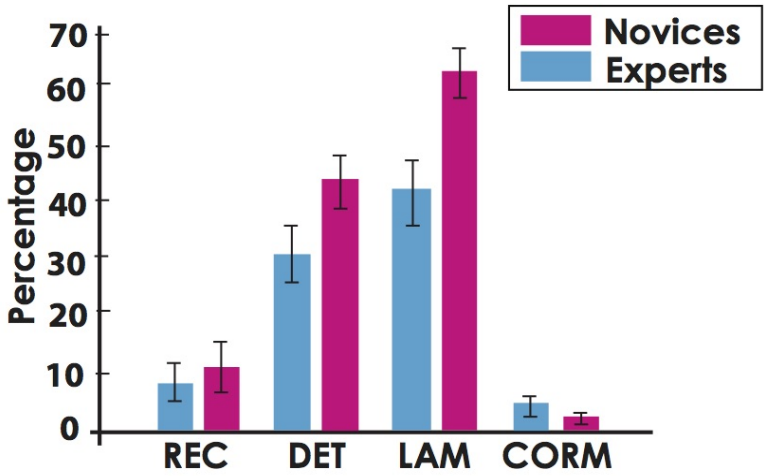

Figure 3.9: Comparison of RQA measures between experts and novices: recurrence, determinism and laminarity are significantly lower for experts than novices; center of recurrence mass was higher for experts. These results indicate that experts refixate or repeat their scanpaths less often and that most of their refixations occur widely separated in time.

along with higher number of fixations per second for the novice group suggests that novices are quickly scanning the scene with less strategy thereby having low fixation duration for individual fixations. On the other hand experts have longer fixation durations meaning they spend enough time on individual fixations to extract the useful information. This supports the low values of RQA measures except for CORM indicating involvement of different type of perceptual strategy by experts in comparison to novices. When viewing dermatology images the high values of CORM could mean that experts initially inspect regions that are most informative or important, then fixate regions that might help them further in their diagnostic path followed by refixations to confirm their inferences. The sensitivity of the analysis to the parameters L and ρ was also tested. The significance tests were unaffected by variations in these parameters.

The number of fixations per second among the individuals can affect the RQA measures. In our work novices ($\mu = 4.5$, SE $= 0.02$) had significantly higher number of fixations per second than experts ($\mu = 3.6$, SE $= 0.02$). The bootstrap technique [Anderson et al., 2013] was used to test if the observed differences were purely by chance. All the RQA measures were significantly different for both the groups from those for random fixation sequences indicating the group behaviors were not random. Motivated by

these results differences in the RQA measures between attending physicians and in-training residents were examined. Due to unequal sample size an iterative test was conducted by comparing the 3 residents with 3 randomly selected attendings. Differences observed were not significant for all the iterations and depended on the individual attending. Larger sample size and statistically stronger tools are required to validate if differences exists between these two groups.

These results suggest that perceptual behavior of experts in a domain can provide additional cognitive information relevant to the user's end goal and benefit the image understanding system, which is the ultimate aim of this work. These results were published by Vaidyanathan et al. [2014].

3.6.3 Multimodal asynchrony

Researchers have investigated how linguistic and visual information are integrated during language processing [Ferreira and Tanenhaus, 2007, Ferreira and Henderson, 2004, Holsanova, 2006]. An observation from prior research is that people do not verbally mention an object's name at the same time as they look at it [Meyer et al., 1998, Griffin, 2004]. To understand if this asynchrony exists in our visual-verbal dermatology dataset DERM I and what factors might influence it we analyzed a subset of eye movement data in the following way.

The subset comprised of eye movement data for dermatologists inspecting 12 dermatological images. An expert dermatologist selected the 12 images with 6 images that were easy to diagnose and 6 images that were difficult to diagnose. For each image, two concrete, frequent clinical attributes, namely *primary morphology* (e.g. *papule*) and *secondary morphology* (e.g. *scale*) were selected for analysis. Automatic identification of the true image regions that represent these clinical attributes is a challenging task. Therefore, to identify these regions accurately manual annotations were used in this case. An expert dermatologist marked the image regions that depicted each attribute. The asynchrony was calculated as the difference between the *first time* ($T_{1^{st}gaze}$) the physician gazed in the marked region and the *first mention* ($T_{1^{st}attribute\_mentioning}$) of the corresponding attribute (Figure 3.10 (top)).

The analysis showed that gaze in the relevant region preceded the verbal reference to it in all cases indicated by the positive values of the temporal asynchrony, i.e. participants consistently looked at the lesion prior to verbally mentioning it. Two one-way between participants ANOVA was conducted to compare the effect of the type of clinical attribute

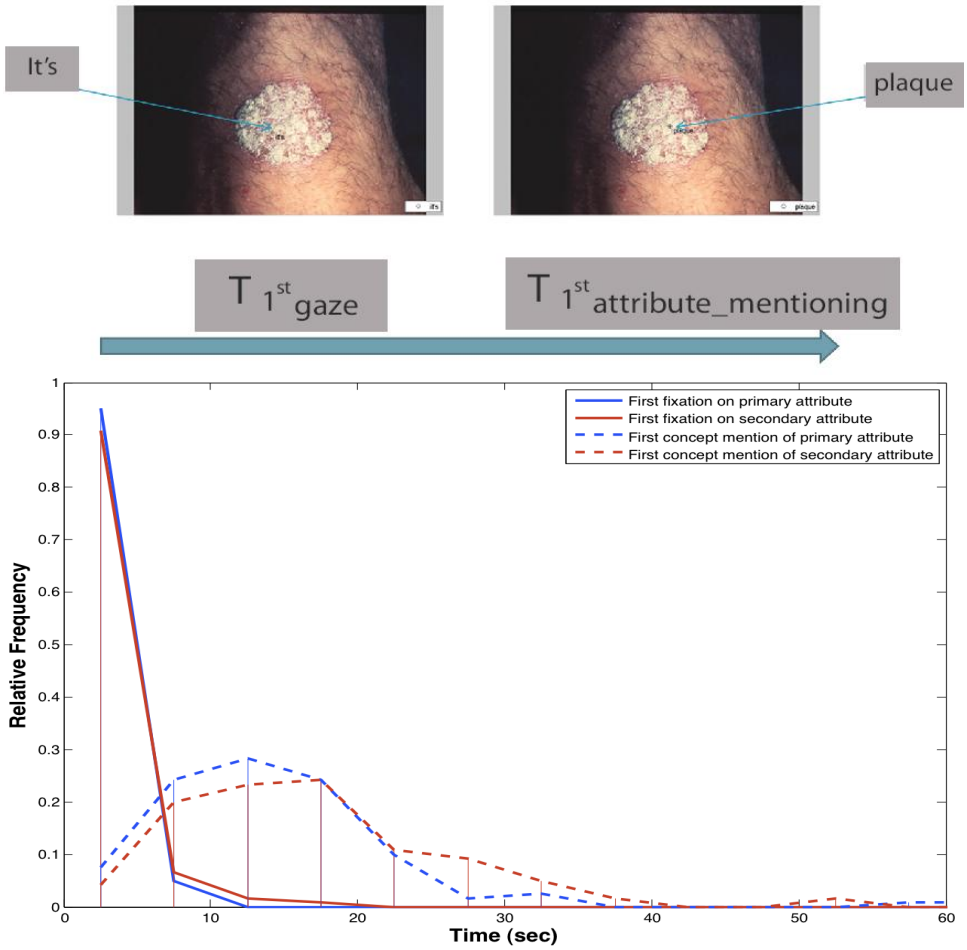

Figure 3.10: Top: Example showing how time of first gaze and first attribute mentioning (utterance) can be used to calculate *Asynchrony*; Here when the first fixation is executed the word *It's* is uttered. At a later point in time the word *plaque* is uttered. Bottom: Histogram in the form of line plot for first fixation and for the first attribute mentioning binned over time scale.

and image complexity on the asynchrony. There was a significant effect of both the type of clinical attribute and image complexity on the asynchrony at the $p < 0.05$ level. Also, the primary attribute showed a shorter time-lag than the secondary attribute suggesting that dermatology experts name the primary attribute prior more quickly than they name the secondary attribute. This further indicates that such variables should be taken into account when modeling the temporal relation between the two modalities.

Additionally, participants spent on average 3 seconds inspecting the image before they began to talk. This suggests that physicians might be trying to obtain an initial holistic view of the image and plan their speech prior to execution. Figure 3.10 (top) shows how the asynchrony measure can be calculated. Figure 3.10 (bottom) illustrates that physicians looked at the regions depicting the two attributes at about the same time but mention the primary attribute prior to mentioning the secondary attribute. This could be because of how they were taught to perform the diagnoses.

These results bring into focus various factors that affect the temporal relation between visual and linguistic information processing. This demands further investigation into mechanisms to fuse information from the two modalities. These results were published by Vaidyanathan et al. [2012, 2013].

3.6.4 Image processing algorithms and multimodal data

We also performed a qualitative comparison of the well-known local feature descriptor called Scale Invariant Feature Transform (SIFT) with the collected fixation data. The union and intersection plots of physicians' visual fixations for two images from the DERM I dataset are shown in Figure 3.11. The union plots were quite similar to the local feature plots. The intersection plots, on the other hand, provided ROIs that are potentially most diagnostic. Data indicate that physicians, due to image center-bias which is an observer's tendency to look at the center of an image [Tatler, 2007], fixate near the center of images for the first few fixations. However, they are very quickly drawn towards the regions of interest. There is a strong connection between lesion location and the eye movements suggesting that during image inspection the effect of viewer center bias is drastically reduced. Results from this analysis suggest that SIFT feature descriptors have moderate amount of overlap with eye movement locations and can potentially be useful in the identification of regions of interest when dealing with large datasets.

It would be beneficial to develop image processing algorithms that are able to extract perceptually important regions. Although SIFT descriptors provided a close approximation

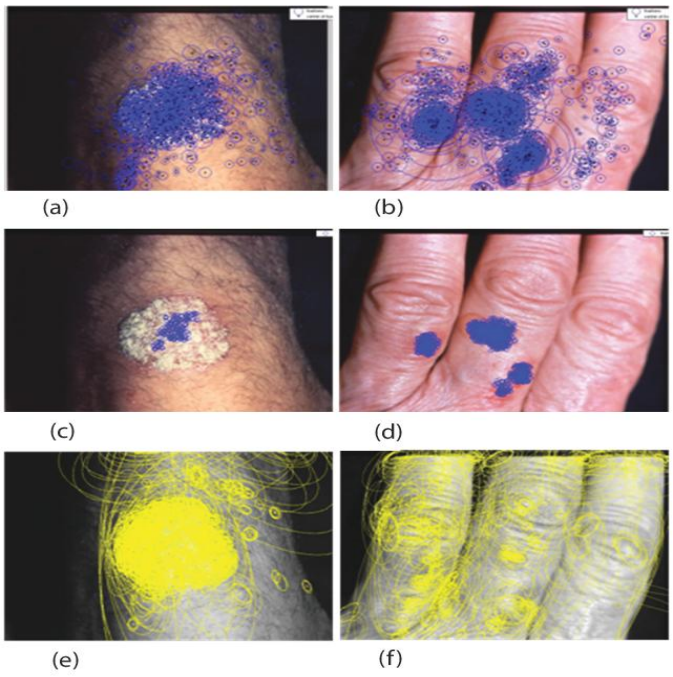

Figure 3.11: Illustration of (a, b) Union of all participants' fixations; (c, d) intersection of 80% of all participants' fixations; (e, f) SIFT plots for the two cases of psoriasis and pemphigus vulgaris, respectively.

to eye movements, the number of these descriptors is huge; thus gaze data can be used to filter unwanted descriptors. In an attempt to achieve this, a correlation between perceptually-relevant image regions obtained through eye movements and individual clusters of image regions identified through k-means clustering was investigated. Judging the segmentation output for various values of k visually, a value of $k = 4$ was selected because higher values lead to oversegmentation. CIELAB color space and the data from the DERM I were used. To measure the correlation between participants' gaze and the segmented image regions, a metric called fixation ratio was defined. This metric was useful in capturing the cluster that would most effectively segment the primary region (lesion) in the image. For each image we calculated the fixation ratio as follows:

1. The original *RGB* images were converted into CIELAB and a-b vectors were used as input for the k-means algorithm, dividing each image into 4 clusters. This generated a segmentation map for each of the 50 images, each with four clusters.

2. The intersection map for each image was overlaid on the corresponding segmentation

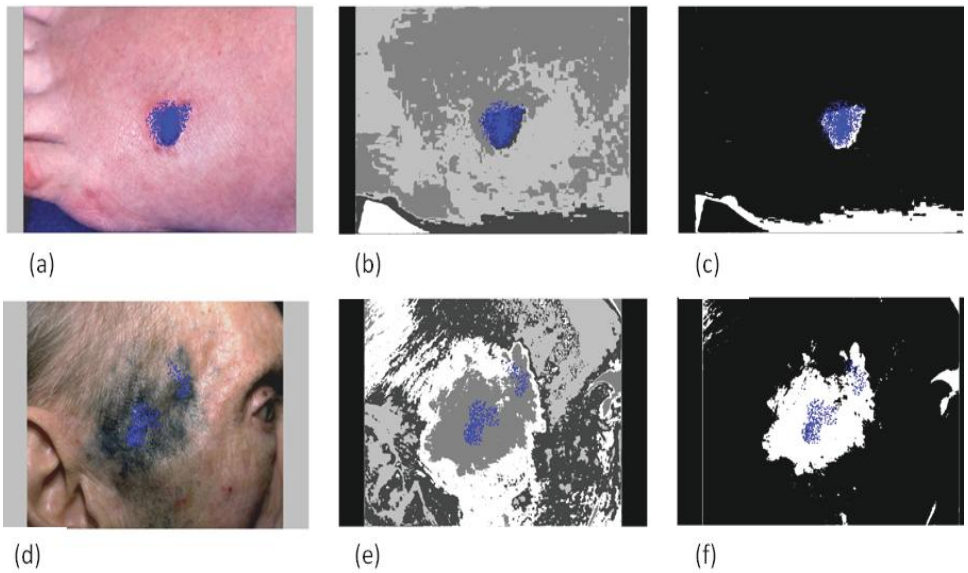

Figure 3.12: Panels (a, d) fixations (blue dots) overlaid on original images. Panels (b, e) fixations (blue dots) overlaid on the segmentation maps of panel (a) and panel (d) respectively. Panels (c, f) cluster picked using the fixation ratio measure.

map to obtain the number of fixations falling in each cluster.

3. These fixations were normalized by the total number of fixations in the intersection map, generating relative fixations per cluster.

4. Similarly, the relative area per cluster for the segmentation map was obtained.

5. Fixation ratio for every cluster was obtained by dividing the relative fixation from step 3 by relative area from step 4.

Visualization of the intersection fixation data overlaid on the segmentation map as shown in Figure 3.12 illustrates that k-means was effective in isolating the primary region (lesion) with high visual interest to the physicians. The high relative fixation on one of the clusters shown in Figure 3.12 and low relative area resulted in a high fixation ratio that also provided a quantitative measure to select this cluster as the most perceptually relevant.

Figure 3.13(a) shows the fixation ratio value of every image in DERM I. From these values it is evident that many images score higher than the average fixation ratio of 3.38. Twenty-two out of 50 images had a value higher than the average indicating the usefulness of k-means. Figure 3.13(b) compares the fixation ratio metric using the two types of

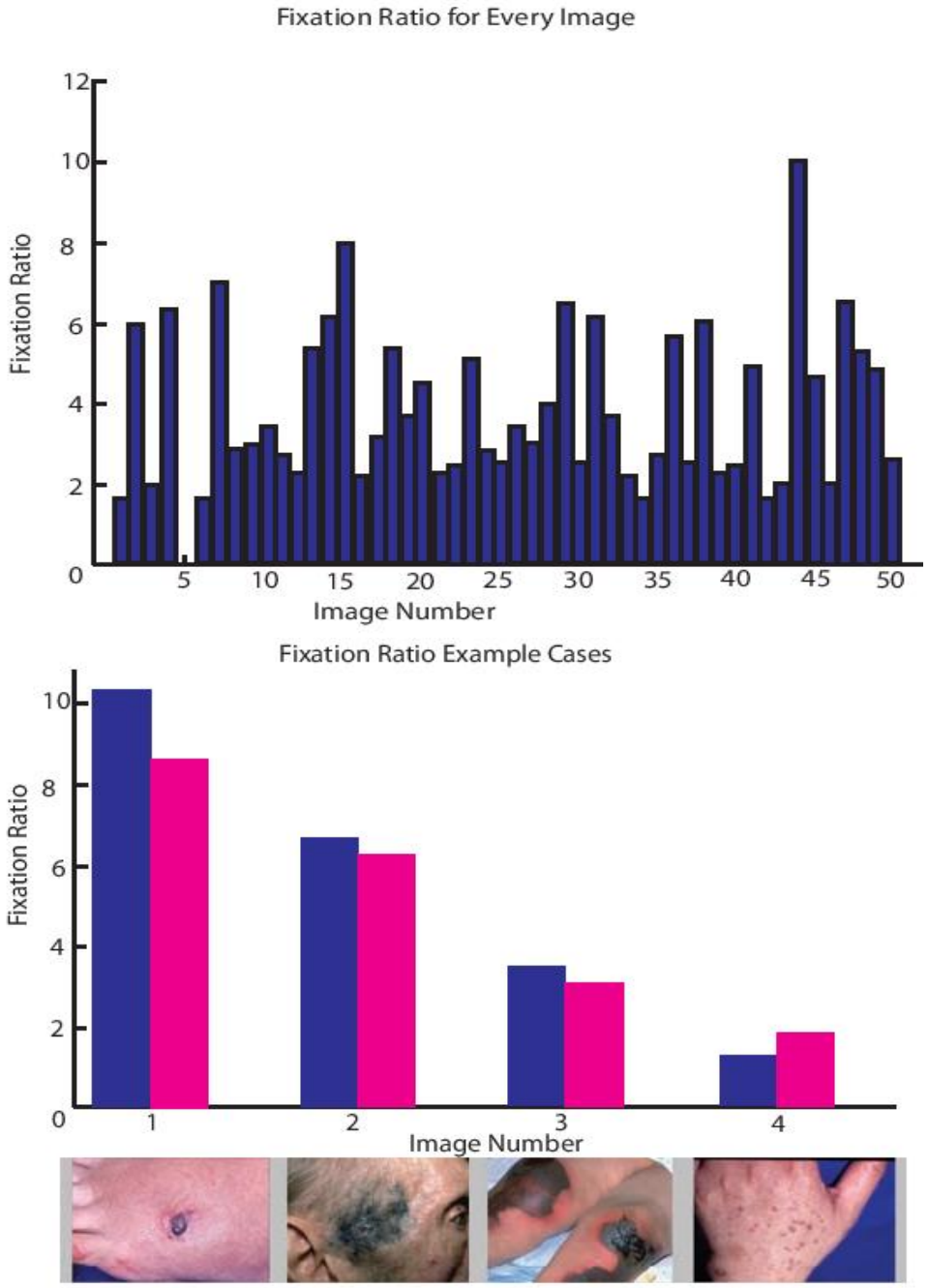

Figure 3.13: (a) Graph showing the fixation ratio for every image; (b) Comparison of fixation ratio generated using the binary (non-foveal) and grayscale (foveal) fixation map for four example images.

fixation maps, one with no involvement of a Gaussian function and the other generated using a Gaussian function. Thus clusters resulting from k-means and SIFT descriptors in conjunction with gaze data can be used to identify perceptually important image regions. Additionally, the fixation ratio method can be used to determine the degree to which existing image processing algorithms can capture the diagnostically relevant image regions. These results were published by Li et al. [2010] and Vaidyanathan et al. [2011].

Therefore, the preliminary work using DERM I and NOV dataset shows that capturing perceptual behavior of experts is important and can help bridge the semantic-gap between perceptually important regions of interests and regions extracted by image processing algorithms. The underlying foundation of the present work is that end users possess special perceptual knowledge and that eye movements and spoken description can provide insights into this expertise or information, which can help image annotation. Additionally, the results indicate the existence of a temporal asynchrony between gaze and speech. The fundamental question that arises now is: What is the relationship between eye movements and spoken description during image inspection? These results motivated the proposed multimodal integration framework and suggested that the relationship depends on various other factors.

3.7 Summary

To sum up, the preliminary result obtained from DERM I and NOV datasets illustrate that experts behave differently. The work was then extended to collect the DERM II dataset. These results motivated us to investigate the asynchrony between visual and linguistic data and integrate them using techniques such as alignment for the purpose of image region annotation. To comprehensively study how bitext alignment can be applied to image datasets of varying scope, we explore its use both on expert domain images (DERM II) and on general-domain images. Thus, in addition, we collected SNAG, a dataset involving general-domain images. This dataset is a unique resource and contribution of this work and is discussed in detail in the next chapter.

4

SNAG: Spoken Narratives and Gaze Dataset

4.1 Motivation

To establish the generalizability of our framework, we must apply it both to specific-domain images (e.g. DERM II) and general-domain images. Thus, we collected SNAG to help investigate whether the multimodal framework would (1) apply to any type of image including general-domain images and (2) be scalable to a larger dataset To limit the vocabulary used for descriptions we recruited native speakers of American English. Additionally, we used automatic speech recognition tools to eliminate human intervention in the framework. Importantly, this dataset is being released to the larger research community, making it the first publicly available dataset that consists of co-captured gaze and spoken image description data.

4.2 Gaze-verbal data collection for general users

Data collection involved 40 native speakers of American English, ranging in age from 18 to 25 years, viewing and describing 100 general domain images. The general-domain images were selected from a larger open-source dataset called MSCOCO (Microsoft Common Objects in Context) [Lin et al., 2014] which consists of more than 300,000 images. Some images from MSCOCO used in our data collection are shown in Figure 4.1. Widely used by the computer vision community, the MSCOCO dataset was created by pooling images from

Figure 4.1: Example images from MSCOCO used in the data collection process. The images vary in number of objects, scale, lighting, and resolution posing challenges to the alignment framework.

various sources, such as Flickr, and crowdsourcing them to obtain segments and captions. The crowd workers were provided with a list of object categories such as *cars* for which they had to identify the object's location in the image and draw its outline. They were also asked to provide short captions for the entire image. The images represent complex everyday scenes containing common objects in their context. For our dataset the primary researcher selected images so that typically they would depict an event with at least one initiator of the event and one target of the action, often respectively known as the *agent* and the *patient* in linguistic semantic role labeling. Of the selected 100 images, 69 images clearly depict at least one event whereas remaining 31 images may not necessarily represent an event. The MSCOCO images vary in number of objects, scale, lighting, and resolution as exemplified in Figure 4.1.

Participants were recruited campus-wide from Rochester Institute of Technology. The participants were given cookies and either a chance to enter a raffle or course credits for their participation. Gaze data was collected using SensoMotoric Instruments (SMI) RED 250Hz eye-tracker attached to a display as shown in Figure 4.2. The reported accuracy of the RED 250 eye-tracker is 0.5 degree. It is a non-intrusive and remote eye-tracker that monitors the participants' gaze. Each image was presented to the participant on a 22-inch LCD monitor (1680×1050 pixels) located approximately 68 cm from the participant. At 68 cm, the full display subtends 38×22 degrees of visual angle. We use a double computer set-up with one computer used to present the image and the other used to run the SMI software iViewX gaze tracking system and Experiment Center 2.3. After each stimulus, a blank gray slide was inserted to ensure that the gaze on the previous stimulus did not affect the gaze on the following stimulus. The blank gray slide was followed by a test slide

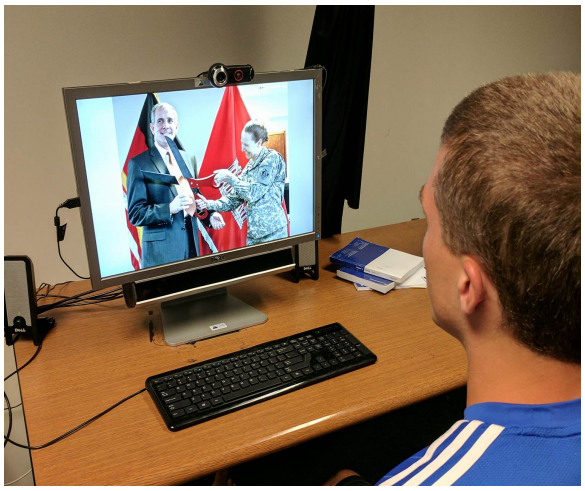

Figure 4.2: Data collection set-up used for the SNAG dataset experiment. The SMI eye-tracker that records the gaze data is attached underneath a display that displays the stimuli. The participant wears a lapel microphone connected to a TASCAM recorder that records the spoken descriptions. The task requires the participant to describe the action in the image to the experimenter.

Table 4.1: Sample raw data as obtained from SMI eye tracker showing from left to right: system timestamp, left-eye horizontal and vertical fixation locations, right-eye horizontal and vertical locations, left-eye and right-eye event, respectively.

Time	L_x[px]	L_y[px]	R_x[px]	R_y[px]	L Event	R Event
7456470899	550.0	406.07	550.0	406.07	Fixation	Fixation

with a small, visible target at the center with an invisible trigger area of interest. Using the test slide we could measure the drift between the location of the target at the center and the predicted gaze location over time that may have occurred due to the participants' movements. A TASCAM DR-100MKII recorder with a lapel microphone was used to record the spoken descriptions. A validation was performed every 10 images and re-calibration applied if the participant's validation error was more than one degree. To approximate the Master-Apprentice data collection method that helps in eliciting rich details, participants were instructed to describe the event in the images to the experimenter. The participants were instructed to "describe the action in the images and tell the experimenter what is happening." Participants were given a mandatory break after 50 images and otherwise smaller breaks if needed to avoid fatigue.

there's a female cutting a Kate
uh she's smiling and has sunglasses on her head
uh the cake has a picture of uh don't know who
also uh an iron man cake
and alcohol maybe champagne
uh she is wearing a black tank top
uh there are plates and other things on the table
and they seem to be in a bar or something

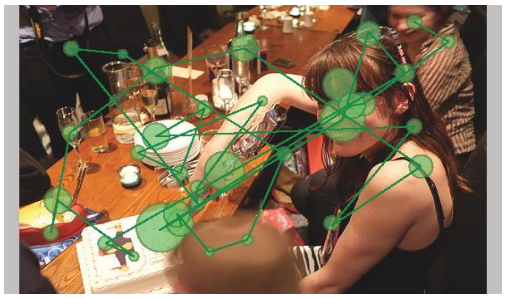

Figure 4.3: Example of multimodal data. On the left is the automated transcription of a participant's spoken description. We can see that apart from the incorrect transcription of *Kate* as opposed to *cake*, the ASR transcription performs quite well. On the right is the eye movement data for the same participant overlaid on the corresponding image. The green circles show fixations with the radius of the circles representing the duration of fixation. The green lines connecting two fixations represent saccades.

Table 4.2: Comparison of mean calibration accuracy for the four datasets. The SNAG dataset is comprised of approximately the same number of participants as the DERM II dataset but consists of more than three times the number of images used in the DERM II dataset.

Dataset	X Mean	X SD	Y Mean	Y SD	Participants	Images
DERM I	0.51	0.13	0.51	0.09	12 (75%)	50
NOV	0.63	0.29	0.70	0.14	12 (100%)	34
DERM II	0.71	0.16	0.81	0.23	26 (86%)	29
SNAG	0.67	0.25	0.74	0.27	30 (75%)	100

4.3 Fixations, narratives, and data quality

The SMI software package BeGaze 3.1.117 with default parameters and a velocity-based (I-VT) algorithm was used to detect eye-tracking events. An example of the detected fixations is shown in Table 4.1. Figure 4.3 shows an example of the scanpath, i.e. fixations (green circles) and saccades (green connecting lines) of an observer overlaid on the corresponding image. The alignment framework exclusively uses fixation data. Nine participants had a mean calibration and validation accuracy greater than two standard deviations in either horizontal or vertical direction. One participant had partial data loss. These 10 participants were removed. The mean calibration accuracy for this dataset is reported and compared to other datasets in Table 4.2. The corpus size is 3000 instances of image descriptions (100 images $\times$ 30 participants), with 13 female participants and

there's a female cutting a **Kate**
uh she's smiling and has sunglasses on her head
uh the cake has a picture of uh don't know who
also uh an iron man cake
and alcohol maybe champagne
uh she is wearing a black tank top
there are plates and other things on the table
uh they seem to be in a bar or something
next

but since we as female wearing sunglasses
that's **intraparty**
uh she's holding a knife
uh wrapped in some sort of uh foil
she seemed to be **fighting** a cake
uh **lastly** for a birthday party
uh next

Figure 4.4: Examples of the transcribed speech for two participants obtained using IBM Speech-to-Text tool. The descriptions belong to the image shown in Figure 4.3. While the narrative on the left from Figure 4.3 has one incorrectly transcribed word (*Kate* where the correct word is *cake*) highlighting that using automated transcription can save manual labor, the narrative on the right shows the limitations of ASR use with more word transcription errors. For the narrative on the right, the correct transcription for the words *but since we as, intraparty, fighting, lastly* are *there seems to be a, in a party, cutting, possibly* respectively.

Table 4.3: Mean, standard deviation, minimum, and maximum number of word tokens, word types, and type-token-ratio over 3000 narratives (30 observers, 100 images) for the SNAG dataset. The high value of mean type-token ratio indicates higher lexical diversity.

	Mean	SD	Min.	Max.
No. of Tokens	55	31	5	295
No. of Types	38	17	5	132
Type-Token Ratio	0.75	0.11	0.41	1

17 male participants. The speech recordings for the 30 participants for 100 images were machine-transcribed using the cloud-based IBM Watson Speech-to-Text service, an Automatic Speech Recognition (ASR) system accessible via a Websocket connection [IBM, 2015]. Example output is shown in Figure 4.3 (left panel). Figure 4.4 shows a additional comparison of output from the IBM Speech-to-Text tool for two observers for the same image as in Figure 4.3 (right). The transcription in Figure 4.4 (left) highlights that modest transcription errors may be accepted given the substantial reduction in the manual labor involved in speech transcription for large datasets. However, transcribed output on the right in Figure 4.4 shows a number of transcription errors thereby indicating that there is still room for improvement in applying ASR in the framework. All of the spoken descriptions for a subset of 5 images from the SNAG dataset were manually corrected

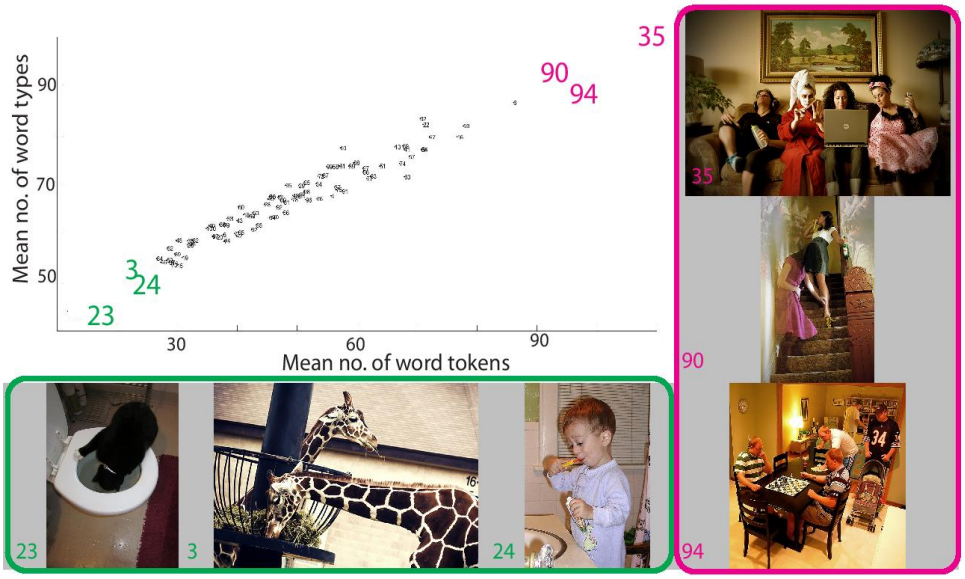

Figure 4.5: Scatter plot showing mean word types vs. mean word tokens for each image across all observers. Each image is a data point. Highlighted images are shown at the bottom (green) and on the right (magenta).

using Praat [Boersma, 2002] to be able to empirically explore the utility of substituting automatically generated transcriptions for careful but laborious manual transcriptions.

First-order descriptive analysis of the gaze and narratives show that the average fixation duration across the 30 participants was 250 milliseconds and average duration of narratives was about 22 seconds. Both these values are lower in comparison to that of the DERM II dataset. The IBM transcribed narratives were segmented into word tokens using the default NLTK word tokenizer. Various measures for the first-order analysis of the narratives were then calculated. Table 4.3 shows the mean number of word tokens and word types, and mean type-token ratio across all the 3000 narratives (30 participants, 100 images) along with the standard deviation, minimum and maximum number of tokens, types, and type-token ratio. The mean number of tokens and the average duration of narratives together suggest that on average observers uttered 2.5 words per second. This value is higher than for the DERM II dataset. The mean type-token ratio of 75% in Table 4.3 suggests that there is significant lexical diversity across the dataset supporting the richness of the dataset. Figure 4.5 shows a scatter plot for the mean number of word types against the mean number of word tokens for the 100 images. The plot is linear since higher number of tokens typically result in higher number of types. Images 23, 3, and 24, highlighted in

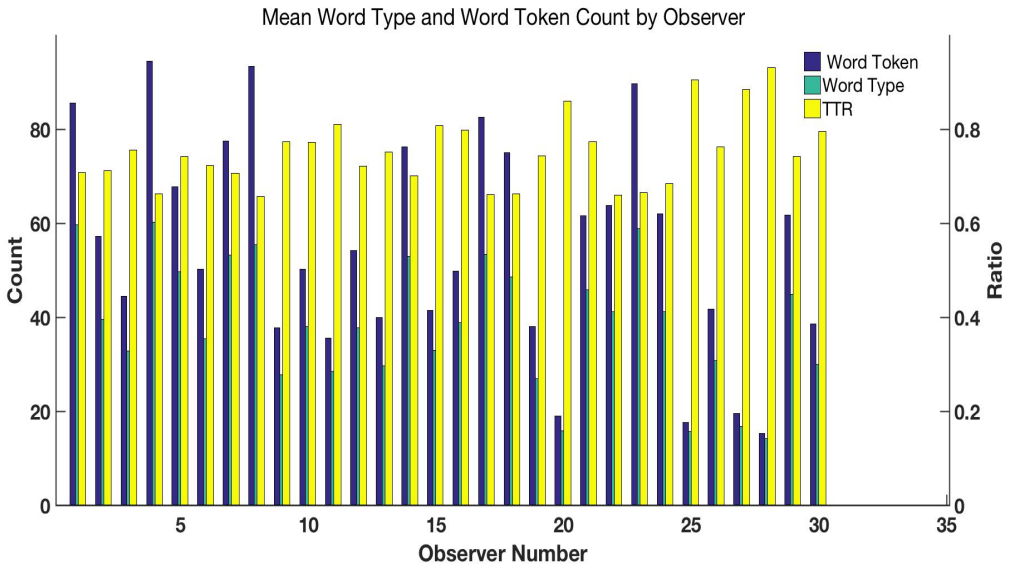

Figure 4.6: Bar plot showing the mean number of word tokens, word types, and type-token ratio (TTR) for each observer across the 100 images. All the observers have a mean type-token ratio greater than 0.6 suggesting stronger lexical diversity. Observer number 28 has the highest mean type-token ratio.

green, have fewer mean word tokens and types than images 35, 90, and 94, highlighted in magenta. For this dataset, this may be due to the number of significant objects in the images where a significant object is defined as an object that occupies a significantly large area of the image. Images 23, 3, and 24 have on average two objects while images 35, 90, and 94 have more than two. Comparing the two extremes image number 23 has two significant objects (*cat, toilet*) whereas image number 35 has around five objects (*female 1, female 2, female 3, female 4, laptop*). The number of significant objects together with the task instruction may have resulted in the distribution obtained in Figure 4.5. Both the datasets suggest that higher number of visually important regions in the image tend to result in higher number of word tokens and types. Figure 4.6 shows the mean word tokens, mean word types, and mean type-token ratio for each observer across all the images. The high values of the mean type-token ratio suggest lexical richness and heterogeneity present in the descriptions provided by the observers.

In both datasets, participants would initially pause and then begin the verbal description but in the DERM II dataset the initial pause was approximately 2.7 seconds longer than in the SNAG dataset. The reason for this may be that the diagnostic task is a

complex cognitive task requiring longer time for the experts to ensure that their spoken description content is correct. Again, task instructions may also have contributed.

4.4 Summary

To sum up, the SNAG dataset is a unique, novel resource that helps us investigate the applicability of the alignment framework on general-domain images where no particular expertise is required to perform the image-inspection task. This dataset can also provide insight into how humans process and describe everyday images involving common objects. Use of automatic speech recognition takes us further in making the framework completely automated. Thus, using both the DERM II dataset discussed in Chapter 3 and the SNAG dataset discussed in this chapter, we explore the applicability of the alignment-annotation framework as discussed in the next chapter.

5

Visual-Linguistic Alignment

This chapter describes an overview of the alignment-annotation framework in Section 5.1. This is followed by Section 5.2, which explains how we extract linguistic and visual units for the DERM II dataset and align them. In Section 5.3, we describe the framework with respect to the SNAG dataset. Sections 5.4 and 5.5 describe how we obtain reference alignments and baseline alignments, which we use for the evaluation study in Chapter 6.

5.1 Overview of framework

The alignment-annotation framework, shown in Figure 5.1, consists of four major steps: 1. collecting multimodal data, 2. collecting and retrieving units of analysis, 3. multimodal bitext alignment, and 4. labeling the image regions. Step 1 involves collection of multimodal data, as described in Chapters 3 and 4. Additionally, the raw audio and gaze data are processed to obtain transcriptions and fixations, respectively, that act as input to step 2. In step 2, we extract the units from the transcripts and fixations. These extracted linguistic and visual units are then fed into the bitext alignment in step 3, where they are aligned. In step 4, image regions are labeled using the output from the alignment.

5.2 DERM II visual-linguistic alignments

5.2.1 Linguistic units

The dermatologists' audio recordings were transcribed verbatim, as shown in Figure 5.2. Visual inspection of the transcripts revealed that the important dermatological concepts

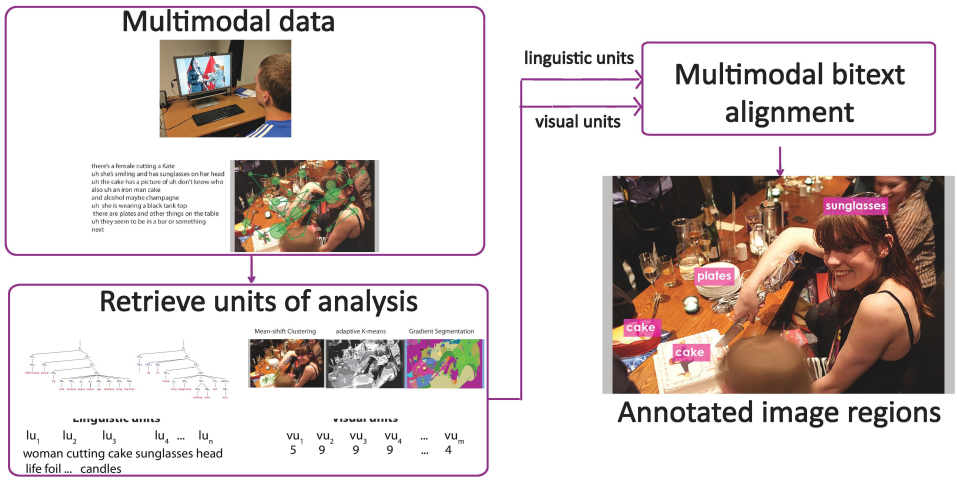

Figure 5.1: Implemented alignment-annotation framework. The SNAG dataset is used as an example to show the framework. The collected multimodal data is processed to retrieve visual and linguistic units of analysis that are then fused using multimodal bitext alignment, resulting in automatically annotated image regions.

used by the experts tended to be nouns and adjectives. After performing minor text normalization, we parsed the transcripts with the Berkeley parser, using the English grammar that is included as part of the code distribution [Petrov and Klein, 2007]. From the parsed output we extracted all types of adjectives (e.g. *red*), singular and plural nouns (e.g. *papules*), singular and plural proper nouns (e.g. *Achilles*), gerunds (e.g. *resolving*), and foreign word (e.g. *erythema*) tokens. These tokens were then filtered to remove stopwords (e.g. *okay*, *the*) along with words used by the observers when following the task-specific instructions to provide a differential, diagnosis, and certainty of that diagnosis (e.g. *diagnosis*, *ninety percent*). Also removed were the names of diagnoses themselves (e.g. *psoriasis*, *basal cell carcinoma*), given that diagnoses correspond holistically to the entire image rather than to a specific image region. Importantly, throughout this pre-processing, the linear order of both the linguistic and visual units was maintained. Figure 5.2 shows the linguistic units extracted for a speaker for the image shown in Figure 5.3.

5.2.2 Visual units

Output from the eye tracker consisted of fixation locations given as x, y coordinates and fixation durations per image per observer as shown in Figure 5.2. We encoded fixations using three different techniques: grid-based image segmentation, mean shift

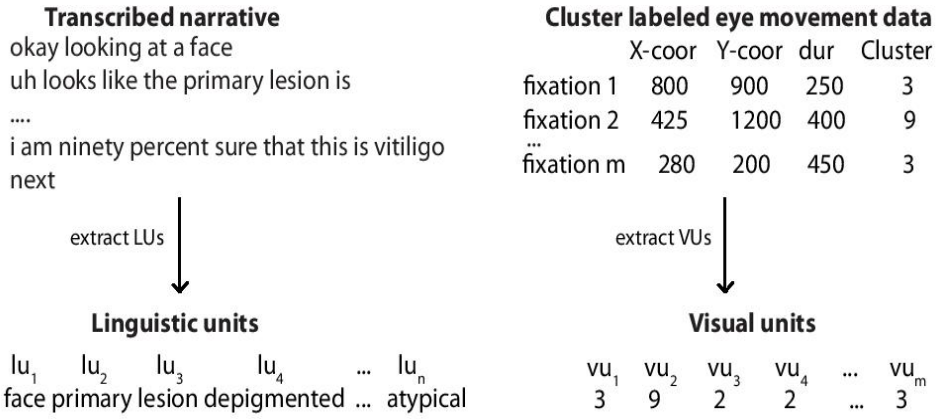

Figure 5.2: The left panel shows an excerpt of a transcribed narrative, the process applied for identifying linguistic units, and the resulting tokens or linguistic units. The right panel shows eye movement data, the process applied in the case of the eye-tracked data, and the resulting gaze-filtered image regions or visual units. The linear order of the units is maintained and reflected in parallel multimodal data sequences. These linguistic and visual units jointly act as input to the Berkeley aligner, taking a bitext alignment approach to associate identified important lexical items and image regions with each other.

fixation clustering (MSFC), and k-means image segmentation.

The most straightforward method to segment an image is to divide it into a grid, as shown in the leftmost panel in Figure 5.3. Each image is 1680×1050 pixels and is divided into a grid of 5 rows and 5 columns. Each cell in the grid is associated with a label that encodes the row and column number for that cell (e.g. 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.

Visual inspection of the scanpaths of observers suggested existence of latent groups of fixations. To explore this further, we used the mean shift fixation clustering algorithm (MSFC) [Santella and DeCarlo, 2004]. The mean shift algorithm is a data-driven method that clusters visual fixations into so-called regions-of-interest. The advantage of using mean shift (MSFC) over other techniques is that MSFC does not require prior knowledge of the number of clusters and additionally is marked by robustness as mean shift is insensitive to outliers. In this work we cluster the fixations spatially but also note that the same method could be used to cluster fixations temporally. For each image, fixations are collected from observers' eye-tracking output. Following this mean shift clustering is applied in

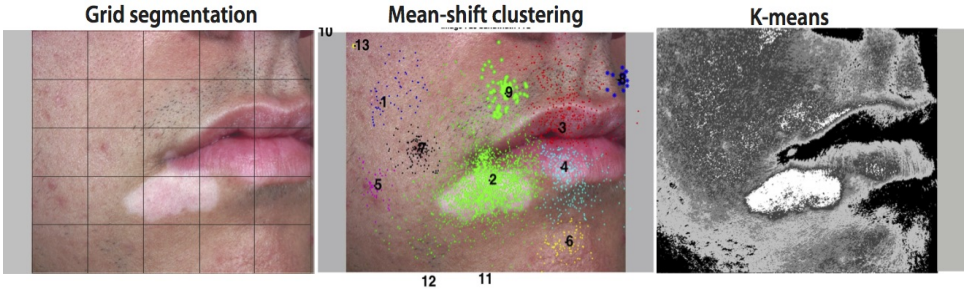

Figure 5.3: The three different fixation labeling techniques used to obtain visual units. Each method clusters/segments the image in a different way following which the fixations are overlaid to extract visual units.

which each fixation is assigned to a cluster of fixations in the same general region of the image. Figure 5.3 shows fixations from all observers for an image clustered here into 13 clusters. Clusters such as 10 in the top left corner of Figure 5.3 contain fixations outside of the image regions mostly due to blinks or track losses by the eye tracker. Such clusters, along with their associated fixations, are removed. For each observer, we then utilized this cluster information to obtain a linearly ordered sequence of visual units (i.e. image regions determined by fixations) that acts as the other input to the alignment algorithm. An example is shown in Figure 5.2. On average for this dataset, the MSFC method yields approximately 10 clusters.

The third method used here is called k-means [Lloyd, 1982]. The k-means clusters image pixels together based on the input features. It is fast, simple, and straightforward to understand. Prior research has shown that *Lab* color features are particularly useful for dermatological images [Bosman et al., 2010]. As briefly mentioned in section 3.6.4, each image is first converted into *Lab* color space, where the L channel represents illumination, the a channel indicates redness-greenness, and the b channel indicates blueness-yellowness in the image. Following this, Lloyd's k-means algorithm is applied, resulting in a segmented image in which each pixel is labeled with the segment label it is a part of. This is shown in the rightmost panel in Figure 5.3. Although the value of k can vary depending on the image and task, we chose k=4, since lower values miss the primary lesion present in the images. Values higher than 4 tend to over-segment many images. The fixation sequences are overlaid on the segmented image and encoded using the segment label they fall within, without loss of linear order.

vah ghar bahut chhota hai	that house is very small
main ghar khareedungi	I will buy a house
vah ghar jal raha tha	that house was burning
yah ek chhota mudda hai	it is a small issue

Figure 5.4: Toy example illustrating the bitext alignment between Hindi and English sentences. The probability of English word *house* being a translation of Hindi word *ghar* increases (black to orange to green) as more parallel sentences containing the two words are added to the training data.

5.2.3 Bitext alignment

Studies have reported that fixations are generated before the end of words and that participants look at an object prior to naming it [Meyer et al., 1998, van der Meulen, 2003]. Additionally, our preliminary analysis showed that there is a temporal lag between when fixations on an object begin and when the person begins naming it. For this reason, visual and linguistic units cannot be aligned merely by considering their time of occurrence. Instead, we require a method that can perform the alignment without making assumptions about the temporal relationship between the units. Conceptually, this is similar to translating one language into another in that the structural characteristics of the source language may not parallel those of the target language. We take advantage of this insight to explore whether a bitext alignment approach can discover meaningful alignments of multimodal data. In statistical machine translation (MT), word alignment models are derived using a parallel corpus of sentences in which each sentence is rendered in two different languages. Figure 5.4 shows a Hindi-English toy example. The principle behind word alignment is as follows: proceed through each pair of training sentences, keeping track of the number of times words co-occur in the two languages. Using these word counts the algorithm builds the probability that a given word in one language (English) is a translation of a word in another language (Hindi). For example, the probability of the English word *house* being a translation of the Hindi word *ghar* increases as more number of sentences containing the two words are processed by the algorithm. In the multimodal scenario of this study, the linguistic (nouns, adjectives, gerunds, and foreign words) and visual (numeric labels of cluster/segments) units extracted for an image

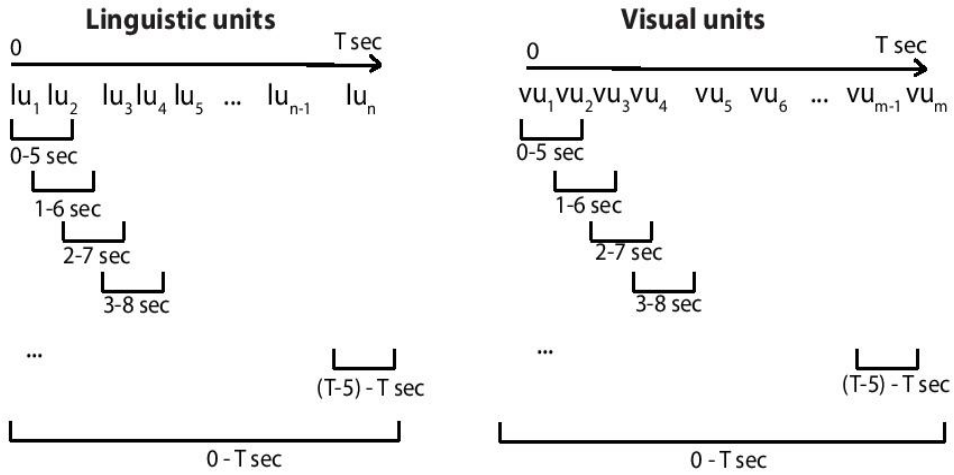

Figure 5.5: Left: Linearly ordered linguistic units obtained from the transcribed narrative. Right: Linearly ordered visual units obtained by labeling fixations using the MSFC algorithm. The labels are different when using other segmentation methods for identifying visual units. Note the linguistic units or visual units are not isochronous. Therefore, the number of linguistic units or visual units between the sliding windows may be different.

representing a pair of "sentences" in the training data. Given the small number of observers per image (29), we get a parallel corpus too small to provide sufficient training data for developing a robust alignment model. It is therefore necessary to increase the size of the training dataset. Utilizing a sliding window of T-seconds where $T = 5$, linguistic and visual units within each sliding window are extracted and added as additional "sentences" or multimodal data pairs to the corpus, as shown in Figure 5.5. Therefore, the number of linguistic or visual units can be different between the sliding windows. By applying the sliding window incrementally, the parallel corpus grows substantially. The original linguistic and visual unit sequence pair, on which the sliding window is applied, is also included in the training data.

Another complication in using this multimodal data is that the sequences of visual units are substantially longer than the accompanying sequences of linguistic units. In order to balance the sequence lengths, we merge contiguous identical visual units (e.g. *cluster3*, *cluster2*, *cluster2*, *cluster3* is converted to *cluster3*, *cluster2*, *cluster3*). This is applied to each sliding window. Subsequently, visual units with the longest fixation duration are selected (keeping the linear order intact) based on the visual-linguistic ratio. The visual-linguistic ratio is defined as $\beta = \frac{Numberofvisualunits}{Numberoflinguisticunits}$, where $\beta = 1$ results in

Linguistic units	Visual units
face	2
face primary	2 6
face primary lesion	2 7 6
...	...
post-inflammatory hypopigmentation atypical	2 4 2
hypopigmentational atypical	4 2
face primary lesion depigmented macule vermilion	2 2 3 6 2 4 7
border involving lower lip corner mouth cutaneous	2 4 2 2 7 6
lip post-inflammatory hypopigmentation atypical	2 4 2 6 2

Figure 5.6: Example training data: A sliding window of 5 seconds is applied to the pair of visual and linguistic "sentences" to expand the data. Subsequently, contiguous visual units are merged and visual units with longest fixation duration are selected. The selected visual units, together with the linguistic units, comprise the training data.

an equal number of visual and linguistic units within each data pair. We also report on the impact of changing the value of T and β as well as the visual unit selection method (α), on the framework's performance[1]. Using the above method the training data for each image increased to approximately 1000 sentences. An example of our training data for the DERM II dataset is shown in Figure 5.6.

We use the Berkeley aligner [Liang et al., 2006] rather than Giza++ [Och et al., 2000] because of its reported higher alignment accuracy and flexibility in testing an existing alignment model on unseen data. One of the biggest strength of the Berkeley aligner is the use of joint training. Further details can be found in Liang et al. [2006]. The Berkeley aligner was run with default parameters settings (2 iterations each of IBM Model 1 and an HMM, joint training, and posterior decoding) with the exception of the posterior threshold used for decoding, which was lowered to 0.1. This value was empirically determined to maximize alignment accuracy on a small held-out set of multimodal data.

[1]We empirically studied the impact of selecting visual units and the values of other parameters in different ways.

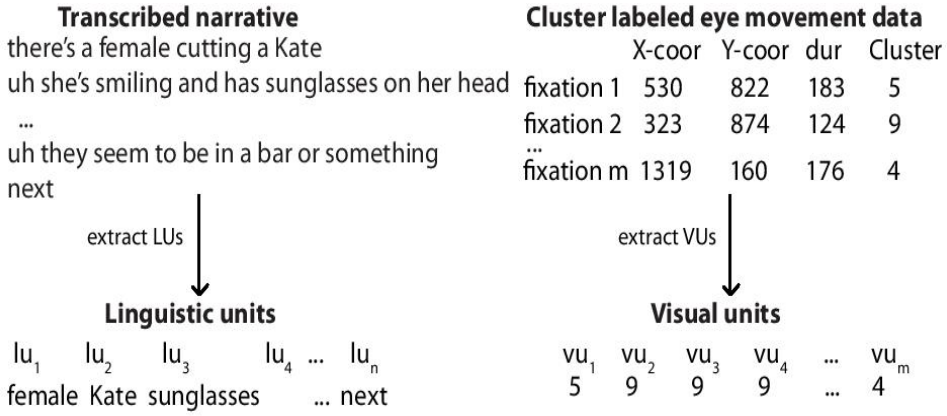

Figure 5.7: Process to extract linguistic and visual units for an image in the SNAG dataset for the MSFC clustering method. The original narrative is automatically transcribed using ASR and linguistic units are extracted. Transcription errors are not corrected manually in order to investigate their effect on the framework. Also, word tokens occurring only once per image such as *wrapped* are removed. This is because word tokens transcribed only once may not necessarily belong to any particular region in the image, or may introduce the idiosyncratic behavior of a participant. Similarly, fixations are labeled based on the cluster they belong to according to the MSFC for the particular image.

5.3 SNAG visual-linguistic alignments

The following section describes in detail the alignment framework for the SNAG dataset.

5.3.1 Linguistic units

As described in Chapter 4, to automate the transcription process, we used IBM Watson Speech-to-Text service for automatic transcription of the audio recordings. Recordings of the descriptions were transmitted as wav files over a WebSocket connection to the Speech-to-Text service which returns transcription results in JSON format. As before, we parsed the original narratives using Berkeley parser and filtered to remove stopwords. Additionally, we removed any word tokens that was transcribed only once for a given image. This is because word tokens with an utterance frequency of one provides less confidence that the word tokens can be associated to a particular region in the image. Frequency of word tokens in the narratives per image is another parameter that needs to be explored in the future. The linear order is maintained. Figure 5.7 shows an example of the linguistic units obtained for this dataset. There are some errors introduced by the ASR system such

Figure 5.8: Original image (top-left), MSFC (top-right), k-means (bottom-left, k=13 for this image) and GSEG (bottom-right) clustering or segmentation output for the image, used for extracting visual units.

as *knife* transcribed as *life*, which we do not correct. One reason behind not correcting these errors is to investigate the resulting effect on the performance of the framework.

5.3.2 Visual units

For the SNAG dataset we used three types of clustering or segmentation methods: mean shift fixation clustering (MSFC), k-means, and gradient segmentation (GSEG) [Ugarriza et al., 2009]. The outputs of the three clustering or segmentation methods are shown in Figure 5.8. The MSFC method is the same as described in section 5.1 for DERM II dataset. For this dataset, on average MSFC yields approximately 11 clusters per image. For the k-means, we do two things differently compared to the DERM II dataset: (1) instead of fixing k to an empirically found value, we applied MSFC to the fixation data for each image and used the number of clusters obtained by MSFC as k, and (2) we use *RGB* and spatial features as input to the k-means algorithm. For this dataset, k means with *RGB* features was visually judged to provide better segments than with *Lab* features. For the sake of clarity, we will refer to this approach for k-means as the *modified k-means*. The GSEG method efficiently integrates spectral intensity, gradient,

and texture information for segmentation purpose. It uses color space gradient information to identify clusters in an image, characterizes the texture in the identified clusters, and applies a region-merging procedure to generate a final segmentation. Sankaranarayanan Piramanayagam, a researcher at Rochester Institute of Technology working on improving the GSEG algorithm, provided us with the non-released version of the toolbox that was applied to the SNAG images with its default values. Further mathematical details about GSEG can be found in Ugarriza et al. [2009].

5.3.3 Bitext alignment

Once the visual and linguistic units were obtained as explained above the same bitext alignment process as used for DERM II dataset was applied. The training data size was increased using the sliding window, and sequence lengths of visual units was balanced. This parallel corpus was treated as input to the Berkeley aligner with the same configuration as that used in DERM II dataset. Effects of various parameters on the output of the framework for this dataset are discussed in Chapter 6.

5.4 Reference alignments

Reference alignments (ground truth) were prepared using a GUI (Figure 5.9) to allow evaluation of the resulting multimodal alignments. This represented the manual alignments obtained by associating each fixation cluster in the case of mean shift fixation clustering and image segment in the case of image segmentation with its corresponding word tokens (linguistic units). Figure 5.9 shows a screenshot of the GUI developed specifically to allow the annotator to perform the manual alignments by drawing borders around image regions and then selecting linguistic units from a pop-up box that contains all the linguistic units for that image. The output from the GUI consists of sets of image pixel coordinates labeled with one or more associated linguistic units, which are then processed to obtain linguistic units corresponding to either fixation clusters in the case of MSFC or image segments in the case of grid-based, k-means, and GSEG. Based on their confidence level, the annotators specify two kinds of alignments SURE (S) and POSSIBLE (P) [Och and Ney, 2003]. SURE alignments define alignments where there is no ambiguity and the annotator's confidence is high. For example, for the image in Figure 5.9, the annotator aligned the word *plates* to the image region circled in black with high confidence. This pair of alignment is therefore added to the set of SURE reference alignment (set S). However, the annotator was not

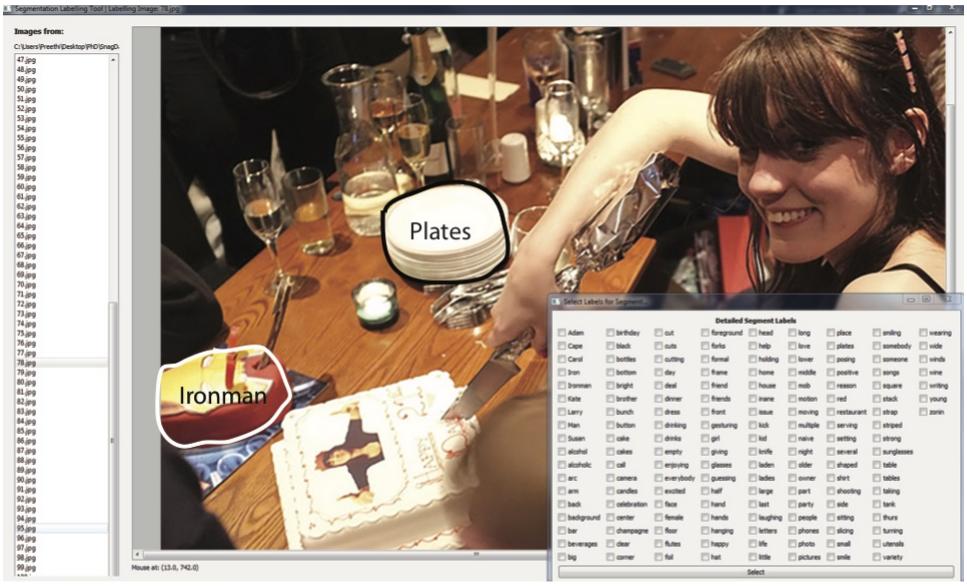

Figure 5.9: Graphical user interface used to acquire reference alignments. The person preparing the manual alignments is able to draw borders around regions and label them with linguistic units. For this image, all pixels within the black border are marked as *plates* in the SURE alignments whereas all pixels within the white border are marked as *Ironman* in the POSSIBLE alignments.

absolutely certain if the word *Ironman* belongs to the region circled in white, thereby adding this alignment pair to the POSSIBLE reference alignment (set P).

For the DERM II dataset, a dermatologist involved with the project from an early stage constructed the SURE manual alignments. The primary researcher of this work indicated the POSSIBLE alignments. While not a dermatologist, the researcher spent prolonged time with these images and dermatological vocabulary. All the manual alignments were done using the post-filtered word tokens. An important observation was that not all the linguistic units present in the narratives were present in the image. Therefore, these linguistic units would also be absent from the reference alignments that is used for evaluation. Only roughly half of the linguistic units present in the narratives are also present in the image and therefore in the reference alignments is indicated in Table 5.1. For example, words such as *rare* and *well-formed* were present in the narratives but not in the image. Therefore, these words were not present in the reference alignments. Given its general-domain nature, the primary researcher of this work performed both the SURE and POSSIBLE manual alignments for the SNAG dataset. For this dataset, the percent of linguistic units present

	DERM II	SNAG
Total no. of linguistic units in narratives	11792	34621
No. of linguistic units in narratives and images	5776	25225
% of linguistic units in narratives and images	48.9	72.8

Table 5.1: Linguistic units present in both the narratives and the images for the general-domain SNAG dataset is higher than for the DERM II dataset. For the knowledge intensive expert domain DERM II dataset, many words such as *atypical* were present in the narratives but not in the image possibly reflecting the use of nonvisual cognitive diagnostic concepts. An example for the SNAG image shown in Figure 5.9 is the word *camera* that was present in the narratives but not in the image.

in the narratives that are also present in the image is close to three-fourths, substantially higher than that for the DERM II dataset. One reason for low overlap in the DERM II dataset could be that the dermatologists were trying to draw inferences based on the visual information given the task instructions. Additionally, the interpretation of linguistic units obtained through spoken description is complex. Therefore, focus is more on manually annotating the *linguistic units* where the interpretation is clear to the annotator. For example, for the image shown in Figure 5.3 few observers uttered the word *patch* instead of *macule*. Intuitively one might say the observers were looking at the correct region and meant *macule*; however, for the purposes of annotation the word *patch* was not considered for manual annotation. This results in lower overlap between reference alignments and linguistic units. Fewer occurrences of such uncertainty was observed in the SNAG dataset due to the general-domain scope.

5.5 Baseline alignments

We compare the performance of the proposed alignment method with two other temporal methods of alignment, namely *simultaneous* and *1-second delay* baselines. Figure 5.10 shows the simultaneous (solid line) and 1-second delay (dashed line) baseline for an example set of visual and linguistic units. Simultaneous baseline alignments are obtained under the assumption that the observers utter the word corresponding to a region at the exact moment their eyes fixate on that region. The 1-second delay baseline assumes that there is a 1-second delay between a fixation and the utterance of the word corresponding to that region, based on prior research [Griffin, 2004]. Although the amount of delay is a parameter that can be varied for comparison against the proposed alignment, we believe

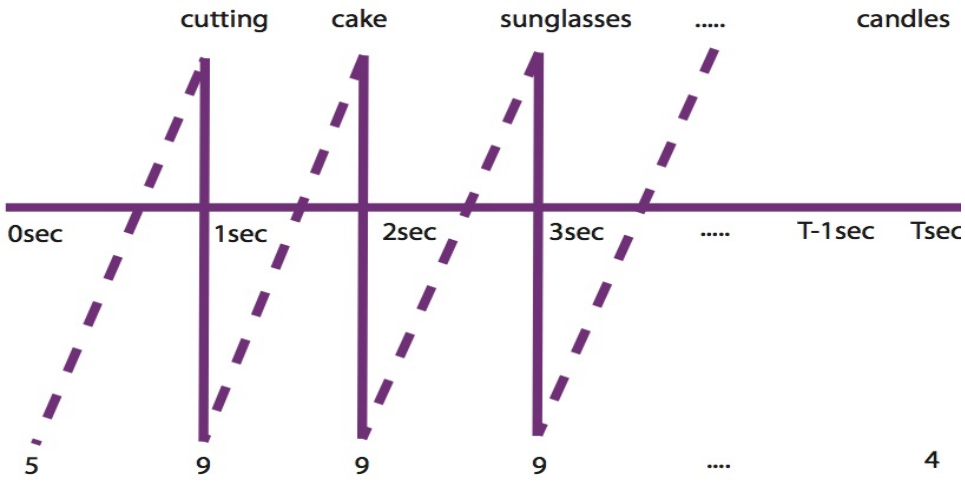

Figure 5.10: Visual units are aligned with linguistic units uttered simultaneously (solid line) and after 1-second delay (dashed line) for the image shown in Figure 5.7.

a fixed-delay will not be capable of aligning words to regions in all cases. This is because prior research has shown that the delay between when a person looks at an object and mentions it depends on various factors such as usage frequency of the object's name and complexity of the name [Griffin and Bock, 2000, Griffin, 2004], and there is variation [Vaidyanathan et al., 2012].

5.6 Summary

This chapter described in detail the alignment-annotation framework proposed in this work. In summary, multimodal data elicited from participants is processed to obtain linguistic and visual units of analysis using various methods that are then aligned using the Berkeley aligner. In the next chapter we discuss the evaluation of the output from the alignment framework against the reference and baseline alignments.

6

Results and Discussion

This chapter presents the evaluation metrics used in section 6.1. Section 6.2 discusses the effects of various parameters and their default values used for the results obtained. This is followed by the analysis and discussion of our results for the DERM II dataset in Section 6.3 and SNAG dataset in Section 6.4, respectively.

6.1 Evaluation of results

Figure 6.1 shows the framework output for a given linguistic and visual "sentence" pair. We use the following metrics and equations from Och and Ney [Och and Ney, 2003] to test how well the framework identifies the correct word-region correspondences compared to the reference alignments:

$$Precision = \frac{|A \cap P|}{|A|} \tag{6.1}$$

$$Recall = \frac{|A \cap S|}{|S|} \tag{6.2}$$

$$Alignment\ Error\ Rate = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{6.3}$$

where A, S and P are number of visual-linguistic unit pairs in the output from the framework, SURE reference alignments that involves no ambiguity in the alignments, and POSSIBLE reference alignments in which alignments may have some ambiguity, respectively. AER is the Alignment Error Rate, which is commonly used to evaluate word alignment in

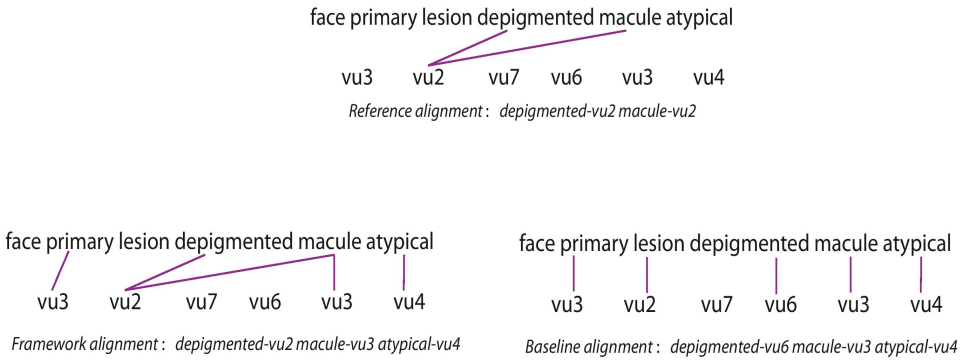

Figure 6.1: Example illustrating output from our framework, the reference alignment, and baseline alignment for a given pair of linguistic and visual "sentences". Linguistic units, such as *atypical*, that do not appear in the image are not present in the reference alignments.

machine translation. A high precision and recall resulting in low AER is considered good. The image regions and their labels change with the segmentation technique being used. Therefore, each segmentation method has its own set of simultaneous and 1-second delay baselines, reference alignment, and alignments from the proposed framework that are used to compute the metrics. In general, the 1-second delay baseline tends to perform as well as or better than the simultaneous match baseline.

A qualitative visualizer was built to visualize the resulting annotations corresponding to the image regions. The visualizer sorts the words in increasing order of frequency of utterance and displays W words on the corresponding image region locations. The number of visualized words W, if needed, can be different for different images. Various results shown and discussed in this chapter use the visualizer with the value of W ranging from 2 to 4 in order to illustrate the output annotations. Low values of W were picked to avoid clutter for illustration of results.

6.2 Effect of parameters

The alignment framework is built based on various parameters ranging from the minimum fixation duration used for the process of identifying fixations and saccades in BeGaze to the value of posterior threshold used in the Berkeley aligner. We experimented with the parameters sliding window (T), visual-linguistic ratio (β) that ensures equal length of

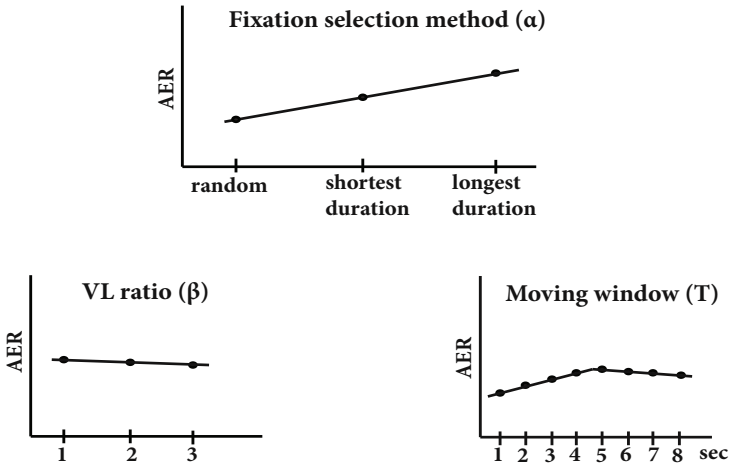

Figure 6.2: General effects on performance per parameter. The effect (positive or negative) reflected all measures. Default values used in this work resulting in high performance are: α = longest duration, $\beta = 1$, and $T = 5$ seconds.

	grid-based			MSFC			k-means		
	Precision	Recall	AER	Precision	Recall	AER	Precision	Recall	AER
Simultaneous	0.32	0.27	0.72	0.36	0.44	0.61	0.39	0.44	0.59
1-second delay	0.34	0.28	0.70	0.38	0.44	0.61	0.40	0.44	0.59
Alignment framework	0.38	0.29	0.68	**0.45**	**0.56**	**0.51**	0.41	**0.56**	0.54
% improvement (over 1-second delay)	4	1	2	7	12	10	1	12	5

Table 6.1: Comparison of alignment performance, average across images. The proposed framework, specially using, MSFC tends to performs the best for the DERM II dataset. The last row shows the absolute improvement (in percentage) achieved by the different clustering or segmentation methods over the 1-second delay baseline.

sequences of visual and linguistic units, and method of visual unit selection referred to as fixation selection method (α). As illustrated in Figure 6.2, the general trend was similar for both dataset. When the longest fixations within a sliding window were selected as visual units the framework's performance was higher. This supports the intuitive notion that participants would fixate longer on image regions that play an important role in achieving the end goal. The default sliding window value of 5 seconds performs the best for the two datasets and higher values do not result in any improvement. Both the visual-linguistic ratio in our framework and the posterior decoding threshold in the Berkeley aligner have a negative effect on the framework's performance as they are increased. Results for effect of parameters for the DERM II dataset are published in Vaidyanathan et al. (2015b).

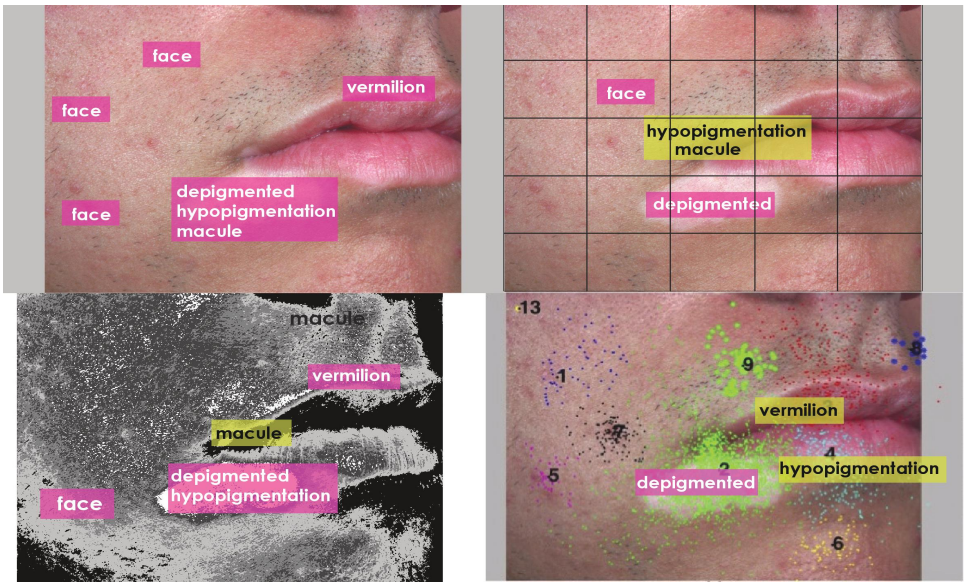

Figure 6.3: Annotations from top-left: The annotator, top-right: grid-based, bottom-left: k-means, and bottom-right: MSFC alignments, respectively. Compared to the annotator's reference alignments, *hypopigmentation* and *vermilion* although incorrectly placed, are still close to their corresponding regions with MSFC. In contrast *macule* with k-means is quite far from where it should be.

6.3 DERM II

Our alignment method, irrespective of the fixation encoding technique, yields stronger performance in comparison to the two baselines. In general, within our alignment method both MSFC and k-means outperform the grid-based method. As indicated in Table 6.1, MSFC achieves absolute improvement of 7%, 12%, and 10% for precision, recall and AER, respectively, over the 1-second delay baseline. On the other hand, k-means when compared to the 1-second delay baseline achieves 1%, 12%, and 5% absolute improvement for precision, recall, and AER, respectively. The results hold on a per-image basis as well, with the MSFC-based alignment approach yielding higher recall and lower AER than baselines in 29 and 28 of the total 29 images, respectively, and higher precision than baselines in 24 of the 29 images. The k-means linked alignment on the other hand yields comparable precision for only 17 of the 29 images.

Figure 6.3 shows an image overlaid with the most frequently used linguistic units with which those regions were aligned by our framework using the three methods. We compare it

to the image when labeled by the annotator (expert dermatologist) in Figure 6.3 (top-left). The labels are generally accurate and well located on the image in the k-means and MSFC cases when compared to the manually annotated image. The grid-based method performs poorly with only one correct label-region association. Labels such as *hypopigmentation*, although incorrect, are still close to the corresponding region in the case of MSFC for this image.

Considering our metrics the MSFC method outperforms the grid-based method in all cases and k-means in many cases. While both MSFC and k-means use fixations as pointers when identifying sequences of visual units from the determined image regions, the major difference is that MSFC identifies the regions by clustering users' fixations based on spatial coordinates without using image features whereas k-means identifies the image regions based on image features without using fixations. Even though k-means in comparison to MSFC uses image features in addition to fixations, it does not use the fixations during the segmentation process. The performance when considering MSFC suggests that we need to include perceptual information to capture the semantics of images from users' perspectives for the process of identifying image regions.

We organized our images into four different groups following Li et al. (2013): *single lesion*, *multiple lesions*, *bilateral lesions*, and *distributed lesions* (Figure 6.4). The category *multiple lesions* had twice as many images as the other categories. Few images in our database posed ambiguity regarding which category they belonged to. The k-means results in 4 segments for each image while MSFC yielded on average 9, 9, 6, and 8 clusters for the above categories, respectively. Interestingly, our results indicate that for most of the single, bilateral, and distributed lesions cases MSFC with bitext alignment performs better than k-means. However, annotations for images with multiple lesions are mostly improved when using k-means with bitext alignment. Figure 6.4 shows some annotations achieved for different cases using MSFC (black) and k-means (red).

Both MSFC and k-means in Figure 6.4 do well for the single lesion image (top-left) but in the multiple lesion image, MSFC does worse than k-means (bottom-right). In this case, MSFC is able to annotate one particular region as *pustule* but k-means, on the other hand, is able to identify all regions associated with *pustule*. When the lesion size gets larger, the k-means method either fails to annotate all of the regions, as in the bilateral case (bottom-left) or annotates the incorrect ones, as in the distributed lesion case (top-right). A visual inspection of the k-means segmented images indicates that, in some cases, color features alone are insufficient and that texture might be playing an important role as

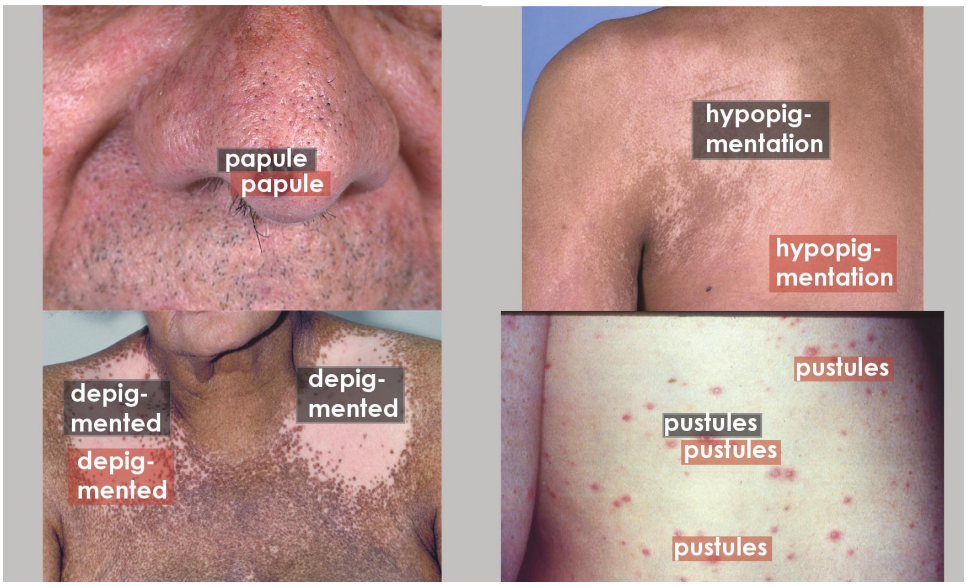

Figure 6.4: Annotation examples for four different kinds of images with MSFC (black) and k-means (red) used to obtain visual units. Top-left: Single lesion case where both MSFC and k-means are comparable with the reference alignment. Top-right: Distributed lesion case where MSFC correctly identifies *hypopigmentation*. Bottom-left: Bilateral lesions where MSFC correctly annotates both parts of the lesion whereas k-means fails to annotate one part. Bottom-right: Multiple lesions case where k-means correctly labels all the regions pertaining to *pustules* and MSFC labels only one. This shows the advantage and disadvantage of using each method.

well, as shown in Figure 6.4 top-right where color is not as strong a feature as texture for differentiating the lesion.

The lower performance of MSFC on images with multiple lesions could be due to the fact that fixations are single image coordinates and sometimes can lie just outside of the tiny multiple lesions, even though the observer might have been paying attention to the lesion itself. This could be solved by considering a region around the fixation approximating the fovea. Also, some of the images with distributed lesions have values very close to the skin in the *Lab* space. This may lead to a poor k-means segmentation output thereby resulting in lower performance of the framework when compared to the MSFC method.

Further investigation reveals that even though MSFC's performance for multiple lesion cases is lower than that of k-means, the best improvement (15% in AER) that our framework achieves over the 1-second delay baseline is for these images with multiple lesions. The best improvement (10% in AER) over the 1-second delay baseline for the

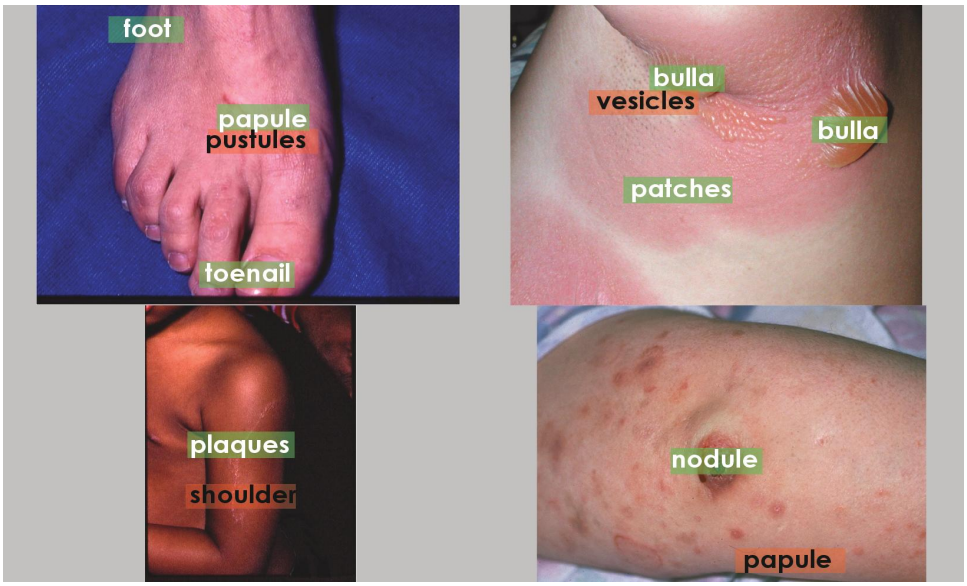

Figure 6.5: Correct (green) and incorrect (orange) labels identified by MSFC. Many labels are correctly aligned even in challenging cases. Top-left: Whereas tiny *pustules* (on the second toe) are incorrectly identified *papule* and *toenail* are satisfactory. Top-right: The system correctly identifies the labels *patches* and *bulla* but does not recognize the tiny collection of shiny *vesicles* underneath the arm. Bottom-left: Body part *shoulder* although incorrectly annotated is still identified as a *part of the* arm. Bottom-right: While the single lesion *nodule* is correctly identified the many scattered *papule* are incorrectly annotated.

k-means method is for images with single lesions.

Figure 6.5 shows more annotation examples from our database using the MSFC method. It correctly identifies single lesions *bulla, nodule, patches* in the top-right and bottom-right panels but misses multiple small *papules* in the bottom-right. The top-left example is an interesting case where MSFC identifies the label *papule*, probably because the *papules* form one long lesion, as opposed to bottom-right where they are scattered. This is confirmed with the observation that MSFC fails to identify *pustules* in top-left since they are scattered (on the second toe).

We also analyzed the performance of the framework for certain concept labels across images. In dermatology the main lesion present in an image is called the *primary morphology*. Across the 29 images, 9 primary morphologies were represented: *plaque, papule, nodule, patch, pustule, bulla, macule, vesicle* and *hypopigmentation* with *plaques* being most frequent and *hypopigmentation* being the least frequent. Each of these primary

morphologies was either present independently or in groups with others. For each label, we calculated precision for each of these labels across all of the images for the MSFC and k-means method. Precision for MSFC in general tended to be higher than that for k-means, particularly for *papule* and *nodule* (0.9). Precision for the rest of the labels ranged from 0.65 to 0.82 for MSFC. An in-depth analysis shows that even though k-means correctly aligned many of these 9 labels, it also aligned them with many non-relevant regions thereby lowering the precision. Using the k-means method, *pustule* and *vesicle* scored slightly better (0.75) than using MSFC.

Therefore, it can be summarized that key labels such as *papule* and *nodule* had high precision for MSFC whereas *pustule* and *vesicle* had high precision for k-means. This indirectly confirms the observation that MSFC performed well on single lesion images since most of the single lesion images had large *papule* or *nodule* as opposed to groups of tiny *pustules* or *vesicles*, in which case k-means tends to perform better. The in-depth analysis of *primary morphology* labels across images is useful in identifying their characteristics across images. We could potentially extract all the regions from all the images pertaining to each of these labels, extract image features and learn both common and idiosyncratic features for each label.

For both MSFC and k-means there is substantial improvement in the recall values over baselines when using bitext alignment across all evaluated image cases. We note that precision is often lower due to linguistic units (*little*) that are not physically present in the image. Additionally there are some linguistic units that correspond to the entire image rather than a specific region (e.g. *face* as shown in Figure 6.3). Such holistic or abstract units are not included in the manually derived reference alignments, resulting in lower alignment precision for these images. In the future, we will incorporate methods to filter such abstract words. These results are published in Vaidyanathan et al., (2015a, 2016).

6.4 SNAG

Using equations stated in Section 6.1, we calculated the average precision, recall, and AER for alignments in the SNAG dataset and compared them against the baselines. The comparison was done for the three clustering or segmentation methods mean shift fixation clustering (MSFC), modified k-means with RGB color features and k equal to the number of fixation clusters obtained by MSFC for each image, and gradient segmentation (GSEG).

The simultaneous baseline's performance measures are similar to the 1-second delay

	MSFC			modified k-means			GSEG		
	Precision	Recall	AER	Precision	Recall	AER	Precision	Recall	AER
Simultaneous	0.42	0.30	0.65	0.49	0.17	0.74	0.41	0.14	0.78
1-second delay	0.43	0.31	0.64	0.50	0.17	0.74	0.42	0.15	0.78
Alignment framework	0.43	**0.50**	**0.54**	**0.56**	0.31	0.60	0.48	0.28	0.65
% improvement (over 1-second delay)	0	19	10	6	14	14	6	13	13

Table 6.2: Average alignment performance across images in the SNAG dataset, for three different clustering or segmentation methods. Our framework with the MSFC clustering method provides the best recall and lowest AER. However, modified k-means provides the best precision. The absolute improvement achieved by the different clustering or segmentation methods over the 1-second delay baseline are shown in the last row.

	MSFC	Modified k-means	GSEG
Precision	62	96	96
Recall	100	100	100
AER	99	100	100

Table 6.3: Number of images for which our alignment framework provides an improvement over the baselines, for each case of clustering or segmentation method. All three methods provide improvement over the baselines for both recall and AER on all images with modified k-means and GSEG providing improvement in precision as well. The total number of images used in the dataset was 100.

baseline for SNAG dataset as well. As shown in Table 6.2 the proposed framework for alignment performs better than either of the baselines. Among the three clustering or segmentation methods, MSFC yields the highest recall and lowest AER. It achieves an absolute improvement of 0%, 19%, and 10% for precision, recall and AER, respectively, over the 1-second delay baseline. The absolute improvement percentages are shown in the last row of Table 6.2. Modified k-means, on the other hand, results in higher precision with an absolute improvement of 6%, 14%, and 14% over the 1-second delay baseline for precision, recall, and AER, respectively. In comparison to MSFC and modified k-means, the performance of GSEG is comparable with an absolute improvement of 6%, 13%, and 13% for precision, recall, and AER, respectively. Table 6.3 shows performance for each clustering or segmentation method based on the number of images. While all three methods yield higher recall and lower AER than baseline for almost all 100 images, modified k-means and GSEG yield higher improvement in precision for 96 images outperforming MSFC.

Visual comparison of reference alignments provided by the annotator with the alignments obtained through our framework for the three clustering or segmentation

Figure 6.6: Top-left: Reference alignments as provided by the annotator. Alignment output when using: Top-right: MSFC, Bottom-left: modified k-means, and Bottom-right: GSEG methods, respectively. Correct alignments are shown in pink whereas misalignments as well as labels not belonging to reference alignments are shown in yellow. Both modified k-means and GSEG align all instances of labels such as *plates* with the incorrect regions whereas MSFC aligns one instance of the label correctly. The visualization tool places the label within the corresponding segment, however, in cases where the segments are small the labels may appear to belong to the adjacent segments too (e.g. *plates* in bottom-right.

methods shows (Figure 6.6) most of the words are correctly aligned (pink) by all three methods. MSFC correctly aligns labels present in the SURE reference alignments such as *cake* and *plates*, yielding a higher recall. It also aligns some of these labels such as *plates* to regions they do not belong to explaining the low precision values. On the contrary, both modified k-means and GSEG misalign labels such as *plates* leading to a lower precision. Also, all three methods align labels such as *camera*, which cannot be grounded to any region in the shown image. Such abstract labels that are not present in either SURE or POSSIBLE reference alignments lower the precision values.

The improvement over the baselines supports that it would be naive to assume simultaneous or fixed-delay correspondence between utterances and eye movement, and underscores the promise of our alignment-annotation approach. This is true regardless of the method used for the identification of visual units or the type of image. As highlighted in Section 6.3, these results indicate that the alignment-annotation framework could in the

Figure 6.7: Example images from category top-left: O=1, with one primary object (*lady*). Top-right: O=2, two primary objects (*girl, bear*). Bottom-left: O=3, three primary objects (*gentleman, army officer, scissors*). Bottom-right: O≥4, four or more primary objects (*person 1, person 2, person 3, person 4, cooler etc*), respectively. Labels in pink indicate all the three methods correctly aligned them. Incorrect alignments are shown in black (MSFC), red (modified k-means), and blue (GSEG). The number of misalignments increases as the images get more cluttered.

future consist of a clustering or segmentation method that uses both fixations and image features during the segmentation process. This is expected to help in reducing the chances of image regions, for example in k-means, representing different concept labels (linguistic units) corresponding to the same region label (visual unit).

We divided the images in the SNAG dataset into four categories ranging from simple to complex, as shown in Figure 6.7. Category *O=1* consisted of images with one primary object to gaze at and describe. For instance, image on top-left of Figure 6.7 consists of one prominent object *lady*. Although there are other objects in the image to look at and describe since there is only one prominent object the annotator categorized this image in *O=1* category. Likewise, category *O=2* and *O=3* consisted of approximately two and three primary objects to gaze at and describe. Category *O≥4* represents images with more than 3 primary objects. There were 16, 37, 12, and 35 images in each category, respectively. The MSFC yielded on average 11, 10, 11, and 11 clusters for the four categories, respectively.

	Precision			Recall			AER		
	MSFC	modified k-means	GSEG	MSFC	modified k-means	GSEG	MSFC	modified k-means	GSEG
O=1 (16)	0.43	0.57	0.47	**0.55**	0.31	0.3	0.53	0.59	0.63
O=2 (37)	0.47	**0.59**	0.51	**0.55**	0.32	0.29	**0.51**	0.58	0.63
O=3 (12)	0.44	0.55	0.48	0.44	0.28	0.25	0.56	0.62	0.67
O≥4 (35)	0.38	0.51	0.44	0.47	0.29	0.27	0.59	0.63	0.66

Table 6.4: Comparison of alignment performance for four different categories of images for different clustering or segmentation methods. These four categories are defined based on the approximate number of primary objects in the image, for example O=1 indicates the images in this category had one primary object to gaze at and describe. Not surprisingly, as the number of primary objects increase, the alignment performance decreases. Also, regardless of the category of image, modified k-means provides the best precision whereas MSFC provides best recall and AER.

Modified k-means resulted in the same number of segments for each category since it uses the number of clusters provided by MSFC. As indicated in Table 6.4, the categorization does not have much of an effect on the general trend of performance of the clustering or segmentation methods. MSFC claims higher recall and low AER values while modified k-means claims high precision values. However, the best performance is obtained for images in category *O=2* followed by category *O=1*. This suggests that the number of objects in an image may affect the aligner's performance. An important point to note is that the above categorization is coarse and may involve subjectivity as it was performed by one annotator, the primary researcher in this case. Further work is required to explore a more defined method of dividing the images based on number of objects and using more than one annotator to reduce any subjectivity.

Figure 6.7 shows the obtained alignments overlaid on their respective images for the four categories. In general, labels are aligned correctly, but we also get additional misalignments, regardless of the clustering or segmentation method used. These misalignments seem to increase in number as the complexity of an image (i.e. number of objects) increases thereby lowering performance. MSFC seems to have less number of spurious alignments compared to both modified k-means and GSEG, possibly because its solely based on the fixation data.

As previously mentioned, MSFC has the advantage of being solely based on fixations thereby reducing the errors introduced due to sharing of image features by various objects. Sharing of image features can lead to common image segment-labels during the segmentation process. For example, in GSEG (Figure 6.8, bottom-left), the man's *coat*

Figure 6.8: Output from top-left: MSFC, top-right: Modified k-means, where k is equal to the number of clusters obtained from MSFC, bottom-left: GSEG, and bottom-right: k-means, where k=4, respectively. Modified k-means and GSEG tend to oversegment leading to multiple segment labels for a given word-label whereas k=4 may lead to undersegmentation in other cases leading to one segment-label shared by various word-labels. A semantic segmentation method built using gaze data and image features may be the solution to this problem.

and part of the *scissors* have the same segment-label. This would lead the framework to incorrectly learn that labels *coat* and *scissors* both belong to the same image region. When comparing k-means with k=4 used in the DERM II dataset and modified k-means, we can observe that k=4 leads to segments that certainly look much cleaner. This is also supported by the quantitative results obtained with $k = 4$, with average precision = 0.56, average recall = 0.46, and average AER = 0.49. Low value of k here is advantageous in some cases, such as *scissors*, which is aligned to one segment label in k-means with k=4 as opposed to approximately three segment labels in modified k-means. For our purposes, the image region corresponding to the word *scissors* need not be segmented into further segments, since our participants do not mention parts or regions of the *scissors*. Similarly, GSEG also tends to oversegment in many cases thereby leading to low values of AER. We still face the problem where both *coat* and *scissors*, despite being different objects, belong to the same segment. This leads to low AER measures. MSFC also faces the same issue in

	MSFC		k-means	
	Precision	Recall	Precision	Recall
DERM II	0.45	0.56	0.41	0.56
SNAG	0.43	0.50	0.56	0.46

Table 6.5: Comparison of precision and recall from the alignment framework for the two datasets for MSFC and k-means with k=4. Precision is generally lower than recall except for the case of k-means with the SNAG dataset.

cases where the algorithm clusters fixations falling on two unrelated regions of the image into one cluster. These observations strongly suggest that our framework would benefit from a segmentation technique that builds on both image features and gaze data.

Interestingly, recall values are higher for the DERM II dataset when compared to the SNAG dataset. Recall values indicate the number of alignment pairs in the reference alignments that are also obtained in the framework's output alignments. One possible reason for high recall values could be that as a result of task instructions DERM II dataset has a precise and limited vocabulary. Due to the nature of the dermatology field, most of the regions in the images usually correspond to exactly one label. For instance, the primary lesion in the image in Figure 6.3 corresponds to the label *macule* and most of the observers mentioned it. On the other hand, due to the general-domain nature of the images in the SNAG dataset, many objects in the images correspond to various labels. For example, for the woman in the image in Figure 6.6, observers mentioned the labels *lady*, *girl*, and *female*. Thus labels that were not mentioned by majority of the observers will have low probability of being associated with the corresponding image region leading to low recall values. Also from Table 6.1 and Table 6.2 we learn that the best absolute improvement over baseline is obtained for Recall for both datasets when compared to Precision and AER.

As shown in Table 6.5 precision values are generally lower than recall values when comparing the framework's performance for MSFC and k-means with k=4 regardless of the dataset. However, this is not true for the case of k-means with the SNAG dataset where precision is higher than recall. Similarly for modified k-means and GSEG the precision is higher than recall for the SNAG dataset. Further work is needed to investigate the reason for this trend.

For the two datasets, we also investigated the effect of the number of clusters obtained from mean shift fixation clustering on the framework's performance. Table 6.6 shows the

	Precision	Recall	AER
DERM II	0.25 (0.29)	0.25 (0.19)	-0.15 (0.43)
SNAG	-0.29 (0.003)	-0.29 (0.003)	0.43 (5×10^{-6})

Table 6.6: Pearson's correlation value (r) and the corresponding significance value (p) between the performance metrics and the number of clusters obtained using MSFC for the two datasets.

	MSFC		modified k-means		GSEG	
	uncorrected	corrected	uncorrected	corrected	uncorrected	corrected
Precision	0.5	0.69	0.6	0.83	0.51	0.71
Recall	0.53	0.55	0.33	0.36	0.28	0.3
AER	0.48	0.37	0.55	0.47	0.62	0.55

Table 6.7: Comparison of average alignment performance across 5 images in the SNAG dataset for *uncorrected* vs. *manually corrected* narratives. There is substantial improvement in both precision and AER for all the clustering or segmentation methods. The MSFC still offers the best AER.

Pearson's correlation coefficient between the number of clusters in the images and the precision, recall, and AER values for the two datasets. Also shown are the corresponding significance values. For the DERM II dataset the correlation between number of clusters and performance metrics is not significant ($p > 0.05$) enough. However, for the SNAG dataset all three metrics are highly correlated with the number of clusters obtained using MSFC. The negative coefficient shows that as the number of clusters increases the performance decreases. This may be due to the fact that less number of clusters mean fewer incorrect output alignments. Further work is needed to investigate the cause of this correlation.

6.4.1 Manual correction vs. ASR only

We also applied our annotation-alignment framework to the manually corrected narratives for 5 images (described in Chapter 4). Table 6.7 shows the performance of the framework with the corrected and uncorrected narratives. For the SNAG data, narratives were on average 60 words in length and on average needed correction of 3 words. There is significant improvement in both precision and AER for all three clustering or segmentation methods between the uncorrected and corrected narratives. Using automated transcription reduces manual labor by a substantial amount, but the performance improvement suggests the limitations of the automated transcription. Therefore, performance could be improved

by using automated transcription followed by manual correction thereby reducing some amount of manual labor. This would also be helpful in better training of the automated transcription tool. Again, the precision for corrected narratives is higher than both uncorrected narratives in the SNAG dataset as well as the DERM II dataset due to overlap percentage of linguistic units with reference alignments. This indicates that we need improved methods to filter out or otherwise handle words that cannot be grounded in regions of the image.

6.5 Summary

The quantitative and qualitative analysis discussed in this chapter show the usability of our alignment framework in achieving meaningful image region annotations from multimodal data. The improvements in alignment accuracy over the baselines for both datasets indicates that the success of the framework is not limited by the type of image used. This is particularly important since it enables the framework to obtain annotations for specific-domain images in a less expensive and laborious manner. Annotations for specific-domain images need experts making it more expensive. Our framework can allow collection of annotation without having an expert go through the laborious task of marking image regions and annotating them by hand. For example, real-time gaze and speech data could be collected while a dermatologist is diagnosing a patient, which can then be processed for annotations.

7

Future Work and Conclusions

This work described in detail a visual-linguistic alignment framework that can be used for annotating image regions. We also reported on a collected visual-linguistic or multimodal data resource with general-domain images, SNAG, valuable for the scientific community. For the DERM II dataset, the individual performances of MSFC and k-means show that they each have strengths that together could be of more value than when individually used. One way to combine the two techniques is to identify the cluster type (multiple, single, bilateral, distributed) of a dermatology image and select a method (MSFC, k-means, etc.) that performs well on the identified cluster type. Likewise, it may also be useful to weigh the output of certain methods for particular concept labels. Another way to combine the strengths of individual methods is to develop a new segmentation algorithm that would combine image features and gaze data during the segmentation process. Gaze data where fixations are convolved with a Gaussian kernel to mimic natural perception may also improve the performance. Additionally, image features such as spatial coordinates and texture should also be included. Future work can involve a method where using k-means with image features such as spatial coordinates, color, and texture, images will be oversegmented into numerous superpixels. These superpixels can then be re-grouped using gaze data resulting in image regions obtained using perceptually important image features.

For the SNAG dataset, MSFC and k-means with k=4 show better performance than modified k-means and GSEG due to oversegmentation. Therefore oversegmentation is an important issue to keep in mind when the new segmentation approach is designed for the framework. Reducing the oversegmentation problem in both modified k-means and GSEG

would result in better performance of the framework. Apart from oversegmentation, images with more number of objects begin to pose a challenge to the segmentation methods. An advantage with the images in the SNAG dataset is that they are general-domain images from the MSCOCO dataset. Several state-of-the-art segmentation methods including deep learning methods have been shown to successfully perform on these images. We can further investigate the performance with DeepMask, a deep learning method and Convolutional Oriented Boundaries, a contour detection and hierarchical segmentation approach [Maninis et al., 2017]. It would also be interesting for the deep learning methods to use gaze data along with other features as input to the neural network.

Currently we focus on extracting mostly nouns and adjectives as linguistic units, which consist of both units that can be grounded in an image and abstract units. We can use abstract concept filtering [Kiela et al., 2014] to remove such abstract units that add to the low values of precision, in the framework. Another method to remove linguistic units that are not present in the image from the narratives is by weighing linguistic units using their frequency or by the percent of participants that mention them. For example in the case of dermatology, for a given image if a participant mentions the incorrect label such as *patch* as opposed to *macule*, using frequency as a weighting factor the label *patch* could be removed or treated differently than the label *macule* for that particular image. In addition, our method for extracting the linguistic units relies on parsing output using the Berkeley parser, which for the DERM II dataset could be improved by training the parser on spoken language data from the biomedical domain. The existing system could be improved further by incorporating knowledge about conceptual relations such as *meronymy*, commonly known as *part-whole* relationships, in both the linguistic and visual modalities. For example, in Figure 6.5 (top-left) the word *foot* corresponds to almost the entire image and therefore most of the regions, whereas the word *papule* corresponds to a particular image region. In this particular image case, a *foot has-a-papule* relation can be seen. Considering this information could benefit the alignment and semantic alignment procedure, for instance in terms of helping to better identify the visual units (including the annotation of foregrounded elements in more narrowly identified image segments with respect to annotation of their embedding larger image region). This would further improve the accuracy of multimodal data alignment and semantic annotation of images.

The multimodal framework can also be used to understand valence-related image content. Exploration of applying the framework to understand how humans react to images of varying emotional content has already begun with an extended set of collaborators

(i.e. Gangji et al. 2017). The obtained image region annotations are useful for a variety of image-based applications. They can assist in image classification and retrieval where it is important to not only have annotations of the image as a single artifact but also have annotations for specific image regions. Captions and descriptions for an image can be automatically generated using the output from the framework. We can also build interactive application systems where a user could be guided through an image-based task using real-time gaze and spoken data. This could be particularly useful, for example, for training dermatology or radiology interns. As mentioned in the introduction, this framework can assist in developing computer applications where when a user gazes at a *painting* in a museum, the computer can highlight areas of the painting where an artist looked at and provide more conceptual information about that area. Last but not the least, all the individual steps in the framework are automated, which facilitates the translation to industry level automation.

From our results it is evident that the proposed alignment framework performs better than the simultaneous and delayed baselines for both the DERM II and SNAG datasets. This shows that integration of multimodal data, specifically visual and linguistic data, is possible using bitext alignment. The above conclusion is supported by both qualitative and quantitative results. The resulting annotations confirm that bitext alignment as employed by our alignment framework can be used to obtain image region annotation. Additionally, the framework's performance also confirms that naturally elicited spoken narratives through the Master-Apprentice model (as opposed to written captions) are valuable for image region annotation.

The qualitative and quantitative results obtained for both the DERM II and SNAG datasets indicate the broad applicability of the multimodal bitext alignment image region annotation framework. This framework does not depend on a specific type of expertise or image type and it can be applied to expert-domain as well as general-domain images. It should also be applicable to other expert domains than the one explored here (dermatology).

Overall, for both datasets, the MSFC clustering method outperforms the other segmentation methods. This indirectly validates the crucial role gaze data can play in an image region annotation framework. Other image segmentation methods such as k-means and GSEG provide comparable values of precision, suggesting that image features are also necessary for modeling image region annotation. Thus, to build an image region annotation framework that can assist in developing advanced image-based application systems we need

to integrate multimodal data elicited from humans with inherent information present in the image. The ability of different segmentation methods to handle different categories of images suggests that an extended framework could benefit by including an ensemble of distinct techniques to address the heterogeneity of images and conceptual regions across images.

The framework's performance on uncorrected narratives suggests that there is potential in using automated speech-to-text transcription tools. However, the improved performance of the alignment framework on manually corrected narratives when compared to uncorrected narratives indicates that automated transcription followed by manual correction may be preferential, at least at times until ASR methods have improved further.

For the two datasets, parameters such as the size of the time window used to expand the parallel corpus did not have major effect on the framework's performance. This probably indicates that the alignment annotation framework relies predominantly on the input data itself and is quite robust to parameters. However, size of the dataset, i.e. number of parallel sentences, affects the values of AER. Using the concept of sliding window aids in lowering the values of AER but significant reduction can be achieved by adding more observers.

The proposed alignment framework shows how we can adapt natural language processing and computer vision methods to creatively integrate visual and linguistic information. This work shows how such a multimodal integration could be used to achieve unsupervised semantic annotations for images. Like most datasets involving multimodal data elicitation from humans, our datasets are modest. Nevertheless, our results clarify our method's promise, and the quantitative metrics we apply and visualized results obtained support our conclusions. With advanced technologies such as virtual reality glasses, wearable eye-trackers, and smartglasses, collecting multimodal data could eventually become straightforward and natural resulting in more data that the alignment-annotation framework and image-based application system could benefit from. Our work is an important contribution toward the highly challenging problem of fusing human-elicited multimodal data sources, a problem that will become increasingly important as such data become more common.

8

List of Publications

(1) Haduong, N., Nester, D., **Vaidyanathan, P.**, Prudhommeaux, E., Bailey, R., and Alm, C. O. (2018). Multimodal alignment for affective content. In *Proceedings of the Workshop of Affective Content Analysis at AAAI*, Forthcoming.

(2) 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*, pages 621-627.

(3) Farnand, S., **Vaidyanathan, P.**, and Pelz, J. B. (2016). Recurrence metrics for assessing eye movements in perceptual experiments. *Journal of Eye Movement Research*, 9(4).

(4) **Vaidyanathan, P.**, Prudhommeaux, E., Alm, C. O., Pelz, J. B., and Haake, A. R. (2016). Fusing eye movements and observer narratives for expert-driven image region annotations. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 27-34. ACM. (**Best Paper Award**)

(5) **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. In *Proceedings of the International Conference on Computational Semantics*, pages 76-82. ACL.

(6) **Vaidyanathan, P.**, Prudhommeaux, E., Alm, C. O., and Pelz, J. B. (2015b). Computational integration of human vision and natural language through bitext

alignment. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 4-5. ACL.

(7) **Vaidyanathan, P.**, Pelz, J. B., Alm, C. O., Shi, P., and Haake, A. R. (2014). Recurrence quantification analysis reveals eye movement differences between experts and novices. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 303-306. ACM.

(8) **Vaidyanathan, P.**, Pelz, J. B., Alm, C. O., Calvelli, C., Shi, P., and Haake, A. R. (2013). Integration of eye movements and spoken description for medical image understanding. In Holmqvist, K., Mulvey, F., and Johansson, R., editors, *Book of Abstracts of the 17th European Conference on Eye Movements*, pages 40-41. EMRA.

(9) Wang, D., **Vaidyanathan, P.**, Haake, A., and Pelz, J. (2012). Are eye trackers always as accurate as we assume? In *Annual Meeting of the Society for Computers in Psychology.*

(10) **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 *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 3-4. AMIA.

(11) **Vaidyanathan, P.**, Pelz, J. B., Li, R., Mulpuru, S., Wang, D., Shi, P., Calvelli, C., and Haake, A. R. (2011). Using human experts' gaze data to evaluate image processing algorithms. In *Proceedings of the IEEE Image, Video, and Multidimensional Signal Processing Workshop*, pages 129-134.

(12) Li, R., **Vaidyanathan, P.**, Mulpuru, S., Pelz, J. B., Shi, P., Calvelli, C., and Haake, A. R. (2010). Human-centric approaches to image understanding and retrieval. In *Proceedings of the IEEE Western New York Image Processing Workshop*, pages 62-65.

Bibliography

[Anderson et al., 2013] Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F., and Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior Research Methods*, 45:842–856.

[Badler, 1975] Badler, N. I. (1975). *Temporal Scene Analysis: Conceptual Descriptions of Object Movements.* PhD thesis, University of Toronto, Toronto, Canada.

[Ballerini et al., 2009] Ballerini, L., Li, X., Fisher, R. B., and Rees, J. (2009). A query-by-example content-based image retrieval system of non-melanoma skin lesions. In *Proceedings of the First MICCAI International Conference on Medical Content-Based Retrieval for Clinical Decision Support*, pages 31–38. ACM.

[Barnard et al., 2003] Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

[Berg et al., 2004a] Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. (2004a). Who's in the picture? *Advances in Neural Information Processing Systems*, 17:137–144.

[Berg et al., 2004b] Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y. W., Learned-Miller, E. G., and Forsyth, D. A. (2004b). Names and faces in the news. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 848–854.

[Beyer and Holtzblatt, 1997] Beyer, H. and Holtzblatt, K. (1997). *Contextual Design: Defining Customer-Centered Systems.* Elsevier.

[Boersma, 2002] Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

[Borji, 2009] Borji, A. (2009). *Interactive Learning of Task-Driven Visual Attention Control*. PhD thesis, Institute for Research in Fundamental Sciences (IPM), School of Cognitive Sciences (SCS), Tehran, Iran.

[Bosman et al., 2010] Bosman, H., Petkov, N., and Jonkman, M. (2010). Comparison of color representations for content-based image retrieval in dermatology. *Skin Research and Technology*, 16(1):109–113.

[Brown et al., 1993] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

[Bullard et al., 2014] Bullard, J., Alm, C. O., Yu, Q., Shi, P., and Haake, A. (2014). Towards multimodal modeling of physicians' diagnostic confidence and self-awareness using medical narratives. In *Proceedings of the International Conference on Computational Linguistics*, pages 1718–1727.

[Castelhano et al., 2009] Castelhano, M., Mack, M., and Henderson, J. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):1–15.

[Chang and Hsu, 1992] Chang, S.-K. and Hsu, A. (1992). Image information systems: Where do we go from here? *IEEE Transactions on Knowledge and Data Engineering*, 4(5):431–442.

[Clarke et al., 2013] 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.

[Coco and Keller, 2012] Coco, M. I. and Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, 36(7):1204–1223.

[Cooper, 1974] Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1):84–107.

[Dahan et al., 2001] Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42(4):317–367.

[Duchowski, 2017] Duchowski, A. (2017). *Eye Tracking Methodology*. Springer-Verlag London.

[Duygulu et al., 2002] 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 *Proceedings of the European Conference on Computer Vision*, pages 97–112.

[Fei-Fei and Perona, 2005] Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531.

[Ferreira and Henderson, 2004] Ferreira, F. and Henderson, J. M. (2004). *The Interface of Language, Vision and Action: Eye Movements and the Visual World*. Psychology Press.

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

[Forsyth et al., 2009] Forsyth, D. A., Berg, T., Alm, C. O., Farhadi, A., Hockenmaier, J., Loeff, N., and Wang, G. (2009). Words and pictures: Categories, modifiers, depiction, and iconography. In *Object Categorization: Computer and Human Vision Perspectives*, pages 167–181. Cambridge University Press.

[Gabbett and Abernethy, 2013] Gabbett, T. J. and Abernethy, B. (2013). Expert–novice differences in the anticipatory skill of rugby league players. *Sport, Exercise, and Performance Psychology*, 2(2):138–155.

[Gangji et al., 2017] 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*, pages 621–627.

[Goldstone, 1998] Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology*, 49(1):585–612.

[Green and Swets, 1966] Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*, volume 1. Wiley New York.

[Griffin, 2004] Griffin, Z. M. (2004). Why look? Reasons for eye movements related to language production. In Henderson, J. M. and Ferreira, F., editors, *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, pages 213–248. Psychology Press.

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

[Guo et al., 2014a] Guo, X., Li, R., Alm, C., Yu, Q., Pelz, J., Shi, P., and Haake, A. (2014a). Infusing perceptual expertise and domain knowledge into a human-centered image retrieval system: A prototype application. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 275–278. ACM.

[Guo et al., 2014b] Guo, X., Yu, Q., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., and Haake, A. R. (2014b). From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding. *Artificial Intelligence in Medicine*, 62(2):79–90.

[Heller, 1988] Heller, D. (1988). On the history of eye movement recording. In Luer, G., Lass, U., and Hoffman, J. S., editors, *Eye Movement Research: Physiological and Psychological Aspects*, pages 37–51. Toronto: CJ Hogrefe.

[Herzog and Wazinski, 1994] Herzog, G. and Wazinski, P. (1994). Visual translator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2-3):175–187.

[Hochberg et al., 2014a] Hochberg, L., Alm, C. O., Rantanen, E. M., DeLong, C. M., and Haake, A. (2014a). Decision style in a clinical reasoning corpus. In *Proceedings of the BioNLP Workshop*, pages 83–87. ACL.

[Hochberg et al., 2014b] Hochberg, L., Alm, C. O., Rantanen, E. M., Yu, Q., DeLong, C. M., and Haake, A. (2014b). Towards automatic annotation of clinical decision-making style. In *Proceedings of the 8th Linguistic Annotation Workshop*, pages 129–138. ACL.

[Hoffman and Fiore, 2007] Hoffman, R. and Fiore, S. (2007). Perceptual (re) learning: A leverage point for human-centered computing. *IEEE Intelligent Systems*, 22(3):79–83.

[Holsanova, 2006] Holsanova, J. (2006). Dynamics of picture viewing and picture description. *Advances in Consciousness Research*, 67:235–256.

[Holsanova, 2008] Holsanova, J. (2008). *Discourse, Vision, and Cognition*, volume 23. John Benjamins Publishing Company.

[IBM, 2015] IBM (2015). IBM Watson Speech to Text. `https://www.ibm.com/watson/developercloud/speech-to-text.html`. (Date last accessed 16-Aug-2016).

[Jaber and Saber, 2010] Jaber, M. I. and Saber, E. (2010). Probabilistic approach for extracting regions of interest in digital images. *Journal of Electronic Imaging*, 19(2):023019–1–023019–13.

[Jain and Vailaya, 1996] Jain, A. K. and Vailaya, A. (1996). Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244.

[Jamieson et al., 2006] Jamieson, M., Dickinson, S., Stevenson, S., and Wachsmuth, S. (2006). Using language to drive the perceptual grouping of local image features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2102–2109.

[Ji and Ploux, 2003] Ji, H. and Ploux, S. (2003). A mental lexicon organization model. In *Proceedings of the Joint International Conference on Cognitive Science*, pages 240–245.

[Johnson et al., 2015] Johnson, J., Ballan, L., and Li, F.-F. (2015). Love thy neighbors: Image annotation by exploiting image metadata. *arXiv preprint arXiv:1508.07647*.

[Just and Carpenter, 1976] Just, M. A. and Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480.

[Just and Carpenter, 1980] Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354.

[Kaiser and Trueswell, 2008] Kaiser, E. and Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, 23(5):709–748.

[Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

[Kiela et al., 2014] Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of Association of Computation Linguistics*, pages 835–841.

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

[Krupinski, 2000] Krupinski, E. (2000). The importance of perception research in medical imaging. *Radiation Medicine*, 18(6):329–334.

[Kumar et al., 2010] Kumar, M., Torr, P. H., and Zisserman, A. (2010). Objcut: Efficient segmentation using top-down and bottom-up cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):530–545.

[Kuznetsova et al., 2013] Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. (2013). Generalizing image captions for image-text parallel corpus. In *Proceedings of Association of Computation Linguistics*, pages 790–796.

[Leigh and Zee, 2015] Leigh, J. and Zee, D. (2015). *The Neurology of Eye Movements*. Oxford University Press, Oxford.

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

[Li et al., 2009] Li, L.-J., Socher, R., and Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036–2043.

[Li et al., 2012] Li, R., Pelz, J. B., Shi, P., Alm, C. O., and Haake, A. R. (2012). Learning eye movement patterns for characterization of perceptual expertise. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 393–396. ACM.

[Li et al., 2013] Li, R., Shi, P., and Haake, A. R. (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*, pages 2187–2194.

[Li et al., 2016] Li, R., Shi, P., Pelz, J., Alm, C. O., and Haake, A. R. (2016). Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes. *Computer Vision and Image Understanding*, 151:138–152.

[Li et al., 2010] Li, R., Vaidyanathan, P., Mulpuru, S., Pelz, J. B., Shi, P., Calvelli, C., and Haake, A. R. (2010). Human-centric approaches to image understanding and retrieval. In *Proceedings of the IEEE Western New York Image Processing Workshop*, pages 62–65.

[Li et al., 2015] Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G., and Del Bimbo, A. (2015). Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *arXiv preprint arXiv:1503.08248*.

[Liang et al., 2006] Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 104–111.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755.

[Lipps and Pelz, 2004] Lipps, M. and Pelz, J. B. (2004). Yarbus revisited: Task-dependent oculomotor behavior. *Journal of Vision*, 4(8):115–115.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1150–1157.

[Malcolm and Henderson, 2010] Malcolm, G. L. and Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, 10(2):4, 1–11.

[Maninis et al., 2017] Maninis, K., Pont-Tuset, J., Arbeláez, P., and Gool, L. V. (2017). Convolutional oriented boundaries: From image segmentation to high-level tasks. *arXiv:1701.04658*.

[McCoy et al., 2012a] McCoy, W., Alm, C. O., Calvelli, C., Li, R., Pelz, J. B., Shi, P., and Haake, A. R. (2012a). Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 95–103. ACL.

[McCoy et al., 2012b] McCoy, W., Pelz, J. B., Alm, C. O., Shi, P., Calvelli, C., and Haake, A. R. (2012b). Linking uncertainty in physicians' narratives to diagnostic correctness. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 19–27. ACL.

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

[Mishra et al., 2009] Mishra, A., Aloimonos, Y., and Fah, C. (2009). Active segmentation with fixation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 468–475.

[Müller et al., 2004] 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. *International Journal of Medical Informatics*, 73(1):1–23.

[Naim et al.,] Naim, I., Song, Y. C., Liu, Q., Kautz, H., Luo, J., and Gildea, D. Unsupervised alignment of natural language instructions with video segments. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1558–1564.

[Och and Ney, 2003] Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

[Och et al., 2000] Och, F. J., Tillmann, C., and Ney, H. (2000). Improved alignment models for statistical machine translation. In *Proceedings of Association of Computation Linguistics*, pages 440–447.

[Oliva et al., 2003] Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Proceedings of the IEEE International Conference on Image Processing*, pages 253–256.

[Pelz and Canosa, 2001] Pelz, J. B. and Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41(25):3587–3596.

[Petrov and Klein, 2007] Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 404–411.

[Phillips and Edelman, 2008] Phillips, M. H. and Edelman, J. A. (2008). The dependence of visual scanning performance on saccade, fixation, and perceptual metrics. *Vision Research*, 48(7):926–936.

[Pollatsek et al., 1993] Pollatsek, A., Raney, G. E., Lagasse, L., and Rayner, K. (1993). The use of information below fixation in reading and in visual search. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(2):179–200.

[Qu and Chai, 2008] 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*, pages 244–253. ACL.

[Rayner, 1998] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

[Richardson and Dale, 2005] Richardson, D. C. and Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6):1045–1060.

[Riche et al., 2013] Riche, N., Duvinage, M., Mancas, M., Gosselin, B., and Dutoit, T. (2013). Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 755–762.

[Roy, 2000] Roy, D. (2000). Integration of speech and vision using mutual information. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2369–2372.

[Roy and Pentland, 2002] Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

[Saber et al., 1996] Saber, E., Tekalp, A. M., Eschbach, R., and Knox, K. (1996). Automatic image annotation using adaptive color classification. *Graphical Models and Image Processing*, 58(2):115–126.

[Santella and DeCarlo, 2004] Santella, A. and DeCarlo, D. (2004). Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 27–34. ACM.

[Scheirer et al., 2014] Scheirer, W. J., Anthony, S. E., Nakayama, K., and Cox, D. D. (2014). Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1679–1686.

[Sensomotoric Instruments, 2016] Sensomotoric Instruments (2016). Sensomotoric Instruments. `https://www.smivision.com/`. (Date last accessed 16-Aug-2016).

[Shadbolt and Smart, 2015] Shadbolt, N. and Smart, P. (2015). Knowledge elicitation: Methods, tools and techniques. In *Evaluation of Human Work*, pages 163–200. CRC Press.

[Shanteau, 1992] Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, 81(1):75–86.

[Shao et al., 2013] Shao, Z., Roelofs, A., and Meyer, A. (2013). Predicting naming latencies for action pictures: Dutch norms. *Behavior Research Methods*, 46:274–283.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

[Shin et al., 2002] Shin, M. C., Chang, K. I., and Tsap, L. V. (2002). Does colorspace transformation make any difference on skin detection? In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, pages 275–279.

[Shyu et al., 1999] Shyu, C.-R., Brodley, C. E., Kak, A. C., Kosaka, A., Aisen, A. M., and Broderick, L. S. (1999). Assert: A physician-in-the-loop content-based retrieval system for HRCT image databases. *Computer Vision and Image Understanding*, 75(1):111–132.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1470–1477.

[Smeulders et al., 2000] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

[Socher et al., 2014] Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

[Spivey et al., 2002] Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., and Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481.

[Srihari, 1995] Srihari, R. K. (1995). Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56.

[Stark and Privitera, 1997] Stark, L. W. and Privitera, C. (1997). Top-down and bottom-up image processing. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 2294–2299.

[Takiwaki, 1998] Takiwaki, H. (1998). Measurement of skin color: Practical application and theoretical considerations. *Journal of Medical Investigation*, 44:121–126.

[Tamura and Yokoya, 1984] Tamura, H. and Yokoya, N. (1984). Image database systems: A survey. *Pattern Recognition*, 17(1):29–43.

[Tanaka et al., 2005] Tanaka, J., Curran, T., and Sheinberg, D. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science*, 16(2):145–151.

[Tanenhaus et al., 1995] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

[Tang et al., 1999] Tang, L. H., Hanka, R., and Ip, H. H. (1999). A review of intelligent content-based indexing and browsing of medical images. *Health Informatics Journal*, 5(1):40–49.

[Tatler, 2007] Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4–4.

[Tavakoli et al., 2017] Tavakoli, H. R., Shetty, R., Borji, A., and Laaksonen, J. (2017). Can saliency information benefit image captioning models? *arXiv preprint arXiv:1704.07434*.

[Thomason et al., 2014] Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., and Mooney, R. (2014). Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics*.

[Treisman and Gelade, 1980] Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

[Ugarriza et al., 2009] Ugarriza, L. G., Saber, E., Vantaram, S. R., Amuso, V., Shaw, M., and Bhaskar, R. (2009). Automatic image segmentation by dynamic region growth and multiresolution merging. *IEEE Transactions on Image Processing*, 18(10):2275–2288.

[Ullman, 2000] Ullman, S. (2000). *High-Level Vision: Object Recognition and Visual Cognition*. MIT press.

[Vaidyanathan et al., 2013] Vaidyanathan, P., Pelz, J. B., Alm, C. O., Calvelli, C., Shi, P., and Haake, A. R. (2013). Integration of eye movements and spoken description for medical image understanding. In Holmqvist, K., Mulvey, F., and Johansson, R., editors, *Book of Abstracts of the 17th European Conference on Eye Movements*, pages 40–41. EMRA.

[Vaidyanathan et al., 2014] Vaidyanathan, P., Pelz, J. B., Alm, C. O., Shi, P., and Haake, A. R. (2014). Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 303–306. ACM.

[Vaidyanathan et al., 2011] Vaidyanathan, P., Pelz, J. B., Li, R., Mulpuru, S., Wang, D., Shi, P., Calvelli, C., and Haake, A. R. (2011). Using human experts' gaze data to

evaluate image processing algorithms. In *Proceedings of the IEEE Image, Video, and Multidimensional Signal Processing Workshop*, pages 129–134.

[Vaidyanathan et al., 2012] 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 *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 3–4. AMIA.

[Vaidyanathan et al., 2015a] Vaidyanathan, P., Prud'hommeaux, E., Alm, C. O., Pelz, J. B., and Haake, A. R. (2015a). Alignment of eye movements and spoken language for semantic image understanding. In *Proceedings of the International Conference on Computational Semantics*, pages 76–82. ACL.

[Vaidyanathan et al., 2015b] Vaidyanathan, P., Prud'hommeaux, E., Alm, C. O., Pelz, J. B., and Haake, A. R. (2015b). Computational integration of human vision and natural language through bitext alignment. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 4–5. ACL.

[Vaidyanathan et al., 2016] Vaidyanathan, P., Prud'hommeaux, E., Alm, C. O., Pelz, J. B., and Haake, A. R. (2016). Fusing eye movements and observer narratives for expert-driven image-region annotations. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 27–34. ACM.

[Vakkari, 2002] Vakkari, P. (2002). Subject knowledge, source of terms, and term selection in query expansion: An analytical study. In *Advances in Information Retrieval*, pages 110–123. Springer.

[van der Meulen, 2003] 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 et al., 2014] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

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

[Walther et al., 2005] Walther, D., Rutishauser, U., Koch, C., and Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1):41–63.

[Waltz, 1980] Waltz, D. L. (1980). Generating and Understanding Scene Descriptions. Technical report, DTIC Document.

[Wang et al., 2012a] Wang, D., Vaidyanathan, P., Haake, A., and Pelz, J. (2012a). Are eye trackers always as accurate as we assume? In *Annual meeting of the Society for Computers in Psychology*.

[Wang et al., 2012b] Wang, X., Erdelez, S., Allen, C., Anderson, B., Cao, H., and Shyu, C.-R. (2012b). Role of domain knowledge in developing user-centered medical-image indexing. *Journal of the American Society for Information Science and Technology*, 63(2):225–241.

[Webber and Zbilut, 1994] Webber, C. and Zbilut, J. P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973.

[Williams and Elliott, 1999] Williams, A. M. and Elliott, D. (1999). Anxiety, expertise, and visual search strategy in karate. *Journal of Sport & Exercise Psychology*.

[Womack et al., 2013] Womack, K., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., and Haake, A. R. (2013). Using linguistic analysis to characterize conceptual units of thought in spoken medical narratives. In *Proceedings of the INTERSPEECH*, pages 3722–3726.

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

[Wooding, 2002] Wooding, D. (2002). Fixation maps: Quantifying eye-movement traces. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 31–36. ACM.

[Yarbus, 1965] Yarbus, A. (1965). *Role of eye movements in the visual process.* Nauka Press, Moscow.

[Yarbus et al., 1967] Yarbus, A., Haigh, B., and Rigss, L. (1967). *Eye Movements and Vision*. 2nd edition Plenum Press New York.

[Yatskar et al., 2016] Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

[Yu and Ballard, 2004a] 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 and Ballard, 2004b] 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*, pages 488–493.

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

[Yun et al., 2013b] Yun, K., Peng, Y., Samaras, D., Zelinsky, G. J., and Berg, T. L. (2013b). Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746.

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

[Zhu et al., 2016] Zhu, H., Meng, F., Cai, J., and Lu, S. (2016). Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27.

[Zitnick et al., 2016] Zitnick, C. L., Vedantam, R., and Parikh, D. (2016). Adopting abstract images for semantic scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):627–638.