Fall 2011

Truthmarker: a tablet-based approach for rapid image annotation

Mark Allen Christopher
University of Iowa

TRUTHMARKER: A TABLET-BASED APPROACH FOR RAPID IMAGE ANNOTATION

by

Mark Allen Christopher

A thesis submitted in partial fulfillment
of the requirements for the Master of
Science degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

December 2011

Thesis Supervisor: Associate Professor Todd E. Scheetz

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

Mark Allen Christopher

has been approved by the Examining Committee
for the thesis requirement for the Master of Science
degree in Biomedical Engineering at the December 2011 graduation.

Thesis Committee: _____
Todd E. Scheetz, Thesis Supervisor

_____
Michael D. Abramoff

_____
Joseph M. Reinhardt

_____
Mona K. Garvin

_____
Gary E. Christensen

ACKNOWLEDGMENTS

I would first like to thank Todd Scheetz and Michael Abramoff. They have been great mentors to me for this and many other projects. I would also like to thank all of the students, staff, and faculty in the CBCB, especially Kyle Taylor whose help with this project has been invaluable. Finally, I want to thank my friends and family for all the support and encouragement they have provided me over the years.

ABSTRACT

The development of automated techniques for the analysis of image data is an important and active area of research. To make progress, this research requires annotations of image data to build and validate models used for analysis. Given this requirement, the development of software tools that streamline the collection of annotations would be of great benefit to image analysis researchers. Such tools should meet the following requirements: rapid generation of annotations for large data sets, annotation and data management that is straightforward for users, flexibility for application to many diverse image datasets, configurability to allow the collection of annotations to be tuned for a specific research goal, and generation of annotation data in a standardized format so that it can be easily parsed and analyzed. Truthmarker was designed as a tablet computer based image annotation tool to meet these requirements. Researchers can configure Truthmarker to fit the needs of a particular study by specifying an annotation model that fine tunes the user interface and resulting data to fit the annotation task. The quality of annotations generated using Truthmarker was evaluated by recruiting medical experts to annotate ophthalmic images for severity of diabetic retinopathy, a leading cause of blindness. These annotations were compared to annotations of the same images assigned using standard desktop computer based tools. The results, as measured by κ statistics and accuracy, indicate that Truthmarker annotations were of equivalent quality compared to those that were created using desktop-based tools.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

Image analysis and computer vision are active areas of research with implications for many other fields including medicine. This research seeks to build models for automated processing and analyses of image data. Developing these models requires not only images but human assigned annotations for the images. These image annotations provide context for and highlight important implicit information of the images. This information is required not only to build models used in image analysis, but also to evaluate. Without annotations for large image data sets, this research could not make progress.

Gathering image annotations can be problematic, though, because it requires human, or even expert, annotation of images. Gathering these annotations for large image data sets can be tedious, time-consuming, and expensive. Streamlining the collection of annotations would allow large, diverse image sets to be utilized in research. This would lead to better models for image analysis to be constructed and help progress related research fields.

Providing software that allows quick annotation of images is important, but does not meet all the needs of researchers. Particular image analysis studies are often focused on a specific type of image or image feature. Usefulness for a wide range of studies requires that the software can be fine-tuned by researchers for a specific annotation task. This allows precisely defined annotation data to be both quickly and reliably gathered.

Truthmarker has been developed as a tablet computer based tool to address the annotation data collection needs of image analysis researchers. This tool can be customized to address the needs of a particular research study and allow annotations to be quickly gathered for large image data sets. The following chapters will discuss important background related to image analysis and annotation, describe the general design

approach of Truthmarker, and detail the implementation. A study performed to evaluate the quality of annotations generated using Truthmarker is also presented.

CHAPTER 2

BACKGROUND

2.1 Image Annotation

Image annotation is the process of associating metadata with a digital image. The annotations might provide data regarding where, how, and when the image was collected, or the annotations could provide semantic information about what the image data actually means. Performing this semantic annotation by applying informative terms or tags to an image or image region provides information that is difficult to infer from the image data itself. The problem is that while images contain large amounts of data, the meaning of this data not explicit. Semantic annotation provides context for image data and allows meaning to be easily accessed. This, in turn, allows large image datasets to be more efficiently stored, queried, and analyzed.

For the purposes of this work, it is useful to make a distinction between two general categories of image annotations: image-level and pixel-level annotations. Image-level annotations consist of information meant to describe an entire image. The general category of an image, whether or not it contains a particular feature of interest, or represents a specific type of scene are examples of image-level annotations. Pixel-level annotations are used to mark-up particular regions of interest (ROI) within an image. These annotations are used to localize individual objects within an image and to segment out ROIs from the background.

2.1.1 Medical Image Annotations

In the case of medical images, annotators are interested associating medically relevant information with particular images. These annotations could consist of image-level determinations such as disease states, patient demographic information, and risk factors associated with an image. They could also consist of pixel-level information

marking ROIs associated with anatomical structures, abnormalities, and markers for disease present within an image.

2.1.2 Computational Image Analysis

Generating annotations for large sets of image is a crucial aspect of research into computational analysis of image data. Image analysis and, specifically, computer vision are well-established fields that seek to develop automated methods for extracting meaningful information from images. This can include categorizing images[1,2], segmenting structures of interest with in an image[3,4], or extracting any other useful information. These fields rely heavily on machine learning work to classify, cluster, and analyze features extracted from image data.

Within the context of machine learning, two general approaches are typically applied: supervised and unsupervised learning. In supervised learning, a set of training data points and associated true category labels are used as input. Using the data and labels, rules for what separating different categories are estimated. Unsupervised learning techniques such as clustering do not use true labels as input. Instead, they attempt to find natural divisions based on the data distribution.[5] When applying machine learning techniques to image analysis, human generated annotations are required as input. In the case of supervised techniques such as classifiers, annotations indicating true labels for image data are needed in order to build and train the models used for classification. Additionally, both supervised and unsupervised techniques require annotations of truth so that quantitative measures of performance that allow researchers to effectively compare techniques can be computed.

Currently, there are well-annotated image datasets such as the PASCAL[6], Caltech-101[7], and Caltech-256[8] publicly available for use in training and testing computational image analysis techniques. These datasets consist of large numbers of images each placed into a category. Example image categories of the Caltech-256 dataset

include vehicles, buildings, and faces. These datasets are useful for general image analysis or computer vision tasks, but there are two important aspects of these datasets that can preclude them from use by many researchers. The first is that the annotations of these images are image-level. The data are closely cropped images of a single object or scene and a category assigned to each image. This limits the usefulness to researchers interested in developing methods for detecting and extracting objects of interest from images. Pixel-level annotations are needed to define the boundaries of these objects. The second aspect is the generic nature of the datasets. While this makes them useful for a large number of computer vision researchers, it prevents them from being useful for domain-specific analyses. For instance, a researcher interested in developing analysis techniques for a particular type of medical image, would have little use for images of airplanes and buildings. These limitations, especially the second, mean that researchers are always in need of new image and annotation data.

2.2 Medical Imaging

Databases of medical images such digital photographs, CT, MRI, X-ray, scanned patient charts, and other images can contain millions images with more being collected constantly. Obviously, these images serve as invaluable diagnostic and prognostic tools in a clinical setting, but they are also extremely valuable for medical and image analysis research. Research in applying existing and developing new image analysis techniques for medical images is currently a very active. The goal of this research is to make the huge stores of medical image data more manageable by applying automated techniques for analyzing the images.

2.2.1 Fundus Imaging

One type of medical image is of particular interest for this work: retinal fundus images. The fundus is the interior surface of the eye that can be viewed or imaged through the pupil. Figure 1 shows a typical fundus image in which the optic disc,

macula, and retinal vasculature are visible. Fundus images are used by ophthalmologists to screen patients for a number of eye diseases.[9]

Figure 2.1: An example fundus image.[10] The image was taken through the pupil with a specialized camera. Several important anatomical structures are visible, including the optic disc (the bright circular region), the macula (the darker circle to the right of the optic disc), and the retinal vasculature (diverging from the optic disc).

One such disease is diabetic retinopathy (DR). This disease is a complication of diabetes and is one of the leading causes of blindness among adults in America.[11] Using fundus images, ophthalmologists can determine the presence and severity of DR based on features of the retinal vasculature, presence of microaneurysms, and other abnormalities. Regular screening of patients at risk for DR has been shown to help prevent loss of vision.[11] Telemedical and internet-based screening programs where many at risk patients are quickly imaged and the images are later viewed by ophthalmologists have been

developed as an efficient way to screen large numbers of patients very quickly.[9] A serious problem for this type technique, though, is the sheer number of images. Recent work in developing computational methods for recognizing DR in fundus images has been performed with the hope of reducing the workload on physicians.[12,13] This work, of course, requires large sets fundus image annotations indicating presence and severity of DR for training and validation purposes.

2.3 Previous Image Annotation Work

Previous work addressing image annotation, especially with regard to medical images, has been concerned with two primary tasks: defining a standard format for semantic annotations and developing tools for performing the annotation. Considering the first task, there are commonly used standards for storage and transfer of medical images and associated metadata such as Digital Imaging and Communications in Medicine (DICOM) and Health Level Seven (HL7). Both DICOM and HL7 are primarily focused on metadata important for data management such as where, when, and by whom the image was acquired.[14,15] There is no standard for including semantic annotation data within DICOM or HL7.

The problem of developing standards for the representation of biologically and medically important semantic annotation of data has been an ongoing challenge for researchers. Large-scale collection of medical images and biological data using high-throughput collection techniques has spurred the development of these standards. Semantic annotation within a defined standard can speed the progress of research by making these large stores of data easier to manage, share, and analyze. The problem with developing these standards, though, is that they require the integration of the disparate terminologies used by different researchers and in different contexts. The Human Physiome Project[16] is one effort that has attempted to create a standard for annotation and modeling of biological data. This project defines ontologies and associated standards for

data representation and is intended as a comprehensive framework for modeling and associating important annotations with biological data. The Human Physiome Project framework provides standards for the representation of biological data of all scales including molecular, cellular, tissue, organ, and whole body data. The ontologies and representations developed as part of this effort could be applied to a wide variety of medical images in order provide a standard for annotation data.

Another proposed annotation standard, has been proposed by Rubin *et al.* based on the predefined medical ontology developed as part of the caBIG project.[17] This project aims to develop a data sharing and collaboration network for cancer researchers. In the standard, each image- and pixel-level annotation is associated with a semantic term taken from a predefined ontology. The utility of these medical standards, though, is limited by their rigidity. Not only are they not applicable non-medical image annotation, they require that researchers using medical images conform to existing medical ontologies. Researchers involved with the caBIG project have also developed software for performing image annotation such as the Annotation and Image Markup (AIM) tool.[18] However, these tools are also limited in their utility because of the requirement that all annotations conform to the ontology.

More general tools that allow annotation of all types of images and that do not require conformation to any particular ontology have also been developed. For instance, LabelME is a web-based annotation tool that allows image- and pixel-level annotation.[19] LabelME and similar tools are meant for annotating large, general image datasets, but do not allow control over what types of annotations can be generated. The lack of structure of these tools prevents expert domain knowledge from being easily incorporated into the annotation process. Ideally, there would be image annotation software more flexible than the rigid AIM tool, but that could allow more structure to be applied to the annotations than web-based tools such as LabelME. In particular, researchers could define set of

annotations, images, and interface elements in order to fine tune the annotation for addressing a specific research question.

2.4 Medical Use of Tablets

2.4.1 Current Applications

Tablet computers including the Apple iPad as well as several tablet PCs[20-22] have recently becomes widely-available to consumers. These modern tablets have a high storage capacity for data while maintaining portability and a compact size. These have made them an attractive alternative to standard desktop or laptop computers in clinical settings. They can allow physicians to access to large stores of patient records and medical images while interacting with patients face-to-face. Research indicating that use of standard desktop computers to view medical records can negatively impact patient perception of care only serves to further incentivize healthcare providers to adopt the use of tablets.[23] Vendors of electronic heath record systems have begun to respond by developing tablet-based software for clinical viewing of medical records and images.[24,25]

2.4.2 Application to Medical Imaging Research

Tablets can provide benefits to research as well as clinical use of medical images. Consider how research involving medical images is often performed. A researcher, or study designer, has a specific research question to address with regard to computational analysis of medical images. The study designer may lack the expertise required to annotate the images for the medical features of interest. Even if the designer has the expertise, he or she may lack the time required to perform annotation or may require multiple independent assessments of the images to provide a robust set of annotations. These considerations require study designers to recruit experts to serve as annotators of medical image datasets. Reducing the effort required of the expert, who may receive many such requests, can make the recruitment process easier.

Tablet computers provide a convenient platform for the delivery of both the image data and the software required to perform the annotation. Annotation software developed for a tablet platform can be installed and large image datasets can be loaded onto the tablet, which can then be handed off to the annotator. With this model, the burden on the annotator is low because no effort is required to install software or access the images; they simply have to launch an application on the tablet. As in the case of clinical use of tablets, the portability of the platform allows annotation to be performed without being tied to a desktop or laptop computer, further reducing the burden on the annotator. The generated annotations can be returned to the designer either by physically delivering the tablet or via the internet.

2.4.3 Concerns of Tablet Use

The use of tablets in clinical or medical research settings raises some concerns. While there is no reason to suspect that medical information represented as text or numeric data would be interpreted differently on a tablet compared to standard desktop computer, the same is not true of medical images. Differences in the displays and user interaction could result in differences in the interpretation medically relevant image features.

Previous work examining the effect of display and environmental characteristics on visual perception has shown that several factors can affect the interpretation of images. First, the resolution of used to display images both in terms of overall number of pixels and pixels per inch has been shown to have a significant impact on the interpretation. Low resolution displays can decrease the performance in visual evaluation tasks as well as increasing the time need to complete the task and the associated visual fatigue compared to higher resolution displays.[26] These effects raise concerns regarding the use of tablets for image evaluation and annotations because current tablet display resolutions are lower than those of many commonly used LCD desktop displays. Second,

display luminance and contrast, which describe the overall brightness and the difference between light and dark areas of a display, respectively, have also been shown to affect image interpretation. Generally, higher contrast and luminance result in improved perceived image quality and performance in visual tasks.[27,28] Here, again, tablets fall short with respect to typical desktop displays raising even more concerns with regarding their use for image annotation. Finally, characteristics of the environment in which the images are viewed can influence interpretation. Specifically, the brightness and color temperature of ambient lighting has implications for evaluation of images.[29] Given their portability, tablets are, of course, more likely to be used in a broader range of environments with varying ambient light and color characteristics. This is yet another concerning aspect of the use of tablets for image annotation.

Recently, the FDA released draft guidelines and is accepting comments regarding the use of tablet for the review of images and other medical data.[30] These guidelines specifically express concern regarding the evaluation of medical images using tablets noting some of display and environmental factors listed above. To be confident in the medical annotations generated using tablet-based systems, comparisons of annotations created using these systems to those created using standard desktop tools need to be performed.

CHAPTER 3

APPROACH

3.1 Problem Statement

The goal of this work is to design an image annotation system that can be used by used by researchers to quickly gather annotations for their image sets of interest. An application called Truthmarker has been developed to address this goal. In order to explain the approach used to design Truthmarker, a more precise statement of the problem is needed.

Truthmarker is designed to meet the following user requirements:

1. Truthmarker should be a tablet-based application. This allows the convenience, portability, and ease-of-use of tablet computers to be leveraged in order to more efficiently gather annotation data.
2. Truthmarker should be flexible enough handle many types of images and annotations. This is required so that it is useful for researchers considering many diverse datasets.
3. Truthmarker should allow researchers to apply domain-specific requirements to the collection of annotations. This that means configurable controls and structure can be applied to image annotation.
4. Truthmarker should provide interfaces for data management and annotation that are easy to understand and use. This reduces the burden associated with annotation and facilitates the rapid generation of data.
5. Truthmarker should produce annotation data in a standardized format. This standardized format should allow annotations to be efficiently parsed, analyzed, and used in research.

The following sections will describe the expected use cases of Truthmarker, the general approaches designed to address user requirements, and explicitly define the system requirements.

3.2 Expected Use Cases

Truthmarker has two general types of expected users: study designers and annotators. Note that these types of users are not necessarily distinct. Indeed, study designers may often act as annotators for their own data. Users acting in these different capacities, though, will interact with Truthmarker in unique ways and have distinct requirements. Consider, again, the example of an image analysis researcher performing a study related to medical images. The researcher would act as the study designer, in this case. They could use their expertise to define the dataset, scope, and analysis of resulting annotations for the study, but may lack the medical knowledge to actually perform the annotation. A medical expert would act as the annotator and apply their knowledge to annotate the images without necessarily having to concern themselves with dataset selection, management, or analysis.

3.2.1 Study Designer

Study designers are individuals using Truthmarker to gather annotation data that will help address a specific research question. These users first assemble an image dataset that requires annotation. Study designers then need to decide on a set of image- and pixel-level annotations and associated semantically-informative terms that are useful for addressing their research question. This image dataset and the associated set of desired annotations define an annotation task. With the annotation task defined, study designers can then load the images and annotation specification onto a tablet. This tablet can then be passed on to an annotator who will perform annotation according to the specification provided by the designer. Once the annotation task is completed, the

annotation data associated with each image returned to the designer and they can use the annotations in a research application.

3.2.2 Annotator

Annotators use the Truthmarker application to actually perform annotation task. Annotators receive a tablet with Truthmarker installed and image data pre-loaded. The annotator then uses Truthmarker to provide appropriate annotations for the task as defined by the study designer. Figure 3.1 diagrams the interactions between the study designer, annotator, and annotation task. The main requirements of annotators are related to the user interfaces provided by Truthmarker during annotation. An interface that is easy to understand and use can help lower the burden and result in more reliable annotations.

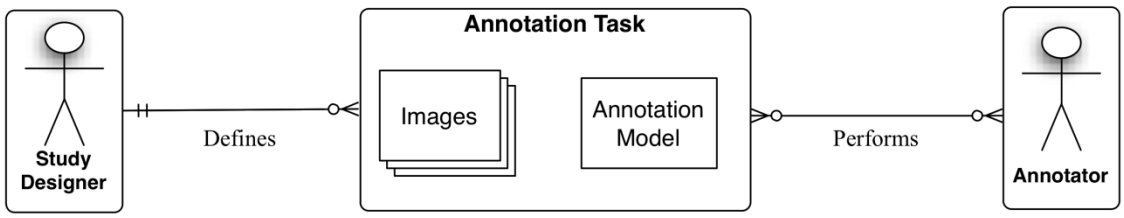

Figure 3.1: This diagram shows interactions between the study designer, annotator, and an annotation task. An annotation task includes a set of images to be annotated and a specification of the desired annotations (annotation model). The study designer defines an annotation task by specifying the image set and model. The annotator then performs the task using Truthmarker. Note that the connectors indicate the cardinality of the relationships. For instance, a study designer could define one or more annotation tasks, but a task would typically have exactly one designer.

3.3 Truthmarker Approach

3.3.1 Annotation Types

A detailed description of the approach employed in Truthmarker to address user requirements necessitates more precisely defining the types of annotation that can be performed. The two annotation categories that have been defined so far, pixel- and image-level annotations, can be further refined into the specific annotation types supported by Truthmarker. These three types are:

- Categorical
- Free text
- Region of interest (ROI)

The categorical type is an image-level annotation that associates a single annotator response to an image. This response is chosen from a pre-defined list of one or more choices. This annotation type is useful for applying annotations that separate images into binary, nominal, or ordinal categories. The free text type is also an image-level annotation, but it allows users to enter any free text into a field rather than selecting from a pre-defined list of choices. These text fields, though, are labeled to indicate the type of information desired. The final type, ROI annotations, is a pixel-level annotation that can be used to mark-up specific areas of an image. The ROI type is useful for outlining objects or marking specific points of interest within an image.

These three annotation types were designed to address the flexibility requirement of Truthmarker. They were chosen so that most any feature of an image could be annotated, making Truthmarker applicable to a wide range of studies.

3.3.2 Annotation Model

To address the configurability requirement and allow meaningful structure to be applied to annotations, task-specific annotation models can be defined. For a particular

task, the annotation model is the set of annotations, their types, associated terms, and any additional control applied to the way in which images are viewed or annotated.

A study designer defines an annotation model by first deciding on the set of image features that need to be annotated. The designer can then map these features to the appropriate annotation types – categorical, free text, or ROI based annotations. The designer can then associate meaningful terms with each annotation. The designer is completely free in the choice of terms associated with each annotation. This allows them to be completely study specific or taken from some defined ontology. In the case of categorical types, the designer can also define the set of categories that the annotator chooses from for each image.

As an example, imagine a study designer wanted to collect annotations for a set of fundus images that specify the location of the optic disc as well as whether any eye disease was indicated by each image. This designer might first define a ROI annotation associated with the term "Optic Disc" that would be used to outline the optic disc within the image. The designer could also define a categorical annotation associated with the term "Disease" and the choices "Yes" and "No". Annotation data for both of these image features could then be collected for the dataset.

In addition the set of annotations, types, and terms, the annotation model can be used to control the way in which the annotation process is conducted. A few examples include controlling the order in which images are viewed, whether annotators can review their previous annotations, and whether annotators can move past one image onto the next before completing annotation.

3.3.3 Dynamic Interface

Truthmarker is designed to map the annotation model to the user interface. Annotations defined within the model are each represented by a user interface element that allows users to apply that annotation to a particular image. Moreover, the different

annotation types (categorical, free text, and ROI) are differentiated by having unique interface elements. The elements corresponding to each annotation are labeled with the associated terms. Finally, the interface allows only annotations that are defined within the model.

Dynamically generating the user interface in this way has two important benefits for Truthmarker. The first is that it provides guidance to the annotator. If there is only one interface element that allows annotation and it is labeled with a meaningful term, then it is fairly clear to the annotator what is expected of them. This addresses usability requirements and simplifies the job of the annotator. The second benefit is related to data quality. The guidance provided to and restrictions placed upon the annotator can help ensure that the required annotations are collected while erroneous or unwanted information is not.

3.3.4 Data Management

With respect to data management, Truthmarker needs to provide users with straightforward mechanisms to load image data along with the associated annotation model and to retrieve the annotation data. Additionally, since generation of usable annotation data is the goal of Truthmarker, providing annotation output that is easy to analyze is central to the utility of the system. Truthmarker is designed with a set of procedures and standard data formats meant to make data management and analysis simple for the users.

Preparing image data for Truthmarker is done by packaging the image dataset into a single directory. The annotation model is defined using a Truthmarker configuration file. Users then transfer the image directory and configuration file to the tablet. Upon start up, the Truthmarker application will discover these files and create the appropriate interfaces to allow annotation.

When the annotator has completed their task, the annotation data for all of the images is compiled into a single file. The file can then be accessed by the study designer by physically connecting the tablet to a computer and transferring this file. Perhaps more conveniently, the annotation data can also be returned to the study designer via the internet.

To make the generated annotation data more useful, a standard format for storage and transmission of the data has been defined. The format is designed to allow the data to be easily machine- and human-readable. Generally, the format maps annotator decisions to the defined annotation model. That is, for each image, annotation choices made by the annotator are associated with the appropriate annotation types and terms.

3.4 System Requirements

Given the user requirements and the general approach designed to meet them, a set of specific system requirements can be enumerated. Truthmarker has the following system requirements:

1. Truthmarker should be developed for the Apple iPad tablet computer. Extensive developer tools, a distribution network, and user familiarity with the platform make it the best choice for tablet-based software.
2. Truthmarker should allow input of image data via iTunes. The process of transferring data to and from an iPad using iTunes is straightforward and familiar to users.
3. Truthmarker should allow resulting annotation data to be retrieved using iTunes for the reasons listed above or via the internet when the tablet is not physically accessible.
4. Truthmarker should use an XML-based standard for defining annotation models. A structured format like XML is fairly straightforward for study designers and methods for validation XML documents are available.

5. Truthmarker should also use an XML-based standard for storage and transfer of the resulting annotation data. In addition to the benefits mentioned above, it allows the data to be easily machine-parsed and used for analysis.

CHAPTER 4

METHODS

4.1 System Architecture

Truthmarker is a data collection tool. Given this fact, the architecture was designed with the way in which data needs to flow between Truthmarker, the study designer, and the annotator in mind. With respect to the study designer, there is a reciprocal flow of data to and from Truthmarker. Data is input by the designer in the form of image data sets and annotation model specifications. The designer then receives data in the form of annotations for their images of interest. Data flow from the annotator, on the other hand, is unidirectional in the form of the annotations assigned to the images. Figure 4.1 illustrates the data interactions between Truthmarker and the users. The following sections will provide detail regarding how the implementation of Truthmarker streamlines these data interactions to provide a tool for rapid and reliable collection of image annotations.

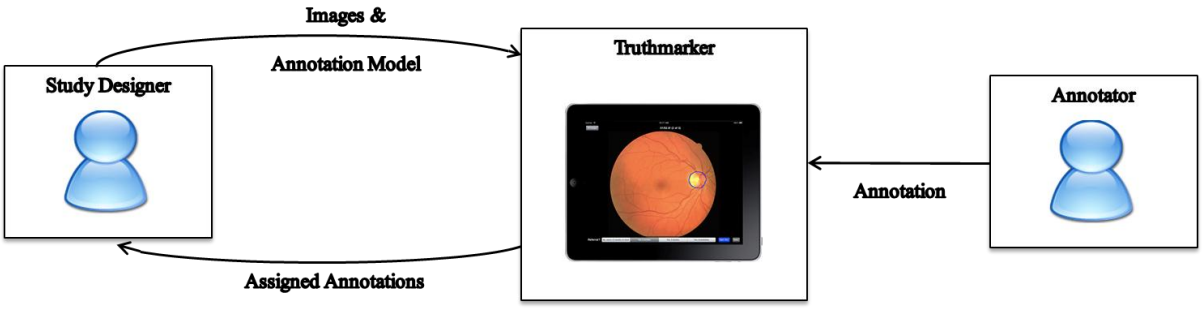

Figure 4.1: This diagram shows the data interactions between the study designer, annotator, and Truthmarker. The study designer inputs data to Truthmarker in the form of images and an annotation model specification. The annotator inputs data by providing annotations for the images. The resulting annotation data is then returned to the designer.

4.2 iPad Platform

Truthmarker was implemented for the Apple iPad tablet computer using the iOS operating system. The choice of the iPad and iOS as the platforms provide several benefits for Truthmarker. First, the iPad is a widely used tablet computer platform, providing a broad user base with access to and familiarity with the platform on which Truthmarker is built. It also provides integrated methods for distribution of Truthmarker. Truthmarker has been made available publicly through the Apple App Store and can be distributed privately to collaborators through the use of Apple enterprise distribution tools.

The iOS platform also provides mature, well-documented developer tools and application programmer interfaces (APIs). The availability of these tools and APIs allowed Truthmarker to be developed quickly without the need to re-implement functionality that existed within iOS. Especially important for Truthmarker was the built-in ability of iOS to handle standard image file formats. This includes functionality for reading the files, displaying the images, and providing standard touch-based zooming and scrolling for images. Additionally, the popularity of iOS has led to the development of useful third-party open source libraries. In particular, third-party libraries were used to handle parsing and writing XML data and for handling of zip archives.[31,32]

4.3 Annotation Project Definition

To perform annotation for a set of images, a study designer would first choose a set of images and define the accompanying annotation model. The annotation model for a project is specified with a single XML configuration file. To make the process of model definition easier, designers are provided with a default XML configuration file that can be used as-is or edited to meet their needs. The following sections will detail how to define the model using a configuration file and then load image data along with the configuration onto an iPad.

4.3.1 Annotations, Types, and Terms

Each annotation to be collected from the images is defined as an element within the XML configuration file. The XML specifies all the characteristics needed to generate user interface elements, create and edit annotations, and save or transmit the resulting annotation data. The three supported annotation types (categorical, free text, and ROI) have a set of shared as well as unique characteristics that can be individually specified. These include the associated term, annotation behavior, and details regarding the user interface elements. Table 4.1 summarizes the set of supported characteristics that can be specified for each annotation.

Within the configuration file, ROI annotations can be assigned a more specific type that indicates how the annotator selects an area within the images. The initial implementation of Truthmarker supports three specific types of ROI annotations: polygon, spline, and point. As the name implies, polygon-based ROI annotations are used for outlining an image region with a polygon shape. The annotator determines the vertices and these are connected with straight lines to form a polygon. Spline-based ROIs are similarly defined, but fit a cubic curve to the vertex points rather than connecting them with straight lines. These are useful for outlining structures with curved rather than straight edges. Finally, point-based annotations are used to annotate a single point within an image.

4.3.2 Process Control

The XML configuration file is also used to specify control applied to the process of annotation. Study designers can use this ability to control annotation in ways that cannot be captured by simply defining the types of annotations allowed for each image. This can allow the designer to guide to the annotator by, for instance, providing an explicit set of instructions that will be displayed during annotation. The designer can also

Table 4.1: The set of characteristics that can be specified for annotations within the configuration file.

XML Tag	**Description**
text	This term is associated with any assigned annotations. The string provided here is incorporated in the user interface and output in the resulting annotation data. A non-empty string must be specified.
required	This boolean value indicates whether the annotator should provide a response for this annotation for all images. If it is omitted, a default value of false is assumed.
option	For categorical annotations, this specifies one of the options that can be assigned to each image. One or more values must be specified for categorical annotations.
textfield	For free text annotations, this specifies the set of fields into which users can enter text. One or more values must be specified for free text annotations.
roitype	For ROI annotations, this specifies the shape used to define the ROI (polygon, spline, or point). A single value must be specified for ROI annotations.
color	This determines the color (using red, green, and blue values between 0 and 255) used to display interface elements associate with the annotation.

Table 4.2: The set of controls that can be applied to the annotation data collection process.

XML Tag	**Description**
email	This is used to specify the email address to which annotation data can be sent once annotation is complete.
annotator	This control is used to associate an annotator name or ID with all assigned annotations. This can be useful for study designers collecting data from multiple annotators.
description	This control can be used to provide a text description of the expected annotations or set of instructions that is displayed to the annotator.
image-order	The order in which images are presented to the annotator is controlled by this option. It can be set to generate a random order, use an order defined by the designer, or use the order in which iOS lists the image files. The last case is the default behavior.
show-preview	This is a boolean value indicating whether or not users should be shown thumbnails for the images within the data set. If not specified, a default value of true indicating that thumbnails should be shown is used.
force-sequential	When true, this boolean value prevents annotators from moving past one image onto the next before completing the required annotations. If not specified, a default value of false is assumed.
allow-review	When false, this boolean value prevents annotators from editing their own annotations for an image once they have completed the required annotations. If not specified, a default value of true is assumed.

exert more direct control over annotation by specifying the order in which images are viewed or by indicating that certain annotations are required for each image. In this case, the configuration could also be set to prevent the annotator from moving past images for which they have not provided required annotations and prevent the annotator from reviewing images that have already been annotated. Table 4.2 describes the currently supported process controls and their corresponding XML tags.

4.4 Performing Annotation

4.4.1 Data and Model Input

Once the image data set and annotation model is defined, the study designer loads this data onto an iPad so that annotation can be performed. This loading process is performed using iTunes file sharing. iTunes file sharing allows users to transfer files onto an iPad into directories that are accessible only by a single application. In this case, the data is transferred to the Truthmarker-specific directory. To load the files onto the iPad, the set of images to be annotated should first be collected into a single directory and the directory then compressed to a zip archive. After connecting the iPad to their computer, users can then drag-and-drop the zip archive and XML configuration file via iTunes file sharing to transfer them to appropriate directory on the iPad. The next time that the Truthmarker application is launched, it will discover the newly loaded data and allow annotation of the images.

Many different data sets and associated configuration files can be loaded onto a single iPad and annotated using Truthmarker at the same time, allowing the study designer and annotator to work on several different annotation projects at once. Distinct configuration files can be associated with each set of images or a single configuration can be used for multiple image sets. This allows one-to-one, many-to-one, or one-to-many relationships between image data sets and annotation models to be enforced. The association of configuration files to image data sets is done using a file naming

convention. Giving a configuration file the same name as a directory of images will cause that configuration to be used for that set of images. Alternatively, giving the configuration file particular pre-determined name, "config.xml", will cause it to be used for any image sets without an associated configuration.

A finer layer of hierarchy can also be applied to image sets annotated using Truthmarker. Within image directories loaded onto the iPad, one layer of sub-directories can be used to group images together. During annotation, images within each sub-directory can be viewed as a group. This allows, for example, medical images collected from the same subject to be viewed and annotated together.

4.4.2 Annotation Interface

When the application is started, the available data is examined and each data set and associated annotation model is mapped to an annotation project. Users are presented with a list of the available projects loaded on the tablet. When a project is selected, a summary of the project is displayed. This summary includes the total number images in the data set, the number that have been annotated, instructions provided in the configuration file, and a set of thumbnails for the images. Annotators can then select a particular image to annotate or choose to continue where they left off with previous annotation of the images. The project selection and summary interface can be seen in Figure 4.2.

Once an image is selected, a new view containing this image presented. The annotator can zoom and scroll within the image using touch controls. The annotator can also move to the next or previous images within the set by using a swiping gesture. In addition to the image, two toolbars are displayed. One shows the name of the current image and the position within the data set. A button allowing the annotator to dismiss the image view and return to the project summary is also provided. The toolbar other contains the interface elements that allow annotations to be assigned to the image. Each

Figure 4.2: This interface is displayed to the user when Truthmarker starts. The list along the left side lists the projects that have been loaded onto the tablet. Each project has a set of images and a configuration file specifying an annotation model. The right side displays a summary of the selected project. Users can begin annotating an image by tapping on the thumbnail or continue where they last left off annotating.

element within this second toolbar is generated using the annotation model defined within the configuration file. Figure 4.3 shows the interface displayed during annotation of an image and the associated XML definition of the annotations.

Interface components built-in to iOS such as radio-style buttons and text fields were used for constructing the interface elements required for categorical and free text annotation. ROI annotation, though, required the design of new interface elements and interactions. In the case of polygon and splines, the annotator defines the ROIs by tapping the image to create individual points used as vertices for the ROI. These vertices

Figure 4.3: The user interface displayed during annotation of an image. Users can use touch controls to zoom in or out, scroll within an image, or move to the next or previous image within the set. The upper toolbar displays some information about the image (filename and index within image set) and provides a button to return to the project summary. The lower toolbar indicates that categorical (A), ROI (B), and free text (C) annotations are being collected. The XML elements defining these annotations are shown as well. The image being viewed is an ophthalmic image known as a visual field.

are connected in order by straight lines or cubic splines fit to the vertices. To complete the ROI, the annotator taps the first vertex again to close the polygon. Point annotations are simpler to create. They only require a single tap to define the point of annotation.

The initial definition of these ROIs, though, is only part of the annotation process. Annotators require the ability to adjust the position and bounds of these ROIs to create accurate annotations. Adjusting the position of an entire ROI can be done by tapping to select the ROI and dragging it to a new position. Adjusting single vertex points within a ROI requires finer control, however, and a new interface element termed the "squid" was designed to provide this fine control. The squid consists of movement handles offset from the position of the vertex. These handles allow users to drag and move vertex points to adjust the bounds of the ROI. The offset prevents the location of the point from being blocked by the annotator's finger. Additionally, the squid contains buttons that allow adding a new or deleting the current point. Removing an entire ROI annotation can be done by deleting each of the points used to define it. Figure 4.4 shows the squid element during use to adjust the bounds of a spline-based ROI annotation.

4.5 Annotation Retrieval and Format

4.5.1 Annotation Data Format

Once the annotation process is complete, the annotation data needs to be returned to the study designer for use in research. For the ease of researchers, an XML-based standard for annotation data was defined for use by Truthmarker so that this data can be easily parsed and analyzed. For the purposes of data transfer and analysis, the annotations for a given project are stored within a single XML document. At the highest level, this document is organized by image. Within each image element are child elements that store the annotations assigned to that image. For all annotation types, these elements contain the terms and display color defined within the configuration file. All annotations also contain time stamps indicating the length of time that the image was

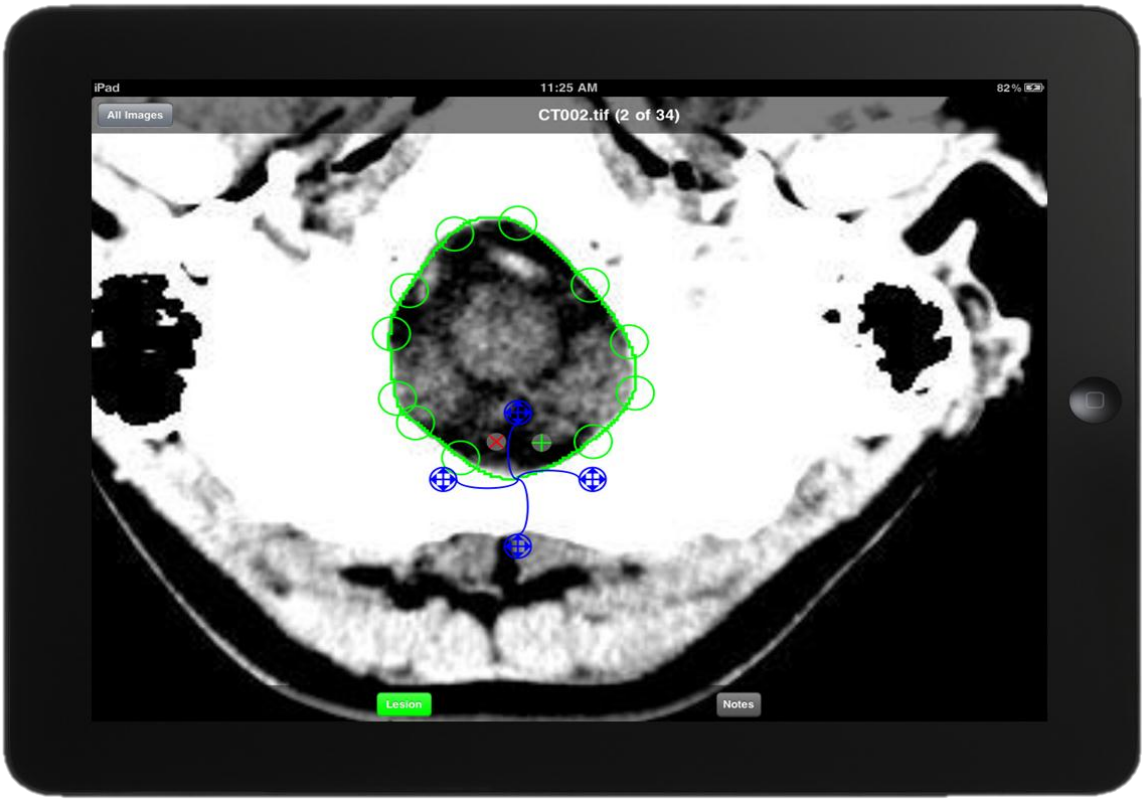

Figure 4.4: A spline-based ROI annotation outlining an area of a CT image is being edited using the squid. The area green line marks the area enclosed by the ROI and each green circle marks a vertex point selected by the user. The squid, seen in blue, is being used to adjust the location of a single point. The four movement handles are indicated by crossing arrows. The red x and green plus sign can be used to delete or add new vertex points, respectively.

viewed before the annotation was assigned. For categorical annotations, the selected option is stored within the annotation element. The text entered by users is stored with elements mapping to free text annotations. Finally, the three ROI annotation types are stored as sets of Cartesian x,y-coordinates that define the ROI bounds (polygon and spline annotations) or point (point annotations).

4.5.2 Data Retrieval

The XML file containing the annotation results can be retrieved either via the iTunes file sharing interface or through email. Data retrieval using iTunes is similar to

the data loading process. Users simply connect to iPad and, using iTunes, drag and drop the file containing the annotations to their computer. Alternatively, if physical access to the tablet is inconvenient for the designer, the annotation file can be sent as an email attachment to an address specified within the configuration file. This, of course, requires the tablet to have an internet connection, but to also have an email account associated with the standard Apple email application installed on the iPad.

4.6 Extensibility

While it is obviously useful to provide a tool for addressing a defined set of requirements, it is even more useful to provide a tool that allows extension of its functionality in case the requirements change. Truthmarker was designed with extensibility in mind so that functionality could be added with as little effort as possible. Specifically, two primary pieces were designed to be modular to allow extension of the basic functionality. The first modular piece determines to the types of annotation projects that can be performed by handling the data types and sources that can be read and annotated. The second relates to annotation functionality and controls the types ROI annotations that can be collected. In each case, these modules were defined by abstract classes and Truthmarker was implemented by inheriting from these abstract classes to provide the current functionality. New functionality can be added by implementing new modules that inherit from the existing classes to extend or override their default behavior.

4.6.1 Projects

Most aspects image as well as annotation data input and output are handled inheritors of a single abstract super class. This abstract class defines functions for reading images and associated annotations from a data source, writing generated annotations to a data sink, and handling some aspects of the user interfaces that display this data but does not provide implementations for these functions. In the current version of Truthmarker, this class is extended to handle data input and output using the iPad file

system. That is, it reads and displays images that have been preloaded onto the tablet and outputs annotation data files to the tablet for retrieval via iTunes or email.

A new module implemented for Truthmarker would not necessarily have to require preloading data onto the iPad. For example, images could be downloaded from some data source via the internet. Similarly, annotation data could read from and sent to some central repository. In addition to handling new data sources and sinks, new data types could be handled. Implementing new methods for reading and displaying data has allowed viewing and annotation movie files.

4.6.2 Regions of Interest

Each of the ROI annotation types supported by Truthmarker (polygon, spline, mark) are implemented by extending a super class representing ROIs. Inheritors of this class have to implement functions to display the ROI on an image, handle user interaction with the ROI, and generate and parse XML data used to store ROIs. Defining new inheriting classes would allow new types of ROI annotations to be collected for images. These could include ROIs defined by bounding shapes such as ellipses or rectangles that are not currently supported or by allowing users define a ROI by free-hand drawing.

4.7 Evaluation

4.7.1 Annotation Protocol

Evaluation of the system was performed by comparing categorical annotations created by retinal experts using Truthmarker to annotations created by those same experts using a desktop computer. To perform the annotation, a retinal expert was asked to review a set of fundus images and assign a grade to each. The grade assigned to each image specified if and when the patient should be referred from primary care to an ophthalmologist for specialist care based on the severity of DR indicated by the image.

The experts used a categorical annotation to choose one of the following grades for each image:

- 0 – No referral, return in 12 months or more
- 1 – Yes referral, 6 months
- 2 – Yes referral, 3 months
- 3 – Yes referral, immediately

Using the annotation, a numeric grade between 0 (DR needing no referral) and 3 (DR needing immediate referral) was assigned to each image with higher grades indicating more severe DR. Figure 4.5 shows the user interfaces displayed to the experts during DR annotation on the tablet and desktop computers. Specifically, the experts were provided with the following instructions to perform the annotation:

> "Imagine that you have only this image for this patient, that the image of the other eye has the same level of diabetic retinopathy, that the patient is in primary care, and that this is all you know about the patient. Based on the retinopathy, select whether the patient with this fundus image should return for imaging in 12 months, or be referred to an ophthalmologist/retinal specialist in 6 months, 3 months or immediately. The 12 month selection is for those with no/minimal retinopathy or anyone who doesn't require a referral to a specialist for at least one year. The 'immediately' selection means that the patient should be referred to a specialist for evaluation and possible treatment immediately. The 3 or 6 month selections may be chosen if you feel the patient doesn't need to be seen immediately but should be followed up in less than one year by an ophthalmologist/retinal specialist, and that it is not safe to simply image the patient again 12 months from now. You should just use your best judgment since there will often be no absolutely right or wrong answer."[33]

During annotation with Truthmarker, the experts used a first generation iPad with a resolution of 1024 by 768 and iOS version 3. Desktop annotation was performed using a high-definition LCD display with a resolution 1920 by 1200. Standard brightness and saturation settings were used. Two experts each performed the annotation task three times: twice using a desktop and once using a tablet. In all cases, the order in which images were viewed was randomized and the experts were masked to any previous

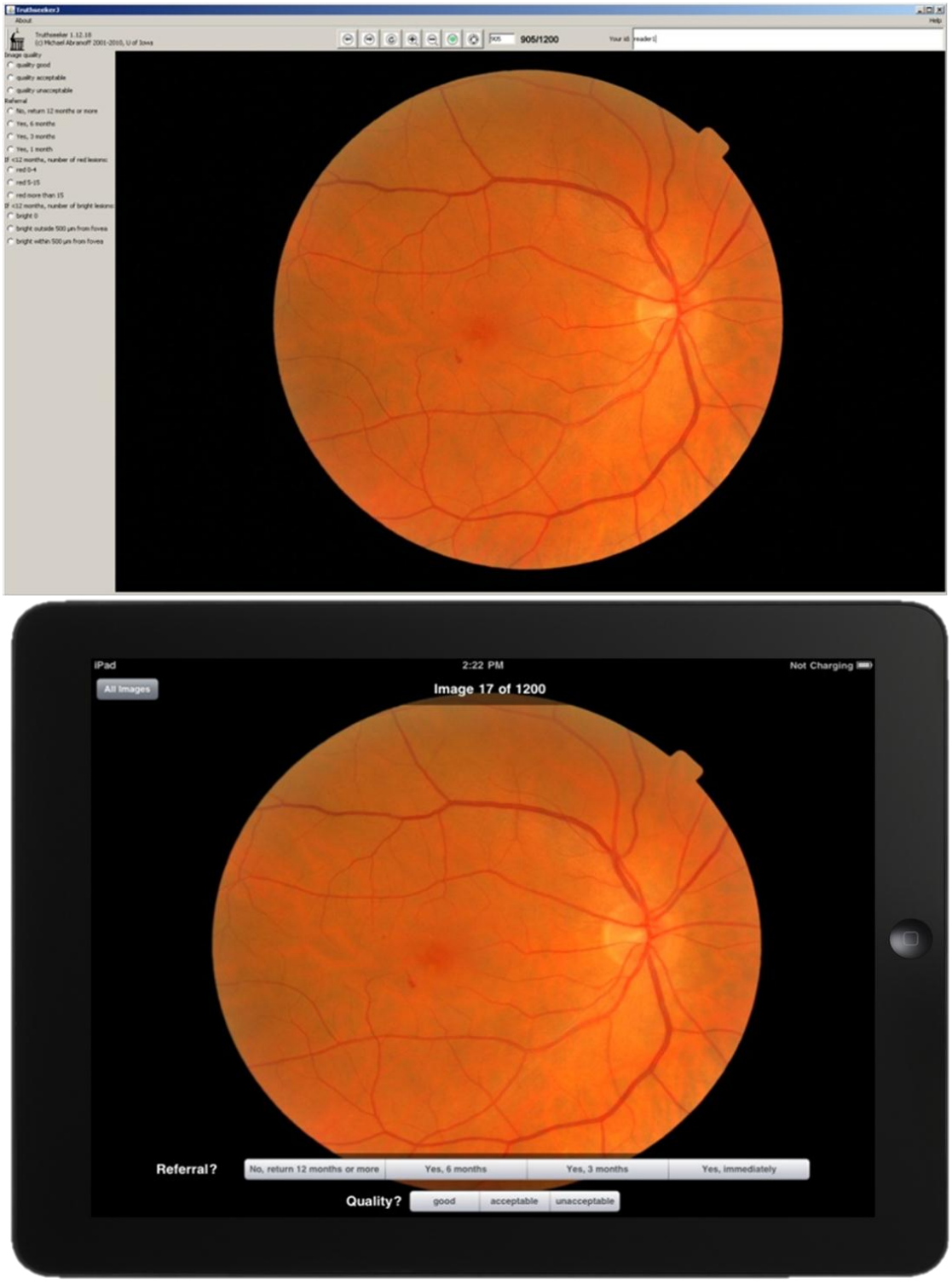

Figure 4.5: The user interfaces for annotation of DR severity using a desktop (top) and tablet (bottom) computer.

annotations. Additionally, one expert performed annotation on the tablet first while the other performed annotation on the desktop first in order to minimize any bias.[33]

4.7.2 Messidor Dataset

A set of high quality fundus images, publicly available as the Messidor dataset[10], was used for the annotation of DR. This dataset consists of 1200 images gathered from diabetic patients at tertiary care clinics in France. The images were captured using a Topcon TRC NW6 non-mydriatic fundus camera with a 45 degree field of view. The images had resolutions ranging from 1440 by 960 to 2304 by 1526 and were in TIFF format with 8 bits per color channel.

4.7.3 Data Analysis and Outcome Measures

The primary statistic used to measure agreement between different sets of DR grade annotations was the κ statistic. The κ statistic is a metric with a typical range of 0 to 1 that is used to measure the agreement between two grading systems assigning binary (unwieghted κ) or ordinal classifications (weighted κ) to a set of observations. A κ statistic with a value of 0 indicates that the graders agreed no more often than expected by chance, while a value of 1 indicates that the graders had perfect agreement.[34,35] Two sets of κ statistics were calculated to characterize intra-observer agreement. First, κ statistics between DR grades assigned by an expert using a tablet to the grades assigned by the same expert using a desktop were calculated to measure cross-platform, intra-observer agreement. Second, intra-observer agreement was also measured by calculating κ statistics comparing the two sets of desktop-based grades for each expert. Inter-observer agreement was measured using κ statistics comparing the grades annotated by the first expert to the grades annotated by the second expert on the same platform.

In all cases, both weighted and unweighted κ statistics were calculated. The weighted statistic was used to compare observers based on the full range of ordinal grades and the unweighted statistic was used to compare observers based on a binary

classification. For the unweighted statistic, a binary classification was applied to the DR grades by setting the grade of "No referral, return in 12 months or more" to 0, while all other grades were set to 1.

In addition to the κ statistics, cross-platform agreement was characterized using a Bhapkar test for marginal homogeneity. The test is performed by calculating a chi-squared value comparing the row and column marginal values in a contingency table. A statistically significant result indicates that the row and column marginal values are different.[36] Specifically, this test compared the distribution of images where the expert assigned a higher DR grade while using a tablet than when using a desktop to the distribution where the desktop-based DR grade was higher. If a statistically significant difference was found between these distributions, it would provide evidence that there existed a systematic bias in the tablet-based annotation with respect to the desktop. A Bhapkar test was applied to both the binary DR classification and the full range of DR grades.

Finally, quality of the tablet-based annotations was assessed by determining the accuracy of DR grades assigned using a tablet with respect to DR grades assigned using a desktop. Accuracy was measured by calculating the sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) each DR grade. These metrics were determined the accuracy for each of three classes of grades: DR grade 1 or higher, DR grade 2 or higher, and DR grade 3.[37] For this analysis, a reference standard was created by averaging the grades assigned by the experts using a desktop.[33]

CHAPTER 5

RESULTS

5.1 Expert Grading Annotations

Contingency tables that characterize the distribution of DR grades assigned by each grader are included as Tables 5.1 and 5.2. These tables compare the distribution of grades assigned by an expert using a Truthmarker on an iPad tablet to the grades assigned by that same expert using a standard desktop computer. The main diagonals of these tables indicate the counts of images where the tablet- and desktop-based grades are in agreement. The values above the main diagonal are image counts where the tablet-based grade was higher than the desktop-based grade and values below the diagonal indicate counts the desktop-based grades were higher. Table 5.1 summarizes the results for the binary DR grades and 5.2 uses the full range of DR grades.

Table 5.1: Comparison of the binary DR grades annotated using Truthmarker to those annotated using a desktop for each expert.

Grader 1	**Tablet DR Grade**		
Desktop DR Grade	**Grade 0**	**Grade 1 or Higher**	**Total**
Grade 0	915 (76.51)	25 (2.09)	940 (78.60)
Grade 1 or Higher	60 (5.02)	196 (16.39)	256 (21.40)
Total	975 (81.52)	221 (18.48)	1196* (100.00)
Grader 2	**Tablet DR Grade**		
Desktop DR Grade	**Grade 0**	**Grade 1 or Higher**	**Total**
Grade 0	727 (60.58)	60 (5.00)	787 (65.58)
Grade 1 or Higher	43 (3.58)	370 (30.83)	413 (34.42)
Total	770 (64.17)	430 (35.83)	1200 (100.00)

Note: Values indicate numeric frequency and (%).

* : Grader 1 missed 4 images so analyses were performed using 1196 images

Table 5.2: Comparison of the full range of DR grades annotated using Truthmarker to those annotated using a desktop for each expert.

Grader 1	**Tablet DR Grade**				
Desktop DR Grade	**0**	**1**	**2**	**3**	**Total**
0	915 (76.51)	20 (1.67)	3 (0.25)	2 (0.17)	940 (78.60)
1	29 (2.42)	23 (1.92)	0 (0.00)	15 (1.25)	67 (5.60)
2	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
3	31 (2.59)	17 (1.42)	1 (0.08)	140 (11.71)	189 (15.80)
Total	975 (81.52)	60 (5.02)	4 (0.33)	157 (13.13)	1196* (100.00)
Grader 2	**Tablet DR Grade**				
Desktop DR Grade	**0**	**1**	**2**	**3**	**Total**
0	727 (60.58)	55 (4.58)	4 (0.33)	1 (0.08)	787 (65.58)
1	42 (3.50)	196 (16.33)	11 (0.92)	17 (1.42)	266 (22.17)
2	1 (0.08)	10 (0.83)	7 (0.58)	9 (0.75)	27 (2.25)
3	0 (0.00)	15 (1.25)	7 (0.58)	98 (8.17)	120 (10.00)
Total	770 (64.17)	276 (23.00)	29 (2.42)	125 (10.42)	1200 (100.00)

Note: Values indicate numeric frequency and (%).
* : Grader 1 missed 4 images so analyses were performed using 1196 images

The data in these tables was used to perform a Bhapkar test of marginal homogeneity. In both cases (the binary and full range of grades), the resulting p-value was greater than 0.5. This indicates that there is no statistically significant difference between the distributions of marginal values in the comparison of tablet- and desktop-based grades.

5.2 Kappa Statistics

The κ statistics measuring intra-observer agreement are shown in Table 5.3. In this table, rows labeled "Tablet to Desktop" indicate κ statistics comparing grades assigned using a tablet to those assigned using a desktop by the same expert. The resulting κ statistics were 0.778 and 0.812 for the two experts. The weighted κ statistics for this comparison were slightly lower at 0.776 and 0.795.

For comparison, rows labeled "Desktop to Desktop" indicate intra-observer κ statistics comparing the two sets of desktop-based grades for each expert. The resulting κ statistics were 0.800 and 0.784. The weighted κ statistics were again slightly lower at 0.796 and 0.768.

Table 5.3: The intra-observer agreement as measured by κ statistics for cross-platform (tablet to desktop) and single-platform (desktop to desktop) DR annotation.

Grader		***κ* (95% CI)**	**weighted *κ* (95% CI)**
1	Tablet to Desktop	0.778 (0.733-0.823)	0.776 (0.734-0.818)
	Desktop to Desktop	0.800 (0.758-0.842)	0.796 (0.758-0.834)
2	Tablet to Desktop	0.812 (0.777-0.846)	0.795 (0.765-0.826)
	Desktop to Desktop	0.784 (0.746-0.821)	0.768 (0.736-0.800)

Inter-observer κ statistics comparing the experts' annotated DR grades for each platform are shown in Table 5.4. The κ statistic measuring the agreement of the two experts using a tablet was 0.544 and the weighted κ was 0.648. For the desktop-based grades, the κ statistic was 0.625 and the weighted κ was 0.675.

Table 5.4: The inter-observer agreement of DR annotation for each of the platforms.

Platform	***κ* (95% CI)**	**Weighted *κ* (95% CI)**
Tablet	0.544 (0.496-0.593)	0.648 (0.605-0.691)
Desktop	0.625 (0.578-0.672)	0.675 (0.637-0.714)

5.3 Receiver Operating Characteristics

Finally, the quality of tablet annotations assigned using Truthmarker was characterized by calculating the grading accuracy using desktop-based grades as a reference standard. The full set of results for the accuracy analysis is summarized in Table 5.5. Worth noting is that the experts achieved AUC values of 0.950 or better in grading accuracy for DR annotated with a grade of 1 or higher using a desktop, an AUC of 0.942 or better for DR of grade 2 or higher, and an AUC of 0.801 or better for DR of grade 3.[33]

Table 5.5: The accuracy of the DR grades assigned by each expert using Truthmarker assessed by sensitivity, specificity, and area under the ROC curve (AUC).

	DR Grade 1 or Higher		
	Sensitivity (95% CI)	**Specificity (95% CI)**	**AUC (95% CI)**
Grader 1	0.872 (0.853-0.891)	0.987 (0.980- 0.994)	0.950 (0.938-0.962)
Grader 2	0.848 (0.828-0.869)	1.0 (0.998-1.0)	0.967 (0.957-0.977)
	DR Grade 2 or Higher		
	Sensitivity (95% CI)	**Specificity (95% CI)**	**AUC (95% CI)**
Grader 1	0.966 (0.956-0.976)	0.911 (0.895-0.927)	0.942 (0.928-0.955)
Grader 2	0.948 (0.936-0.961)	0.976 (0.967-0.985)	0.974 (0.965-0.983)
	DR Grade 3		
	Sensitivity (95% CI)	**Specificity (95% CI)**	**AUC (95% CI)**
Grader 1	0.991 (0.9985-0.997)	0.611 (0.585-0.640)	0.801 (0.779-0.824)
Grader 2	0.987 (0.980-0.994)	0.728 (0.704-0.754)	0.859 (0.840-0.879)

CHAPTER 6

DISCUSSION

6.1 Addressing the Requirements

The overarching goal guiding the development of Truthmarker was to provide a flexible tool that allows researchers to quickly and reliably gather image annotation data. Addressing this goal, of course, required consideration of the use cases and user requirements. The two types of users considered, study designers and annotators, have both shared and distinct requirements for Truthmarker.

Both of these user types, for instance, require that annotation with Truthmarker is straightforward and can be completed quickly. This was achieved by implementing a tablet-based system with user interface elements tuned for the annotation task. For the annotator, the convenience of a tablet and guidance provided by the interface makes annotation both easier and quicker. The study designer benefits as well because easier annotation makes recruitment of experts less difficult and results in more reliable annotations.

In other cases, though, designing Truthmarker necessitated finding a balance between competing requirements. Truthmarker was required to be applicable to as many different image data sets and annotation types as possible while also allowing domain-specific knowledge to be applied to the collection of annotation data. In case, a balance was struck by allowing study designers to define their own annotation models tailor-fit to specific annotation task. Unfortunately, defining an annotation model does present a problem for the designers. Designers are required to edit the XML by hand. This means that designers are required to be familiar with XML in general and the standard used by Truthmarker.

6.2 Evaluation

Considering the study performed to validate annotations created using Truthmarker, the results support the idea that these annotations are equivalent to those created using a standard desktop computer at least with respect to the data set considered. The κ statistics measuring annotations of DR grade assigned by a single annotator on different platforms ($\kappa = 0.778$ and 0.812, weighted $\kappa = 0.776$ and 0.795) were not significantly different than κ statistics measuring annotations assigned by a single grader using a desktop ($\kappa = 0.800$ and 0.784, weighted $\kappa = 0.796$ and 0.768). Furthermore, these values were significantly higher than the inter-observer κ statistics ($\kappa = 0.625$, weighted $\kappa = 0.675$). This indicates that any differences in annotation introduced by the use of Truthmarker on a tablet have a smaller effect than the existing differences between annotators.

The results of the Bhapkar test of marginal homogeneity and the accuracy assessment also suggest that the annotations DR grade were equivalent across the platforms. The Bhapkar test showed no statistically significant difference between the platforms indicating that there was no obvious bias in the ratings assigned using a tablet. The sensitivity, specificity, and AUC assessment of tablet-based DR annotations with respect to desktop-based annotations were similar to previous results of DR grading accuracy of ophthalmologists.[38,39]

There are limitations, though, to the evaluation of tablet-based annotation presented here. The power of the evaluation work is limited by the fact that only two annotators were studied. In addition to the small sample size, this work only considered annotations for a single disease and image modality. Studies involving larger numbers of annotators and more diverse data sets are needed to confirm these results. This work serves only as a pilot study comparing tablet-based to standard desktop-based annotations, albeit one with encouraging results.

6.3 Future Work

Future work on Truthmarker will focus on adding functionality for handling new types of data and annotations as well as for defining annotation models. With regard to the latter, providing an interface within Truthmarker used to define or edit annotation models would make it a more attractive tool for study designers. The XML configuration files would be generated based on interactions with the user interface rather than through hand editing of XML. Handling new types of data and annotations would also, of course, make Truthmarker a more attractive annotation tool. For instance, extensions for three-dimensional images like those produced by some medical imaging modalities would make Truthmarker useful for new areas research. Additionally, news ways to interact with data would improve Truthmarker. Integration with image databases could make data management easier for researchers. Truthmarker could automatically pull in images from the database and return the assigned annotations.

CHAPTER 8

CONCLUSIONS

Truthmarker addresses several of the challenges associated with gathering large amounts of image annotation data. It is a tablet-based system that can be generally applied in image analysis research, while still allowing the data collection process to be customized to address the needs of a particular study. The quality of annotations generated using Truthmarker was evaluated using expert annotations of DR severity for fundus images. The results indicate that for this medically important task, Truthmarker was equivalent to annotation performed using standard desktop tools. Future work on this system includes streamlining the process of configuring Truthmarker for a particular annotations task, providing support for more diverse sets of image data to be annotated, and supporting the collection of additional annotation types. Truthmarker can be a valuable tool for image analysis researchers. It provides a platform for rapid and efficient collection of the annotations that are required for their research.

REFERENCES

1 Varma, M. & Ray, D. in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on Computer Vision. 1-8.

2 Gehler, P. & Nowozin, S. in Computer Vision, 2009 IEEE 12th International Conference on Computer Vision. 221-228.

3 Ion, A., Carreira, J. & Sminchisescu, C. in International Conference on Computer Vision (Barcelona, Spain, 2011).

4 Sharon, E., Galun, M., Sharon, D., Basri, R. & Brandt, A. Hierarchy and adaptivity in segmenting visual scenes. Nature **442**, 810-813 (2006).

5 Bishop, C. M. Pattern Recognition and Machine Learning. 3-4 (Springer, 2006).

6 Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision **88**, 303-338 (2010).

7 Fei-Fei, L., Fergus, R. & Perona, P. in Computer Vision and Pattern Recognition (2004).

8 Griffin, G., Holub, G. & Perona, P. The Caltech-256. Caltech Technical Report (2007).

9 Abramoff, M. D. & Suttorp-Schulten, M. S. Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. Telemed.J.E.Health **11**, 668-674 (2005).

10 MESSIDOR: Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology. TECHNO-VISION Project (2005). <http://messidor.crihan.fr/download-en.php>.

11 Facts About Diabetic Retinopathy (2009). <http://www.nei.nih.gov/health/diabetic/retinopathy.asp>.

12 Abramoff, M. D. et al. Automated early detection of diabetic retinopathy. Ophthalmology **117**, 1147-1154 (2010).

13 Abramoff, M. D. et al. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. Diabetes Care **31**, 193-198 (2008).

14 Mildenberger, P., Eichelberg, M. & Martin, E. Introduction to the DICOM Standard. Eur Radiol **12**, 920-927 (2002).

15 Quinn, J. An HL7 (Heath Level Seven Overview). J AHIMA **70**, 32-34 (1999).

16 Hunter, P., Robbins, P. & Noble, D. The IUPS human Physiome Project. Pflugers Archiv : European journal of physiology **445**, 1-9 (2002).

17 Rubin, D. L., Mongkolwat, P., Kleper, V., Supekar, K. & Channin, D. S. Annotation and Image Markup: Accessing and Interoperating with the Semantic Content in Medical Imaging. Intelligent Systems, IEEE **24**, 57-65 (2009).

18 Channin, D. S., Mongkolwat, P., Kleper, V., Sepukar, K. & Rubin, D. L. The caBIG annotation and image Markup project. Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology **23**, 217-225 (2010).

19 Russell, B. C. & Torralba, A. LabelMe: A database and Web-Based Tool for Image Annotation. International Journal of Computer Vision **77**, 157-173 (2008).

20 Samsung GALAXY Tab Opens a New Chapter in Mobile Industry, <http://galaxytab.samsungmobile.com/press/pressrelease.html> (2010).

21 Motorola Xoom, <http://www.motorola.com/Consumers/US-EN/Consumer-Product-and-Services/Tablets/ci.MOTOROLA-XOOM-US-EN.overview> (2011).

22 HP Slate 500 Tablet PC - Overview and Features, <http://h10010.www1.hp.com/wwpc/us/en/sm/WF05a/321957-321957-64295-3841267-3955550-4332585.html?jumpid=re_r602_slate_body_psg_oct10_product> (2011).

23 Margalit, R. S., Roter, D., Dunevant, M. A., Larson, S. & Reis, S. Electronic medical record use and physician–patient communication: An observational study of Israeli primary care encounters. Patient Educ Couns **61**, 134-141 (2006).

24 ClearPractice Announces General Release of Nimble™ – A Comprehensive EMR for the iPad™, <http://www.clearpractice.com/ehr/press-9-28-2010.cfm> (2010).

25 ifa presents the iPad version, <http://www.ifa4emr.com/index.php?view=article&catid=9:home&id=75:home&format=pdf> (2010).

26 Ziefle, M. Effects of display resolution on visual performance. Human factors **40**, 554-568 (1998).

27 McLean, M. V. Brightness contrast, color contrast, and legibility. Human factors **7**, 521-526 (1965).

28 Oetjen, S. & Ziefle, M. A visual ergonomic evaluation of different screen types and screen technologies with respect to discrimination performance. Applied ergonomics **40**, 69-81 (2009).

29 Lin, P. H. & Kuo, W. H. Image quality of a mobile display under different illuminations. Perceptual and motor skills **113**, 215-228 (2011).

30 FDA. Draft Guidance for Industry and Food and Drug Administration Staff - Mobile Medical Applications, <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm263280.htm> (2011).

31 Moskowski, K. ZipKit: An Objective-C Zip framework for Mac OS X and iOS, <https://bitbucket.org/kolpanic/zipkit/wiki/Home> (2011).

32 Google Data APIs Objective-C Client Library, <http://code.google.com/p/gdata-objectivec-client/> (2011).

33 Christopher, M. et al. Validation of bablet-based evaluation of color fundus images. In preparation (2011).

34 Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull **70**, 213-220 (1968).

35 Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. Biometrics **33**, 159-174 (1977).

36 Bhapkar, V. P. A Note on Equivalence of 2 Test Criteria for Hypotheses in Categorical Data. J Am Stat Assoc **61**, 228-& (1966).

37 Chiang, M. F. et al. Remote image based retinopathy of prematurity diagnosis: a receiver operating characteristic analysis of accuracy. Br J Ophthalmol **90**, 1292-1296 (2006).

38 Suansilpong, A. & Rawdaree, P. Accuracy of single-field nonmydriatic digital fundus image in screening for diabetic retinopathy. J Med Assoc Thai **91**, 1397-1403 (2008).

39 Scanlon, P. et al. The effectiveness of screening for diabetic retinopathy by digital imaging photography and technician ophthalmoscopy. Diabet Med **20**, 467-474 (2003).