WHITE PAPER

The Annotation Exchange Format for Images (AXFI)

Overview

Image annotation and segmentation is an important analytic process in many scientific, technical, and commercial fields. Nonetheless, there are few standard formats for describing and representing image annotations, and those which do exist tend to be used in specific, relatively narrow contexts.[1] This is not a new observation; Daniel L. Rubin *et. al.*, in 2007, note that:

> Images contain implicit knowledge about anatomy and abnormal structure that is deduced by the viewer of the pixel data, but this knowledge is generally not recorded in a structured manner nor directly linked to the image. [Moreover,] the *terminology* and *syntax* for describing images and what they contain varies, with no widely-adopted standards, resulting in limited interoperability. The contents of medical images are most frequently described and stored in free-text in an unstructured manner, limiting the ability of computers to analyze and access this information. There are no standard terminologies specifically for describing medical image contents — the imaging observations, the anatomy, and the pathology. [N]o comprehensive standard appropriate to medical imaging has yet been developed. A final challenge for medical imaging is that the particular information one wants to describe and annotate in medical images depends on the *context* — different types of images can be obtained for different purposes, and the types of annotations that should be created (the "annotation requirements" for images) depends on that context. For example, in images of the abdomen of a cancer patient (the context is "cancer" and "abdominal region"), we would want annotations to describe the liver (an organ in the abdominal region), and if there is a cancer in the liver, then there should be a description of the margins of the cancer (the appearance of the cancer on the image). (http://cedarweb.vsp.ucar.edu/wiki/images/d/d9/R_19.pdf, pp. 1-2).

These challenges inspired **AIM** (the "Annotation and Image Markup" project), which "provides a solution to the ... imaging challenges [of]: No agreed upon syntax for annotation and markup; No agreed upon semantics to describe annotations; No standard format ... for annotations and markup" (https://wiki.nci.nih.gov/display/AIM/Annotation+and+Image+Markup+-+AIM). However, **AIM** itself has not been widely adopted outside the specific field of cancer research and cancer-oriented image repositories.

One obstacle to formalizing image-annotation data is that annotations have a kind of intermediate status, neither intrinsic parts of an image nor merely visual cues supporting the presentation of the image witin image-viewing software. Many applications exist which allow markup or comments to be introduced with respect to an image. From the application's point of view, these annotations are part of the application display, not part of the image — analogous to editing comments that might be added to a text document by a word processor or **PDF** viewer, which are records of user actions, not intrinsic to the document itself. Indeed, one mechanism for recording image annotations in **DICOM** (the "Digital Imaging and Communications in Medicine" format) is via "presentation state." The presentation state includes all details about how the image currently appears to **DICOM** workstation users, such as Radiologists, which could include markings they have made to indicate diagnostically

[1] Current formats include AIM (Annotation and Image Markup), CVAT XML (CVAT is the Computer Vision Annotation Tool), DICOM-SR (Digital Imaging and Communications in Medicine Structured Reporting), PASCAL VOC XML (Pattern Analysis, Statistical modelling and ComputAtional Learning Visual Object Classes), and COCO JSON (Common Objects in Context).

significant image regions or features. Insofar as image-annotations are considered to be artifacts of image-viewing software, rather than significant data structures in their own right, there is less motivation for imaging applications to support canonical annotation standards.

Nevertheless, in many scientific and technical areas image annotations *are* significant; they are intrinsic to the scientific value of a given image as an object of research or observation. Image regions, segments, and features have a semantic meaning outside the contexts of the applications that are used to view the corresponding images, which is why it is important to develop cross-application standards for describing and affixing data to image annotations.

It is also important for image-annotation models to be broadly applicable and multi-disciplinary. While image analysis serves different goals in different contexts (segmentation of microscope images to detect cancer cells serves different ends than segmentation of camera snapshots to study traffic patterns), there is always a possibility of analytic techniques developed in one subject area to be applicable for other image-processing problems, even if the practical outcomes desired of the analyses are very different. Furthermore, certain computational domains are similar enough to image analysis to warrant inclusion in a general-purpose image-annotation framework, even if the underlying data is not "images" in the conventional sense (not, for instance, captured via photographs or microscopy). For example, **PDF** document views, Flow Cytometry data plots, and geospatial maps subject to Geographic Information Systems (**GIS**) annotations may all be considered images — by virtue of a semantic significance attributed to color and to geometric primitives as a way of characterizing phenomena observed or modeled through their data — even though such resources are not acquired by ordinary "image-producing" devices. The "Pantheon" project, characterized as a "platform dedicated to knowledge engineering for the development of image processing applications,"[2] offers one of the few attempts in bioimaging literature to rigorously define "imaging" and "image processing" in the first place. The problem of defining "images" as such, and therefore delineating the scope of image annotation, is addressed below (section ...). In brief, **AXFI** considers the realm of imaging to be more general than just graphics obtained by a direct recording of the optics of some physical scene via cameras, microscopes, or telescopes. That is to say, the image acquisition process is not necessarily one where data is generated by an instrument which produces a digital artifact by absorbing light, so that geometric and chromatic properties of the image are wholly due to the functioning of the acquisition device.

Systematically identifying the scope of "image annotion" is important for **AXFI** because doing so clarifies the sorts of domains whose semantics could reasonably be incorporated into **AXFI**, as well as the sorts of applications which would be reasonable candidates for supporting **AXFI** annotations (i.e., the capability to parse **AXFI** data and represent it vis-à-vis the relevant images). For example, if immunoflourescent Flow Cytometry (**FCM**) data plots are classified as images, then the numerical properties of the "channel" axis, with notions of "decades" and a "log/linear" distinction, become relevant to the **AXFI** vocabulary for representing spatial dimensions and magnitudes. In general, **AXFI** uses paradigms and terminology from "Conceptual Space Theory" as part of the process of formalizing geometric and dimensional notions.[3]

When defining the scope of **AXFI**, it is also important to distinguish the *data models* encapsulated by **AXFI** resources from the file formats where **AXFI** data is encoded. The choice of one file type or another to represent a data structure — **XML** versus **JSON**, for instance — does not fundamentally affect the data thereby communicated. Therefore, it is important to formalize data models in such a way that numerous different serialization languages might actually be used to share/express the data. However, in practice, format-specific standards, such as **XML** Schema Definitions, are often used as the basis for formalizing and enforcing compliance with data models. Therefore, **AXFI** cannot be complete unconcerned with the structure and requirements of files which convey **AXFI** data. This

[2] See https://hal.archives-ouvertes.fr/hal-00260065/document.
[3] See http://idwebhost-202-147.ethz.ch/Publications/RefConferences/ICSC_2009_AdamsRaubal_Camera-FINAL.pdf or https://arxiv.org/pdf/1801.03929.pdf.

is particularly true because AXFI seeks to incorporate the data models of other specifications, such as AIM and GATING-ML, which are formalized via XML specifications. Although AXFI is not primarily XML-based, in short, it intends to be (in the relevant contexts) compatible with XML languages that rely on XML schematization. Further details on how AXFI manages the relation to XML and other serialization languages are provided below (section).

A further detail that should be clarified prior to expositing a formal outline of AXFI is that of how image-annotations originate. Sometimes, of course, annotations are manually introduced on images by human users of image-viewing software. On the other hand, automated image segmentation — or similar algorithmic or AI-driven image processing without human intervention — yields partitions of images into regions, or identification of semantically important locations in an image, therefore generating annotations computationally. In short, AXFI should support both human-generated and computer-generated annotations. This becomes complicated, however, insofar as image-processing may yield analyses which overlap with annotation objectives but may not intrinsically produce annotations in the conventional sense. For example, an algorithm to count the number of cars in a highway picture may rely on statistical analysis of some quantitative image feature — such as "zero-crossings" — without in fact producing determinate image segments.

It is important to remember that image processing operations do not always act directly on images themselves; sometimes algorithms are based instead on mathematical complexes derived from the image, but with their own quantitative properties. For instance, color-valued pixels may be replaced by matrices measuring the derivative of some image-feature field in eight directions around each point (an example would be "Sobel kernels" applied to the image intensity function). For a given "semantic" task — that is, an image-processing objective whose end-result is not just image-related data but some empirical observation — image segmentation, or other analyses yielding annotations, are a means to an end: one *way* to count cars is to delineate the edges of distinct cars in distinct segments, and then count the number of segments which result. However, statistical image-analysis may produce largely accurate results for such semantic tasks, given large image corpora (e.g., estimating traffic flows from highway cameras), without yielding artifacts such as human-visible segment representations. Or, in a different domain, AI-powered analysis of FCS (Flow Cytometry Standard) data could establish a largely accurate count of "events" (i.e., discrete FCS measurements of light-scattering and/or flourescent properties of cellular-scale entities) without manual "gating" (referring to the conventional practice of scientists using geometric annotations of FCS data-plots to isolate and thereby count different event-types). An AI-powered analysis of image features, or likewise of Flow Cytometry data, may yield calculations similar to those which for *human* users are achieved via image segmentation, manual gating, and similar operations which clearly yield annotation data. For AXFI, the complication arises when these AI mechanisms do not themselves yield results that would normally be considered annotations, but rather yield the desired empirical results for which the annotations would be a preliminary step — e.g., an approximate count of the number of cars in a highway photo, without a precise segmentation of the image marking the cars' respective borders.

AXFI ultimate considers these AI-related comlications to be issues resolvable at the level of software, rather than stadardized annotation models *per se*. An image annotation is, among other things, a visual (or viewable) record of some image-processing activity. If a radiologist manually clarifies a report to the effect that a given CT shows a tumor by circling the area where the tumor is visible, he or she is using the image annotation to communicate to others the thought-process which motivated the diagnostic conclusion. This is different than an AI-driven processor which would automatically demarcate an image segment outlining the tumor and use geometric properties of that segment to derive a pathological finding. In short, the data conveyed in an annotation — an image segment, rendered precisely, or rendered indirectly via a circle or polygon around the segment — may be *intrinsic* to an image-processing operation: it may be data acquired *at one stage* in an analytic workflow. However, annotations may also be *retroactive*; if a radiologist circles a tumor, he or she has completed (at least mentally) the image analysis, and is using the analysis to summarize what occured in the course of the analysis. Therefore, any image-processing task can be associated with

ex post facto annotations which summarize the process even if they are not intrinsic to it.

To continue the example of counting cars from a highway photo, an **AI**-powered observation could be retroactively *justified* by providing an image segmentation where the number of car-segments matches the **AI** count. In lieu of precise segments, it may be simpler to provide location-points for the "geometric center" of each car, or the points furthest apart in the direction of each car's front-to-back — these may be the statistical signals used for the car-counting process. Anlogously, facial recognition does not need to rely one segmenting out regions (eyes, nose, lips), but rather can be based on distances between individual points (such as the inner corners of each eye). But in any case, depending on the analytic algorithm used, it is often possible to identify some spatial/geometric feature or object that can be visualized in the image context, and that summarizes or legitimizes the analytic operation. This summarial data, then, can provide *retroactive* annotations which allow human viewers to understand and review the algorithmic process. In short, simply because image-processing tasks may not generate annotation data as part of their interal activity, it is still possible (and may be desirable) for the software operationalizing these tasks to implement annotation generators, where the resulting annotations document the operations for the scientific record and/or summarize them in **GUI** objects for the benefit of human viewers.

In general, then, **AXFI** distinguishes human-generated from computer-generated annotations, and moreover leaves open the possibility that some computer-generated annotations are *retroactive*: that instead of being internal to an imaging computation they are indirectly produced subsequent to such a computation, for purposes of documentation and validation.

With these preliminaries concerning the scope of **AXFI** clarified, the following sections will (1) outline the kind of data encoded via **AXFI** and (2) the relation between **AXFI**, data types, and applications/libraries that may use **AXFI**.

Part I: Outline of the AXFI Data Model

Types of Image Annotations