|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-H-038-R01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | Brasilia, 22-24 January 2020 |
| **DOCUMENT** |
| **Source:** | De Montfort University, Costa Rica Institute of Technology; ADA (Germany) |
| **Title:** | Proposal for a data labelling standard and a public data labelling tool |
| **Purpose:** | Discussion |
| **Contact:** | Saul Calderon RamirezCosta Rica Institute of Technology/De Montfort University | Tel: +50683059141Email: sacalderon@itcr.ac.cr  |
| **Contact:** | Henry HoffmannAda Health GmbH | Email: henry.hoffmann@ada.com  |

|  |  |
| --- | --- |
| **Abstract:** | This paper proposes the development of a public web-based fast image ground truth authoring tool (GTAT). Image ground truth authoring tools are key to generate training and validation data for image segmentation and classification systems. This paper provides a short review of similar publicly available GTATs, the features, and the shortcomings, in order to spot the key features missing for a public GTAT to the community. Based on the desired features, we aim to develop a free and open GTAT in the future. This document provides a first draft of a list of requirements for a data labeling standard and the development of a community-based and public web-based ground truth authoring toolNOTE – This revision *replaces* H-038. |

1. **Technical requirements**
	1. Functionalities

The functionalities provided below are a preliminary list of requirements. Some of them are already implemented in a prototype described below. General goal: organize/plan/execute/monitor collection/creation/review of test data in a global, representative, scalable, safe, and trustworthy way. The following is a list of basic functionalities proposed.

* + 1. A canvas that allows manual image annotation with different brushes and tools.
		2. Web tool: the ground truth authoring tool must be web based, to ensure the tool’s portability, and open to the community.
		3. Flexible storage format: the format must allow its usage with ground truth comprising several (thousands) of ground truth masks, preventing performance degradation.
		4. A semi-automatic algorithm for mask initialization: the tool must implement a technique for initializing the foreground masks, and allow the user to make adjustments to the masks.
		5. A semi-automatic algorithm for mask readjustment: for video segmentation, the masks drawn for the first frame can be readjusted for next frames. An algorithm that automatically estimates such readjustment is useful for a GTAT, as implemented in Sensarea.
		6. Collaborative support and crowd-sourcing support: the tool must allow several concurrent users to work on the same project.
		7. Evaluation module: the tool must provide means to evaluate metrics for the segmentation algorithms, including but not limited to: sensitivity, specificity, accuracy, and F-score.
		8. Change history: the changes made by other users in the markings of the images are shown in the tool.
		9. Support for measuring the base-lines (community health workers, nurses, GPs, experts, ...) in the different levels of the different countries.
		10. If necessary, multi-language support.
		11. Protection against any access to benchmarking data.
			1. If necessary, encryption.
		12. Distributed backup
		13. Manipulation protection
			1. It must not be possible to manipulate data in the database without leaving traces.
		14. Timeline
	1. Meta-data representation

[TO DO]

Labeling information representation must support generic and pluggable specific medical concepts. Information should be provided regarding:

* + 1. Sample information: equipment used, model, and relevant calibration parameters.
		2. Annotator information: organization, experience, and background. Management of hierarchical entities system
			1. Benchmarking administrators, countries, clinics, departments, and annotators
			2. Specification of field of expertise (e.g. ICD10 based)
			3. Specification of level of expertise
			4. Performance scoring system (how often other annotators disagree)
			5. Resource management (so that it is clear how many cases an entity can do in which time frames)
			6. Access tracking
			7. Interface administration of sub entities for all levels
		3. Label description: class; used standard for clinical categorization with reference
		4. Label information: pixel positions; manually modified.
	1. Data requirements

 Image formats supported; support flexible multimodal annotation/labeling

Plugins for all relevant modalities

* + 1. 2D image area segmentation + labeling
		2. 2D image boundary segmentation + labeling
		3. 2D image entity counting + labeling
		4. 3D image area segmentation + labeling
		5. 3D image boundary segmentation + labeling
		6. 3D image entity counting + labeling
		7. Case vignette editor enforcing ontologies for symptoms, conditions, and factors
		8. ...

For nD imaging

* + 1. Possibly with modality specific helpers “magic wand” for some image task
		2. Pan, zoom, and rotate
		3. Dynamic contrast, dynamic coloring
	1. Active learning functionalities

A list of unsupervised/semi-supervised techniques supported.

1. **Ethical requirements**
	1. Privacy requirements
		1. DICOM anonymization is required for the usage of DICOM images.
		2. If necessary, GDPR-compliant data request handling (e.g., deletion requests by patients) with explicit documentation, tracking of the source of every case and label - including acquisition metadata for imaging (all relevant machine parameters, settings, etc.) documentation of any known AI company that had already access to parts of the data (for real cases) build in case/task based feedback/discussion mechanism access/ rights and role management minimizing visibility to the absolute minimum web based
	2. Licensing
2. **User requirements**
	1. Intended users

[TO DO]

* 1. Registration procedure

[TO DO]

* 1. Allocating tasks to users

[TO DO]

* 1. Collaborative flow
		1. Shared metadata management
			1. Flexible definition of meta-data dimensions
			2. Corresponding dynamic UI
			3. Basic metadata on case type (synthetic vs. real)
		2. Task management system
			1. Raw data set importing, ticket generation tool
			2. Assignment of annotator tickets for
				1. Synthetic case creation
				2. Real case annotation
				3. Peer reviewing of cases
			3. Assignment of case-distribution volume tickets for higher level entities (e.g,. clinic xyz; create 1000 cases with this age, sex, disease, and real/synthetic distribution)
			4. Corresponding notification system
				1. For instance, via email
			5. Task inbox at all levels (but mainly for annotators)
			6. Time tracking (relevant for funding, payment)
		3. Peer review management
			1. Automatic review selection
			2. Dynamic scheduling of more peer reviews in case of conflicts
			3. Automatic conflict resolution
		4. Progress/monitoring system
			1. On all levels
			2. High-resolution distribution analysis to see if the planned data sets distributions along all the defined dimensions have been achieved
1. **Testing**

 A list of suggested performance and usability metrics are to be enlisted.

* 1. Benchmarking data collection campaigns management system
		1. Creation of new campaigns
		2. Planning of case-numbers, time-frames
		3. Definition of the distributions along all configured dimensions (e.g. age, sex, region, ethnics, diseases, symptoms, multimorbidities, genetic predispositions, risk factors (smoker, pregnancy, ...)
			1. Possibly connection of statistical databases where to draw the country level statistics on age, ethnicity distributions etc.
		4. Tools for statistical representativeness / stability estimation
		5. Definition of peer review details
			1. Number of reviewers
			2. Agreement/conflict resolution thresholds
		6. (definition of additional exclusion criteria e.g. pregnant women, kids < x months)
		7. Benchmarking export data
			1. Once all data for an benchmarking iteration was collected and annotated

**Prototype: Insight GT for medical image annotation**

Available at https://insight-gt.hulilabs.xyz/canvas/. For region initialization, SLIC superpixels are computed (Achanta et al., 2010). A javascript implementation of the algorithm can be found at Tangseng et al. (2017).

Insight GT: An example of a GT tool using the developed standard

<https://www.researchgate.net/publication/336927724_Insight_GT_a_public_fast_web_image_ground_truth_authoring_tool>

<https://insight-gt.hulilabs.xyz/canvas/>

Additional considerations

* documentation/tracking of all necessary GDPR relevant patient consent tracking
* definition of the different iterations of such a tool
* technology options
* frontend
* JS react + redux + redux-saga?
* backend
* discussion of software process for building such a tool
* git
* open source
* discussion of who could run / host the tool
* discussion on who could/should organize/setup the benchmarking campaign and who has the right to export a dataset
* discussion how the exported dataset is then transferred in a safe way to the benchmarking system

The following the proposed functionalities implemented so far in our prototype available at https://insight-gt.hulilabs.xyz/canvas/. For region initialization, SLIC superpixels are computed [15]. A javascript implementation of the algorithm can be found at [16].

* { A canvas that allows manual image annotation with di\_erent brushes and
* tools, as seen in [9, 17, 18], aided by region initialization, using super pixels
* or a similar algorithm.
* { Web tool: The GT authoring tool must be web based, to increment the tool
* portability, and open to the community.
* { Flexible storing format: The format must allow its usage with GT compromising
* several (thousands) of GT masks, avoiding performance degradation.
* { A semi-automatic algorithm for mask initialization: The tool must implement
* a technique for initializing the foreground masks, and allow user to
* make adjustments to the masks.
* { A semi-automatic algorithm for mask readjustment: For video segmentation,
* the masks drawn for the \_rst frame can be readjusted for next frames. An
* algorithm which automatically estimates such readjustment is useful for a
* GTAT, as implemented in Sensarea.
* { Collaborative support and crowd-sourcing support: The tool must allow several
* concurrent users working on the same project.
* { Evaluation module: The tool must provide means to evaluate metrics for the
* segmentation algorithms, including but not limited to: sensitivity, speci\_city,
* accuracy, F-score, among others.
* { Change history: the changes made by other users in the markings of the images are shown in the tool.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_