Data Management Plan

A major drive of the proposed work is to establish new methods for taking advantage of new kinds of annotation for achieving better progress in machine learning and machine vision. While our proposed work includes our own initial forays into this new space, we anticipate the greatest possible impact will be achieved by fully engaging and leveraging a much broader community of investigators. As such, a significant component of the proposed work is the creation and development of publicly-sharable material – both datasets and software.

This plan is based on NSF's policy on the dissemination and sharing of research results within a reasonable time. In accordance with this policy, this plan does not include preliminary analyses (including raw data), drafts of scientific papers, plans for future research, peer reviews, or communication with colleagues.

Furthermore, data to enable peer review and publication/dissemination and/or to protect intellectual property may be temporarily withheld from distribution and other proposed data management. This plan will make certain that the data produced during the period of this project is appropriately managed to ensure its usability, access and preservation.

1 Project Deliverables

In particular, we will deliver the following:

- **A** repository of human psychophysical measurements corresponding to standard machine vision sets. For face detection, we plan to target the Annotated Faces in the Wild (AFW) and Face Detection Database (FDDB) data sets. For attributes, we will target the PubFig set. In all cases, the TestMyBrain infrastructure is carefully designed to collect no identifiable human subject data so as to avoid privacy concerns. The operation and dissemination of data derived from TestMyBrain has been reviewed by the Harvard Committee on the Use of Humans and Experimental Subjects.

- **A** reference implementation of oracle-assisted, deep-annotated machine learning. Code for this implementation will be made available under a creative commons license in a publicly-accessible repository (e.g. on GitHub.com).

- **C**ode for running experiments (client-side javascript) and code for analyzing psychophysical data. As a matter of standard practice, code for running psychophysical experiments (suitable for running on Mechanical Turk as well), will be made publicly available under an MIT software license. Likewise, code required to analyze this data (e.g. to generate psychometric curves) will also be released.

- **A** face detection data set with realistic occlusion. A new data set of faces in realistic occlusion conditions will be collected. Care will be taken to work with the Harvard IRB to allow collected images to be shared with the research community.

2 Data and Code Sharing Timeline

We anticipate regular releases of data and code as they are completed. In keeping with the overall schedule presented in the Project Description found of page C-14 of the proposal, we anticipate the release of initial attribute-related psychophysical data to take place in month 30, with periodic subsequent releases as protocols are refined and iterated. Likewise, we anticipate the initial release of face detection-related data in month 30, and the initial release of software in month 36.

3 Formats and Access

Psychophysical data will be shared in standard formats (e.g. plain text, hdf5, csv), without subject information. Data release will include both raw subject data as well as derived data (e.g. psychometric curves computed from the raw data) along with the code produces the derived data. Files will be hosted on servers controlled by Harvard University, and will be discoverable both through the TestMyBrain.org website and through websites of the individual investigators.