Live link:
https://docs.google.com/document/d/1ee5J6ahatOOvX88ZCrm5dWWeV57ArNyowIWL77Hf5Ho/edit?usp=sharing

Using Perceptual Image Hashing/Machine Learning to Identify Like Images for More Refined Archival Appraisal and Selection

Lora Woodford

June 21, 2017

With the advent of fully digital photography it has become commonplace for photographers – both casual smart phone users and trained professionals alike – to capture images at a scale that dwarfs prior practice. University events for which campus photographers would have, in the past, snapped dozens of pictures, are now recorded in hundreds (if not thousands) of raw digital files. As we approach the third full decade since the commercial availability of digital cameras, archives and other cultural heritage institutions must continue their missions to preserve records of enduring value, while also acknowledging that both real resource limitations (storage costs, human capital, etc.) and the archival principles of appraisal and selection[1] warn against the practice of simply keeping every photograph snapped. Digital curators rely on many existing digital forensics applications to, among other things, identify exact duplicates (typically based on cryptographic hash values), extract and analyze file metadata, and identify potentially confidential information (e.g. credit card numbers, social security numbers, etc.); however, no existing tool currently allows for archivists to identify near-duplicate digital images based on image visual content. To be viable and widely applicable in the digital archives community, such a tool would have to:

1. Allow a user to point to a directory of images and recursively compare the *visual content* of those images (likely using perceptual image hashing[2]) to one another in a many-to-many scenario and alert the user to groupings of images that are most likely to have like content;
2. Allow for complete transparency as to how content grouping was achieved for each batch of images processed and allow for adjustments in sensitivity (e.g. a digital preservation friendly configuration file that identifies threshold values used that can be packaged with archival information packages (AIPS));
3. Allow for the tool to be run locally without having to upload all images to an off-site server given both the sheer size of and potential restricted content in the type of born-digital image collections archives routinely acquire.

It seems likely that a tool or tools already exists and is widely used in allied fields such as GIS, data management, audio-visual fields, machine learning, etc., but these tools are not widely known to archivists.

[1] The Society of American Archivists' online glossary defines "selection" as: "The process of identifying materials to be preserved because of their enduring value, especially those materials to be physically transferred to an archives." See: https://www2.archivists.org/glossary/terms/s/selection

[2] See: http://www.phash.org/