Multimedia Metadata-based Forensics in Human Trafficking Web Data

Chris A. Mattmann[1,2], Grace Hui Yang[3], Harshavardhan Manjunatha[2], Thamme Gowda N[2], Andrew Jie Zhou[3], Jiyun Luo[3], Lewis John McGibbney[1]

[1]Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109 USA

[2]Computer Science Department
University of Southern California
Los Angeles, CA 90089 USA

[3] Department of Computer Science
Georgetown University
Washington, DC, 20057 USA

mattmann@jpl.nasa.gov, huiyang@cs.georgetown.edu,

ABSTRACT

Crawling the web for ads related to human trafficking yields a wealth of traditional static and dynamic web page content. Ads in which victims are trafficked by their predators are often littered with textual signals such as physical characteristics (hair, body type, race, ethnicity, etc.), and location information (city, state) that can aid law enforcement and non governmental organizations in identifying victims and intervening to aid them. The ads also carry a strong multimedia footprint, as there are increasingly images and videos to accompany the ad text. Many groups have researched computer-vision (CV) based approaches to analyze these images and videos to extract meaningful features that when combined with the physical characteristics and location information can greatly aid in thwarting these activities. However CV approaches are computationally expensive and further they may not be able to discriminate between images and videos taken in similar lighting and color situations as they rely greatly on image analysis which is limited to scene properties. Our team has investigated and created a preliminary system that takes advantage of image and video multimedia metadata – or information about the actual images and videos typically recorded when they are created. This metadata can be leveraged to provide valuable discriminatory signal to aid in search and relation of multimedia data while saving computational cost and also providing a novel complimentary feature set when CV techniques are unable to differentiate between images and videos.

CCS Concepts

• **Information systems~Information retrieval** • **Information systems~Content analysis and feature selection** • Information systems~Similarity measures • Information systems~Information extraction.

Keywords

Apache Tika, Jaccard Similarity, Metadata, Multimedia, Forensics

1. INTRODUCTION

Human trafficking (HT) is a global modern phenomenon in which modern day slavery is enabled through the public Internet. Women, men, and children are sold for sexual slavery, labor-based slavery and for other potentially criminal activities [1]. The ease of use of the Internet as a large bulletin board complete with the ability to upload images and videos to describe the victims has contributed greatly to the growth of HT throughout the world. The size of publicly available HT data is estimated to be around 60 million ads, and 40 million images and 10s of thousands of videos already eclipsing the ability for single computers to process and analyze the information.

Law enforcement, non governmental organizations (NGOs) and other groups are interested in collecting publicly available ad postings, multimedia information (images, video, etc.) and other data to assist in identifying human trafficking victims, and in finding and prosecuting predators responsible for the trafficking. This data can be collected by using web crawler software to download public ad and bulletin board data from HT websites. Crawling the web for ads related to human trafficking presents an interesting challenge both in scale in terms of the number of web sites and number of ads and multimedia, but additionally in terms of search and information retrieval. Ads for trafficking victims regularly have textual based signals that can aid law enforcement and NGOs including physical information about the victim such as race, ethnicity, gender, body and hair type, etc., along with physical and forensic information about the location in order to connect buyers or "johns" through the Internet to the trafficking victim. Textual signals are an important element that can be used to identify victims and relate them together. Manual approaches for analyzing textual signals in HT ads remain in use by NGOs and law enforcement today.

Beyond textual signals are the rich multimedia images and videos present that can be leveraged along with the textual ad data for search and retrieval in the HT domain. Current research is focused on computer vision (CV) techniques [2] and in advanced approaches such as machine learning and deep learning on CV [3] to relate together images and videos and to relate objects in both to scenes, and to places and things. Some limitations of CV based approaches to multimedia search are their computational cost. Even if resources are available to leverage CV, CV techniques are typically based on image analysis and thus relations made between images, videos, and further they may not be able to discriminate between dissimilar images and videos taken in similar lighting and color situations. This is most often seen in cropped images and videos that CV techniques have difficulty discerning.

Our team has investigated and created a preliminary system that takes advantage of image and video multimedia *metadata* – or information about the actual images and videos typically recorded when they are created. The system exploits content creation metadata as it propagates throughout the image and video lifecycle: from creation to editing and manipulation and to dissemination. Metadata includes useful properties related to both the physical properties of the content (RGB color space; whether or not the flash fired;) to geo-location, to information about the instrument that captured the multimedia (camera/phone Make/Model; Serial number;) to other information such as creator, date/time the multimedia was generated. Our system, *Image Space*, leverages metadata to overcome CV-only oriented approaches, and to compliment image and video analysis techniques, with metadata-based forensics to better combine multimedia with ad-based data in the HT domain. In particular, our techniques have shown to identify image and video relationships to ads, and to identify relationships between victims and predators in ways otherwise not possible without our approach.

The rest of this paper is organized as follows. Section 2 describes Image Space, our metadata forensics toolkit for multimedia. Sections 3 and 4 describes our algorithmic approach for relating images and videos together based on multimedia metadata and on domain dynamics. Section 5 concludes the paper by identifying next steps and future work.

2. A METADATA FORENSICS TOOLKIT FOR MULTIMEDIA

We have constructed the ImageCat and ImageSpace architecture depicted in Figure 1. ImageCat – short for "Image Catalog" and shown in the bottom middle of Figure 1 – is an Extract, Transform and Load (ETL) system to automatically create a search index of Image and Video similarity metadata descriptors by extracting that information from a collected set of multimedia data (10s of millions of images/videos). Though we have fielded ImageCat in the HT domain, it is also applicable to any multimedia data and images collected on the Internet. ImageCat uses Apache Tika [7] to automatically identify multimedia files and their type, and in turn to invoke open source, third party parsing libraries on the multimedia data. For example, EXIFTool is a useful Perl-based program that extracts EXIF image and video metadata – Camera and scene properties, make, model, color space information, etc. – and is integrated into Tika. FFMPEG extracts video information such as bitrate, tracks, scene properties, etc. – and is integrated into Tika. Besides EXIF metadata and other descriptors, Apache Tika integrates the Tesseract [4] library to perform Optical Character Recognition (OCR) and to extract text descriptors from multimedia data as well – as shown in the middle right portion of Figure 1. We will detail the use of OCR later in this section.

Apache Tika is integrated into the Apache Solr search indexing system via a plugin called "SolrCell" that automatically runs Tika on the server side during the indexing process. To perform indexing of the information from crawled web data, a list of image file paths, possibly as large as tens of millions of images, is presented to ImageCat (left side of Figure 1), and those images are run through Apache OODT [5] a data-flow oriented ETL system. OODT's File Manager (FM) tracks and records file

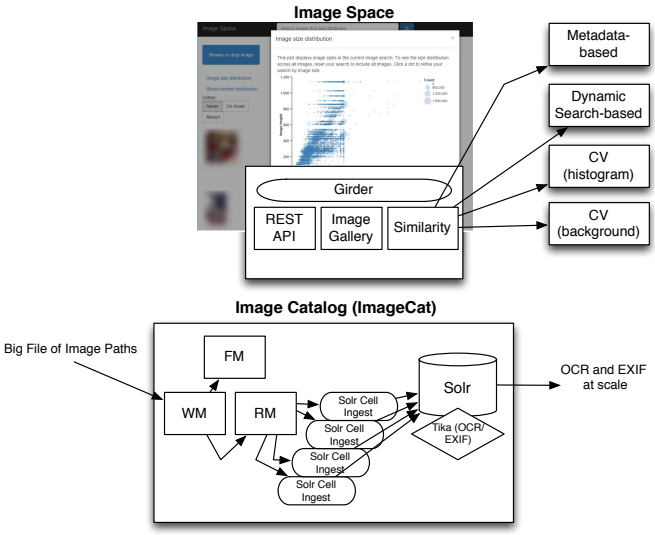

Figure 1. The Image Space and Image Cat architecture.

locations and metadata; its Workflow Manager (WM) manages provenance of the ETL pipelines, and the Resource Manager (RM) schedules the ETL jobs on a large cluster or the cloud. Each job in ImageCat is an ingest job that sends the image or video file to SolrCell for server-side Tika-based extraction and for building the inverted index and catalog. OODT records all the provenance information from file to ETL to resource scheduling. This is useful information especially since ImageCat can quickly be constructed, and/or torn down depending on adaptations to Tika and when changes or fine-tuning are necessary.

OCR and EXIF metadata are particularly important to images and videos crawled from the HT domain. Victims usually hold placards containing vital information like phone number, email address, which can be parsed by Tika and Tesseract. Although the OCR from Tesseract contains noise, partially parsed text can be leveraged as keywords in search against Solr and has been found to be effective.

ImageSpace is a front-end web application for ImageCat. It connects directly to an ImageCat and can be used for searching and querying browsing large collections of multimedia information in an efficient manner using Apache Solr. ImageSpace is built on top of the Girder web framework, and it presents a REST API (shown in the upper left of Figure 1) for interactively extracting multimedia metadata with Tika; for performing similarity comparisons based on metadata and CV techniques, and by temporally associating user's queries as we will explain in Section 4. ImageSpace supports both text-based search and multimedia based searches against any of the extracted metadata properties. ImageSpace provides an interactive density plot feature that visualizes the distribution of images (based on Image Size). Similarly ImageSpace also provides an interactive histogram feature (binned on camera serial numbers) to refine your current image search session or across all images in the index. The Text based approach searches based on the parsed textual content (e.g., OCR) of images and the value of other image metadata attributes. Common text-based multimedia search queries include email address, etc. The image-based approach allows users to search for similar images by uploading a query image onto the ImageSpace application, or by selecting one of the images from the text based search results. Upon uploading or selecting an image, the metadata attributes of that image are displayed along with the various similarity metric options to

search upon using the chosen image as a query. The similarity metrics are indicated in the upper right portion of Figure 1, and correspond to image content (histogram), background of the image, size/resolution of the image, and other metadata attributes. An experimental temporal image search is described in Section 4. Our multimedia similarity metric is described in the next section.

3. METADATA BASED MULTIMEDIA SIMILARITY

Our team has also derived a metadata-based similarity metric allowing for clustering and grouping of multimedia data forensically, without relying on computationally burdensome computer vision (CV) techniques. The approach is derived from an observation that metadata creators and content creators leave a forensic footprint that propagates through the content dissemination process – from authorship and tool (e.g., Camera; Video Recorder; Phone) to editing (Photoshop, GIMP, etc.) to delivery (web-server Headers etc.) to browser or image/video reader.

Algorithm 1. Tika Jaccard Similarity

```
1  input: directory of files d
2  output: scores s for all files in d
3
4  goldSet:= {}
5  allMetadata:= {}
6  scores:= {}
7
8  for file in d:
9    text, metadata:= tika.parse(file)
10     goldSet:= goldSet ∪ metadata.keys
11     allMetadata[file]:= metadata
12
13 goldenSetSize:= |goldSet|
14
15 for file in allMetadata.keys:
16     overlap:=|allMetadata[file]∩ goldSet|
17     score:= overlap / goldenSetSize
18     scores[file]:= score
19
20 return scores
```

The approach, shown in Algorithm 1, builds upon the *Jaccard* similarity algorithm [1]. Leveraging an Image or Video's metadata *footprint*, the algorithm works as follows. We leverage the extracted Tika metadata and text descriptors from ImageCat. A golden feature set is computed by iterating all the files in a given directory and extracting out metadata key names using Tika (e.g., *EXIF Flash*, *RGB Color Space*, *Camera Make*, *Camera Model Serial Number*, etc.) and/or their discretized value space (line 10 in Algorithm 1). The features are collected across all multimedia files, as is the per file metadata. A second pass through the set of files in line 15 shown in Algorithm 1 is performed so that for each set of per file extracted metadata, the intersection of each file's space can be computed and compared with the entire set of features in the golden feature set (read: *all metadata property names and/or values* across the entire set of collected data). This allows a distance metric to be derived that shows each files computed distance from the golden set, allowing for metadata-based multimedia characterization.

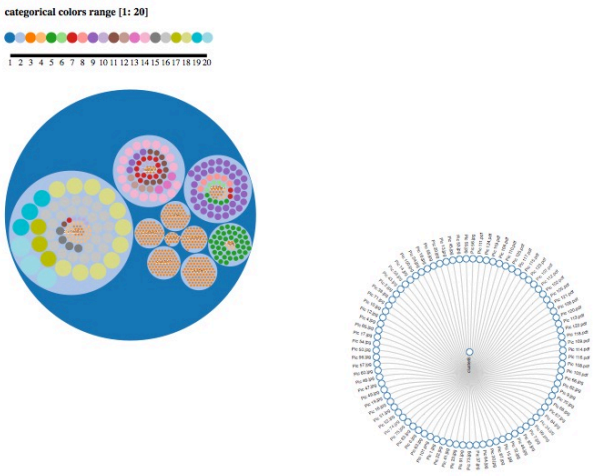

Figure 2. Image Similarity Computed from H/T data.

Once the scores are computed from Algorithm 1, a simple clustering threshold technique can be applied to visually create image, video and multimedia clusters based on metadata names and/or value spaces. These clusters can then be visually represented using our technique built upon the Data-Driven Documents (D^3) framework [8] as shown in Figure 2. In the upper left are the cluster groups derived from Jaccard's coefficient differencing. Each group represents a set of items that share the same metadata features. In the bottom right are the actual images and/or videos present in the cluster. Color in the upper left clusters are per feature and indicate the number of times that metadata field is present in the cluster, and the density of the circles are used to represent the number of files in the cluster.

What we find in practice using this technique is that it naturally groups multimedia files that were either edited, created, captured, or modified in the same vein and/or fashion. This can derive and provide new meaning unable to be discerned visually or using CV techniques. For example, scenes and pictures, and videos in which the objects in the scene and/or background are not related by color or by histogram or by general look and feel can be related in the sense that the pictures were captured using the same camera; or same camera type; or same camera settings which provide information as to the content creator. In the HT domain, this is typically the predator (and also depending on the stage of editing, the victim) and so metadata multimedia forensics can be a useful tool in augmenting existing CV-oriented techniques with new ways to associate multimedia information with textual ad-based features – the domain of trafficked weapons yields similar results. In the next section we will describe another approach that our group is pioneering to relate images together based on the way that users query ImageSpace and ImageCat for multimedia information.

4. DYNAMIC SEARCH OF HUMAN TRAFFICKING DATA

Dynamic search is an emerging topic in Information Retrieval (IR) research [11]. In dynamic search, we model dynamic systems that change over time or a sequence of events using artificial intelligence and reinforcement learning.

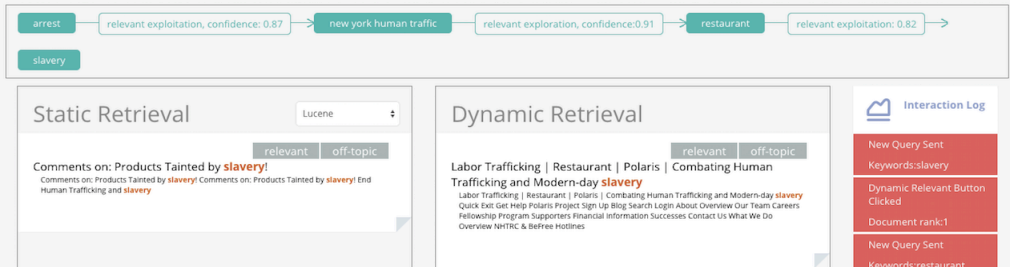

Figure 3. Dynamic Search Engine Built for Human Trafficking Dataset

In a dynamic search setting, a user issues multiple queries during a session to accomplish a search task. The common process is that the user sends a query, gets the top ranked documents, changes her query and sends it again. As a result, a series of queries, a series of retrieved documents, and rich user interactions will be generated, until the session stops when the user finishes her search task. During a session, the temporal dependencies between queries are presented in that way that the previous queries and the previously obtained search results will influence how the user issues the current query and how the search result rankings could be optimized for the current query.

We have preliminary integrated dynamic search into our ImageSpace application. Figure 3 shows the interface of a dynamic search engine developed for the HT dataset and integrated as a similarity metric into ImageSpace. Our dynamic search metric supports the digital forensic process that requires detailed, iterative, and complex queries and is based on the metadata we extract using SolrCell and Tika. We have demonstrated that the dynamic search process is more effective than using a static off-the-shell search engine especially when users are able to use ImageSpace and query it over time.

For instance, in a case to discover all the massage parlor multimedia data related to arrests in the United States, the user entered a series of queries into ImageSpace. The first query is "arrests". The user scans the images and videos and finds that the second image is relevant. After clicking the second image, the user found the phrase "New York" in the extracted second iamge and used it to form the second query "New York human trafficking". In the third returned result, a video, the user found a name "Christopher Robert" – a predator whose name was present in the metadata authorship from a picture taken with his camera - and used it as the next query. Because the dynamic search engine always keeps the context in mind, the search for a person named Christopher is not out of the scope of human trafficking, while a static search engine would return persons outside of this criminal domain on the top of the document list without saving this context. We are excited by the preliminary results from applying this approach to multimedia data and look forward to next steps.

5. CONCLUSION AND FUTURE WORK

We have described our approach to multimedia forensics in the Human Trafficking domain and more generally for relating images and videos together with text when querying information collected from the web. While the early results from our approach appear promising, a number of pertinent areas remain unexplored.

The cosine distance algorithm is under evaluation for more precise metadata clustering. In addition, we are exploring the effectiveness of dynamic search similarity relating to traditional CV approaches and open datasets from the ACM Multimedia conference. We are also adding additional video similarity metrics including the Pooled Time Series approach. Finally we are addressing some limitations of the approach including discerning the optimal metadata for clustering; finding more scalable ways (e.g., REST-ful services) to invoke additional extractors and choosing an optimal set of extractors for the same content types.

6. ACKNOWLEDGMENTS

This work was supported by the DARPA XDATA/Memex program. In addition, the NSF Polar Cyberinfrastructure award numbers PLR-1348450 and PLR-144562 funded a portion of the work. Effort supported in part by JPL, managed by the California Institute of Technology on behalf of NASA.

7. REFERENCES

[1] New Search Engine Exposes the Dark Web, http://www.cbsnews.com/news/new-search-engine-exposes-the-dark-web/, Accessed: November 2015.

[2] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision* 8.2 (1992): 99-111.

[3] M. Abadi, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. tensorflow.org, 2015.

[4] R. Smith. An overview of the Tesseract OCR engine. *IEEE ICDAR* 2007.

[5] C. Mattmann, et al. A reusable process control system framework for the orbiting carbon observatory and NPP Sounder PEATE missions. *IEEE Space Mission Challenges for Information Technology*, 2009.

[6] Jaccard Index, https://en.wikipedia.org/wiki/Jaccard_index, Accessed: November 2015.

[7] C. Mattmann and J. Zitting. *Tika in Action*. Manning Publications, 2011, 256 pages.

[8] M. Bostock, O. Vadim and J.effrey Heer. D³ data-driven documents. IEEE Transactions on Visualization and Computer Graphics, 17.12 (2011): 2301-2309.

[9] Foto Forensics, http://fotoforensics.com/tutorial-meta.php, Accessed: November 2015.

[10] M. A. Anoop. Image forgery and its detection: A survey. IEEE International Conference on Innovations in Information, Embedded and Communication Systems, 2015.

[11] Jiyun Luo, Sicong Zhang, Hui Yang. Win-Win Search: Dual-Agent Stochastic Game in Session Search. In Proceedings of the 37th Annual ACM SIGIR Conference (SIGIR 2014).

[12] Hui Yang, Marc Sloan, Jun Wang. Dynamic Information Retrieval Modeling. Tutorial in the 37th Annual ACM SIGIR Conference 2014 (SIGIR 2014). Gold Coast, Australia.