Internet Image Archaeology: Automatically Tracing the Manipulation History of Photographs on the Web

Lyndon Kennedy
Dept. of Electrical Engineering
Columbia University, New York, NY 10027
lyndon@ee.columbia.edu

Shih-Fu Chang
Dept. of Electrical Engineering
Columbia University, New York, NY 10027
sfchang@ee.columbia.edu

ABSTRACT

We propose a system for automatically detecting the ways in which images have been copied and edited or manipulated. We draw upon these manipulation cues to construct probable parent-child relationships between pairs of images, where the child image was derived through a series of visual manipulations on the parent image. Through the detection of these relationships across a plurality of images, we can construct a history of the image, called the visual migration map (VMM), which traces the manipulations applied to the image through past generations. We propose to apply VMMs as part of a larger internet image archaeology system (IIAS), which can process a given set of related images and surface many interesting instances of images from within the set. In particular, the image closest to the "original" photograph might be among the images with the most descendants in the VMM. Or, the images that are most deeply descended from the original may exhibit unique differences and changes in the perspective being conveyed by the author. We evaluate the system across a set of photographs crawled from the web and find that many types of image manipulations can be automatically detected and used to construct plausible VMMs. These maps can then be successfully mined to find interesting instances of images and to suppress uninteresting or redundant ones, leading to a better understanding of how images are used over different times, sources, and contexts.

Categories and Subject Descriptors: H.4 [Information Systems Applications]:Miscellaneous

General Terms: Algorithms, Experimentation

Keywords: Internet Image Mining, Image Manipulation History, Perspective Discovery

1. INTRODUCTION

Archaeologists gather artifacts and objects from past civilizations in order to learn more about the people and cultures that have inhabited the planet throughout history. In our modern society, we create and archive many types of artifacts at an increasingly growing rate. Specifically, with the proliferation of the world wide web and the continued growth and simplification of web publishing tools, it has become amazingly easy for any of us to share and publish our ideas online. These stories and ideas are frequently enhanced by the inclusion of photographs and images which support the viewpoints and messages of the authors. In many cases, through the use of photo editing software, the images can become highly manipulated, with portions added, removed, or otherwise altered. How do these manipulations affect the meanings conveyed by the images and the documents that they accompany? If we are cognizant of these effects, can we conduct "archaeological digs" in online image repositories to gain a better understanding of the beliefs held by image authors? Can we use this knowledge to design systems that enhance image browsing and retrieval?

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

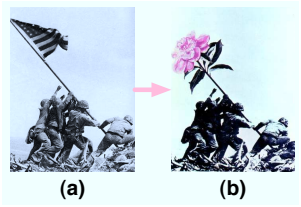

Figure 1: Image meaning can change through manipulation.

In Figure 1a, we have the famous photograph, *Raising the Flag on Iwo Jima*, which was taken by Joe Rosenthal shortly after the World War II battle in Iwo Jima in 1945. Immediately, the image was reproduced in newspapers all across the United States. By and large, the intention of distributing the image was to spread national pride and convey a highly patriotic and supportive view of the United States' use of military force. In Figure 1b, we see a photomontage made for an anti-Vietnam War poster in 1969 by Ronald and Karen Bowen, which replaces the flag in the soldiers' hands with a giant flower. The objective here is to take the originally pro-war image and subvert its meaning into anti-war imagery.

1.1 Visual Migration Map

These image manipulations do not exist in a vacuum, of course. Each image has a history and context that grows over time. We hypothesize that it is not the typical case that users are exposed initially to some original version of the image and decide to derive an image directly from the original. Instead, users may be exposed to some version of the image that is already derivative in some respect. Perhaps it is cropped, scaled, or modified with overlays. And frequently, they may also be exposed to text surrounding the image, which conveys a story and shapes the user's in-

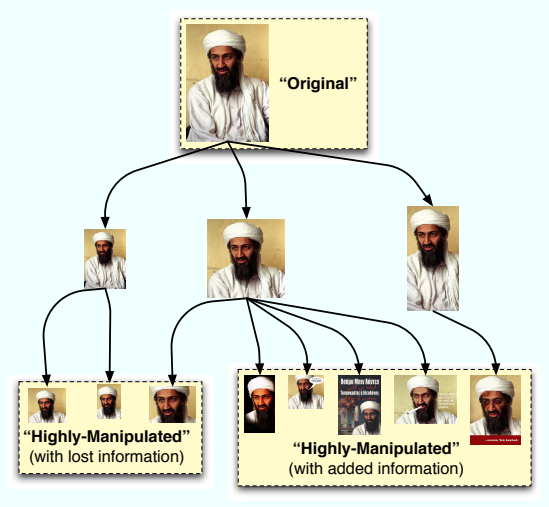

Figure 2: Hypothetical Visual Migration Map (VMM) showing the manipulation history of an image.

terpretation of the image. So, in effect, it is unlikely that every manipulated image is directly descended from the original image. Instead, each image is likely the result of many generations of manipulations and changes in meaning and context. In Figure 2, we see a hypothetical view of what the manipulation history, or visual migration map (VMM), of an image might look like. At the top, we have the original image. Children of images can be spawned through any number of manipulations. And then those children can, in turn, spawn more children. We expect this tree to have a number of characteristics that can be helpful for image exploration or search. These characteristics, the "original" and "highly-manipulated" images are shown in Figure 2.

Original versions of images would be thought to be the closest to the very first instance of the photograph. These would be highest-resolution, contain the largest crop area, and be subject to the least manipulation. These versions may be the most relevant for some cases in image search.

Highly-manipulated images are the ones falling the most generations away from the "original." On the one hand, there are images which are highly-manipulated in terms of simply having information removed, through excessive cropping and down-scaling. These are sometimes not of much interest, since they do nothing to enhance or change the meaning conveyed in the original version. On the other hand, there are images which are highly-manipulated in the sense that they have a great deal of external information overlayed on top of the original image. These are actually quite likely to be of interest to the user, since the meanings of the images may have been significantly altered.

In this work, we develop a framework and component techniques for automating the image exploration process. We contend that, given many manipulated instances of a single image, and information about the visual migration map between these instances, we can identify points along the evolution of the image which may be of particular value to a viewer or a searcher. In particular, we hypothesize that the above-described "original" and "highly-manipulated" versions of the image will be interesting to users. We take the stance, however, that it is difficult, if not impossible, to obtain a true history of image manipulations, since, without explicit information from image authors, we simply have no definitive proof of a parent-child relationship between any two images (i.e. we cannot know if one image was directly derived from the other). Furthermore, it is infeasible to obtain every single copy of an image. A web crawler might be able to find and index nearly all the images on the web, but, surely there are many images that were never digitized or posted online, which leaves gaps in the image history.

Nonetheless, we suggest that the results of a web image search for many important people or photographs will yield many dozens of copies of important images, which is sufficient for mining manipulation histories. We further propose that, despite the fact that we cannot truly know the parent-child relationships between two images, the low-level pixel content of images gives significant clues about plausible parent-child relationships. This entire process can be automated. The VMMs that emerge from this automated process are different from the true VMM of the image in many ways. Specifically, the parent-child relationships are merely *plausible*, and not necessarily true. An equally important aspect is that *implausible* manipulations (where no links are created) are detected much more definitively. If an image is not in the ancestor path of a manipulated image, then there must be information in the image extending beyond what's contained in the higher level image. Given these characteristics, we find that these automatically-constructed VMMs have many of the important characteristics necessary for building search and browsing applications and the "original" and "highly-manipulated" images found are exactly the images of interest that we are looking for.

1.2 Internet Image Archaeology

We test this visual migration map framework in the context of a larger internet image archaeology system (IIAS), shown in Figure 3. The IIAS framework takes in a set of related images (such as the results of a web image search), finds candidate sets of duplicate images derived from common sources, and then automatically extracts the visual migration map. The VMM can then be applied to many interesting applications, such as finding particular versions of the image or exploring the perspectives that the images convey.

We find that these search results frequently contain many repeated instances of the same image with a variety of manipulations applied. We examine the most-repeated images within these results and develop a series of detectors to automatically determine if particular edits (such as scaling, cropping, insertion, overlay, or color removal) are present. We find that many of these edits (scaling and color removal) are detectable with precision and recall both above 90%. The remaining edits are sometimes detectable, with precision and recall values in the range of 60-80%. Using these atomic detectors, we construct a plausible edit history for the set of images. We find that, despite the errors in the individual edit detectors, the system constructs manipulation histories that are highly similar to histories manually constructed by humans. Furthermore, the automatic histories can be used to surface "interesting" images from within the set. We show that the system has strong performance in retrieving "original" and "manipulated" images and that these highly manipulated are often correlated with differences in viewpoints being expressed by authors. Mining these image changes and their associated viewpoints could be of great interest to many different users, such as information ana-

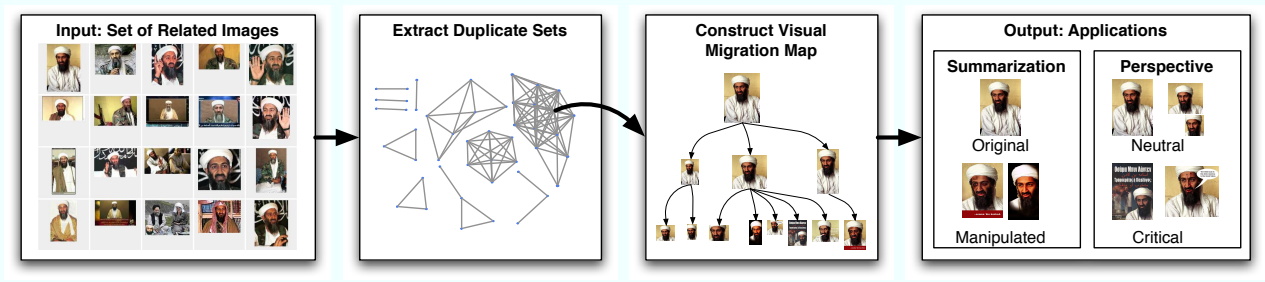

Figure 3: Proposed Internet Image Archaeology System (IIAS) framework. Given a set of related images, sets of duplicates or edited copies of various kinds can be extracted. For each set of edited copies, a visual migration map, representing the history of the image can be leveraged to summarize or explore the images and perspectives contained within the set.

lysts interested in tracking public responses to international figures, social scientists involved in understanding the user behavior and media use patterns on the Internet, or professionals looking to assess the way that celebrities or products are being received by the public.

The unique contribution of this work is a framework for an internet image archaeology system, which can be used to surface interesting images and viewpoints from a set of related images. The IIAS framework's key component is the visual migration map, which automatically extracts the manipulation history from a set of copies of an image. A key insight in this work is that image manipulations are directional, meaning that a positive detection result for one of these manipulations implies that one image might have been derived from the other. If all of the detectors agree about the direction of an edit, then we can establish plausible parent-child relationships between images. Across many different images, these relationships give rise to a graph structure representing an approximation of the image manipulation history, which can in turn be used to surface interesting images in exploration tasks.

The remainder of this paper is organized as follows. We discuss the proposed framework and system in Section 2 and give details about automatic component technologies and implementations in Section 3. In Section 4, we discuss our experimental results and propose some applications in Section 5. In Section 6, we discuss related work and offer some conclusions and thoughts on future direction in Section 7.

2. MINING MANIPULATION HISTORIES

Given a set of related images, either vast (i.e. every image on the web) or confined (i.e. the top results of an image search), we would like to discover interesting images and perspectives. To achieve this, we first extract the visual migration map, which will surface cues about the history of these images. We propose that such a rich understanding of the history of related images can be uncovered from the image content via pairwise comparisons between each of the related images in the context of a plurality of instances of the image. In this section, we lay out the intuition that we will rely upon to automatically detect VMMs.

2.1 Pairwise Image Relationships

Given two images, we can ask: "are these images derived from the same source?" This implies that a single photograph was taken to capture a scene and that various copies of images can then be derived from this source via a series of physical or digital manipulation operations. The figures throughout this paper give many examples of the typical appearances of pairs of images derived from the same source. It is sufficiently feasible to detect whether or not two images are derived from the same source using existing copy detection approaches [3, 9]. What is interesting about these related copies of images is that the operations applied to derive new copies give rise to artifacts within the image content which can tell us a great deal about the history of the image.

2.1.1 *Directional Manipulations*

Once we have established that two images are copies of each other, it remains for us to determine whether or not one is descended from the other. The key intuition behind this work is that image manipulations are directional: it is only possible to derive the more-manipulated image from the less-manipulated image. Below, we have enumerated a number of possible edits that can be detected between a pair of images and the directionality implied by each manipulation. Visual examples of each are shown in Figure 4.

Scaling is the creation of a smaller, lower-resolution version of the image by decimating the larger image. In general, the smaller-scale image is assumed to be derived from the larger-scale image, as this usually results in preservation of image quality. **Cropping** is the creation of a new image out of a subsection of the original image. The image with the smaller crop area is assumed to have been derived from the image with the larger crop area. **Grayscale** is the removal of color from an image. We generally assume that the grayscale images are derived from color images. **Overlay** is the addition of text information or some segment of an external image on top of the original image. It is generally assumed that the image containing the overlay is derived from an image where the overlay is absent. **Insertion** is the process of inserting the image inside of another image. Typical examples might be creating an image with two distinct images placed side by side or by inserting the image in some border with additional external information. It is assumed that the image resulting from the insertion is derived from the other image. Of course, there are exceptions in the directions of each of these manipulations: it is possible, though not ideal, to scale images *up*, or an overlay could be *removed* with retouching software. Still, we assume the directions that we have specified are true in most cases. This list is also not exhaustive. There are other types of manipulations due to format changes (such as JPEG to Tiff), compression quality, and contrast enhancement. While we might build detectors for these manipulations, we instead focus on the operations that change the meaning of the image,

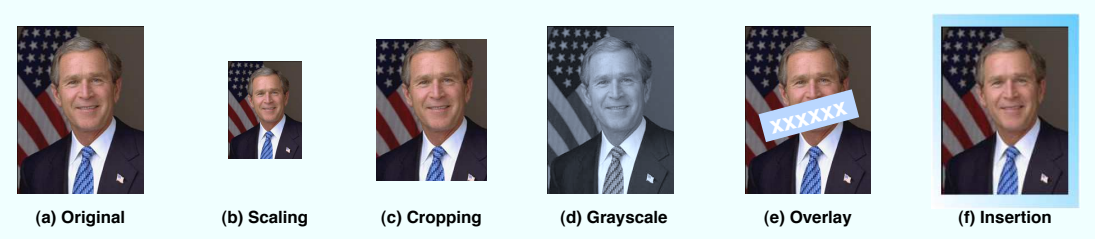

Figure 4: Example outcomes (b-f) resulting from possible manipulations on an original image (a).

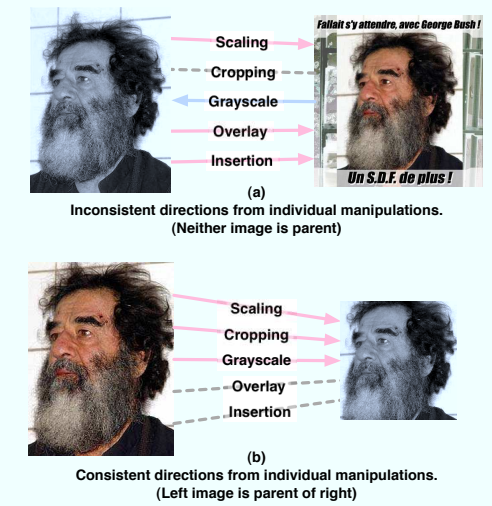

Figure 5: Examples of multiple directional manipulation detectors and their relative levels of consistency.

rather than just the perceptual quality. Later, we will show that this is sufficient for building compelling applications.

2.1.2 *Checking Manipulation Consistency*

If we have the directions of the five above-mentioned manipulations for images A and B, it remains for us to evaluate whether or not they make sense all together. Each method can give one of three possible results about the parent-child relationship between the two images: 1) the manipulation indicates A is the parent (or ancestor) of B, 2) the manipulation indicates that B is the parent (or ancestor) of A, or 3) the manipulation is not present, giving no information about the relationship. If these detection results all agree on the directionality, then it is plausible that there is a parent-child relationship present. If not, then most likely there is no such relationship. Also note a parent-child relationship does not assert one image is the immediate source from which the other one is derived. There could be intermediate generations of copies between the two.

In Figure 5, we show some examples of how image manipulations can either be consistent or inconsistent. At the top, in Figure 5a, we show a case where the detectors are giving conflicting stories. The scaling, overlay, and insertion cues indicate that the right image is the child, while the grayscale cues would suggest just the opposite. (Also, there appears to have been no cropping). The contradictory stories being told by the manipulation detectors indicates to us that neither image is derived from the other. It is more likely that each is derived from some other parental image, and the two are cousins or siblings in the manipulation history.

In Figure 5b, we see an example where the individual cues are in agreement. The scaling, grayscale, and cropping cues indicate that the right image is derived from the left one, while no insertion or overlay effects appear to be present. The cumulative effect of all of these detections (and their agreement) is that it is plausible that the left image is, indeed, the parent of the right one.

2.2 Contextual Manipulation Cues

The precise directionality of certain types of manipulations cannot be detected from a single pair of images, alone. Comparing Figures 4a and 4e, a human can easily tell that 4e contains an overlay. A machine, however, would only be able to discern that the two images differ in the region of the overlay, but would not necessarily be able to infer which image contains the original content and which one contains the overlay. By considering all of the images in Figure 4, an automated algorithm would see that most images have content similar to Figure 4a in the region of the overlay, which would imply that Figure 4e is the outlier, and is likely to have been the result of an overlay manipulation.

This context provided by a plurality of instances of the image is also needed to obtain information about the manipulation history. After we have detected each of the manipulation cues and evaluated their consistency, we are essentially left with a consensus-based decision about the existence (and direction) of parent-child relationships between pairs of images. We take each of the images to be nodes and form directed edges between nodes based on these detected parent-child relationships. The interpretation of this graph is that, where a directed edge exists between two image nodes, it is plausible that a series of manipulations resulted in one image being derived from the other.

2.3 Visual Migration Maps

Depending upon the nature of the original pool of images used to conduct this manipulation history detection, the emergent structure can be quite different. If the pool is diverse (perhaps drawn from web image search results), then we would expect to find several dfferent connected components of different (non-copied) images found within the pool. If the pool is rather homogeneous (perhaps a human manually provided a set of known copies), then we would expect to find a single connected component covering all of the images. Regardless of the structure of the pool at large, each individual connected component leads to a resulting VMM.

In general, there will be redundancy in the graph structure of each VMM. In practice, our detection approach will result in a structure like Figure 6a, since it is plausible for an image to have been derived from its parent or its parent's parent, there will be links formed between an image and each of its ancestors. In determining the actual VMM of an image,

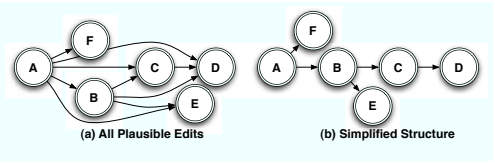

Figure 6: Simplification of redundancy in VMMs.

either path between two images is equally plausible. From a practical point of view, however, we are equally unsure of how truthful either path is. We simplify the structure by always choosing the longest path between two images, resulting in structure similar to 6. This is not necessarily better than any other graph simplification, but it is practical in that it retains important aspects, such as the sink and source nodes, and assumes that each image inherits the manipulation history of its parents.

Similarly, our automatic detectors may be faulty and result in cycles in the graph. We can handle these cycles by removing the offending edges based on some criteria (e.g., the confidence score in detecting each manipulation operation). This may lead us astray from the true VMM of the image, but the resulting structure will remain sufficient for image discovery applications in the IIAS framework. In our experiments, we do not find cycles in our automatic VMMs.

3. AUTOMATIC MANIPULATION DETECTION COMPONENTS

In this section, we will discuss automatic algorithms that we have used to implement the manipulation detection methods. A framework for the system is shown in Figure 7. We divide the detection methods into context-free detectors, where all we need is the two images and the manipulation detection can be done directly, and context-dependent detectors, where we need to gather information from other images to determine the exact nature of the edits occurring.

3.1 Context-Free Detectors

3.1.1 Copy Detection

The first step in the automatic system is to ascertain whether or not the two images are copies of each other, namely the two images are derived from the same source image through distinct manipulation operations. In principle, any generic image near-duplicate method can be applied here. In our current implementation, we adopt a simple but sufficiently effective method for copy detection (which we first applied to detected repeated views of landmarks in [5]) and focus on the evaluation of the novel concepts of IIAS and VMM. We begin by extracting scale invariant feature transform (SIFT) [8] descriptors for each image. These features capture local geometric properties around interest points within the image. SIFT descriptors are invariant against a number of distortions, such as scaling and rotation, and robust against a number of other transformations. They are highly distinctive and occurrences of descriptors of a real-world point represented across different images can be matched with very high precision. Each image has a set of SIFT descriptors, I_S, where each descriptor is a tuple of (x_i,y_i,$\mathbf{f}$), where x_i and y_i are the spatial X-Y coordinates of the interest point in the image, and $\mathbf{f}$ is the 128-dimensional SIFT feature vector describing the local geometric appearance surrounding the interest point. Given two images, A and B, we exhaustively search all pairwise point matches between the images. We detect a matching set when the euclidean distance between the points' features, $D(f_{A,i}; f_{B,j})$ falls below a given threshold. Matching points between A and B are then retained in a set, $\mathbb{M}_{A,B}$, which consists of a set of tuples, ($x_{A,i}$,$y_{A,i}$,$x_{B,j}$,$x_{B,j}$), marking the locations of the matching points in each image. We then apply a threshold on the number of matching points, $\mathbb{M}_{A,B}$, in order to get a binary copy detection result. In our experiments, we have set this threshold equal to fifty, since some of our initial observations have shown that this yields precise detection. (For comparison, each image in our data set contains between several hundred and a thousand interest points.)

3.1.2 Scaling

An important piece of information that emerges from the above-described copy detection approach is the set of matching points, $\mathbb{M}_{A,B}$. Assuming no image rotation is involved, the scaling factor between the two images, $SF_{A,B}$, can be estimated directly as the ratio in the spatial ranges of the X-Y locations of the matching points:

$$SF_{A,B} = \frac{max(x_A) - min(x_A)}{max(x_B) - min(x_B)} \quad (1)$$

The same estimate can be computed for the Y-dimension to account for disproportionate scaling. We apply a threshold to $SF_{A,B}$ for a binary detection of scaling. A more principled approach might be to apply random sample consensus (RANSAC) [2], which has been frequently used for image registration in computer vision and remote sensing. We can utilize the above estimation to normalize the scales and align the positions of two images. It implies that A is $SF_{A,B}$ times larger than B, so we can generate B', which is at the same scale as A', by scaling and interpolating its pixels by a factor of $SF_{A,B}$. In addition, simple shift operations can be performed to align the interest points (and corresponding pixel content) of B' with A. Such scale normalization and position alignment can then be later used to conduct pixel-level comparisons to detect other manipulation artifacts.

3.1.3 Color Removal

To implement the color removal detector, we start by estimating whether each image is grayscale. In the trivial case, the image is stored as a grayscale file, so we can see unambiguously that the image is contained in a single channel. This accounts for 50% of the grayscale images that we encounter. The others are grayscale, but are stored in regular three-channel (RBG, YUV, etc.) files. For these cases, we analyze the differences between the red, green, and blue channel values for each pixel. For grayscale images, we expect these differences to be zero. We calculate the mean over all of these channel differences and take images below a certain threshold to be grayscale. Once we know whether each of the two images are in color or in grayscale, we can then estimate the direction of the color removal edit.

3.2 Context-Dependent Detectors

The nature of certain types of manipulations cannot be detected directly from just a pair of images. Consider the case of two images: one is an original instance of the image and the other contains a portion of overlayed image content. Given just the two images, we could most likely compare

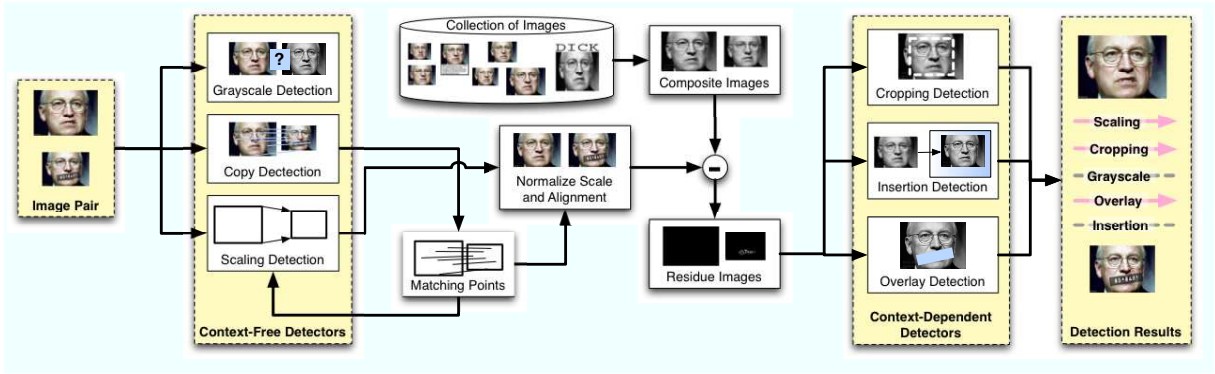

Figure 7: Proposed system architecture for automatically detecting various types of image manipulations. In the first stage, copy, scaling, and grayscale effects can be detected using only the two images. The scaling details can then be used to align the images against other available versions of the image to infer details about cropping, insertion, and overlays.

pixel intensities and discover that there is some area where the images have different content, but how can we know which image is the original and which one is the derived version? Consider the images in Figure 4. If we only had the images in 4a and 4b, all we would know is that one has a larger crop area than the other. But, if we only had images in 4a and 4f, we would reach the same conclusion. Given just two images, we can detect if one has a smaller crop area than the other, but what does that actually tell us? Perhaps the smaller image resulted from cropping the larger image, or maybe the larger image resulted from inserting the smaller image in a larger scene.

To address these problems, we look at image differences in the context of all of the other copies of the image that we have available. We consider the set of all images, $\mathbb{I}$, within a connected component obtained via analysis of the copy graph construction described in Section 2.2. Suppose that we would like to evaluate an image I_A. The image has a set of "neighbors," $\mathbb{I}_{A,N}$, which are the images that have been detected as copies of I_A. Within this set, we can use the method described in Section 3.1.2 to normalize the scales and offsets between each image in $I_{A,N}$ and I_A, yielding a scaled-shifted version of the images, $I_{A;N''}$, such that they can be composited on top of each other with pixel-wise correspondences. We can construct a composite image:

$$I_{A,comp} = \frac{\sum \mathbb{I}_{A,N''}}{|\mathbb{I}_{A,N}|} \tag{2}$$

where each pixel in the composite image, $I_{A,comp}$ is essentially the average of the values of the corresponding pixels in the images in the neighbor set $I_{A,N}$. This composite image gives us contextual information about the typical appearances of areas of the image across many different copies of the image. We can compare the content of I_A against the composite content in $I_{A,comp}$ to find regions of I_A that are atypical. We do this by finding the residue between the two:

$$I_{A,res} = |I_A - I_{A,comp}| \tag{3}$$

where the residue image, $I_{A,res}$, is the absolute value of the pixel-wise difference between the image and the composite of its neighbors. We apply a threshold to $I_{A,res}$ to binarize it. Now, if we wish to compare I_A against some other image, I_B, we can similarly produce composite and residue images $I_{B,comp}$ and $I_{B,res}$ and use the residue images $I_{A,res}$ and $I_{B,res}$ as proxies for evaluating the pair I_A and I_B.

In Figure 8, we see some examples of the appearances of the composite and residue images. The composite images sometimes still show traces of manipulations that are present in other images, but are largely true to the original content of the image. The key intuition behind the resulting residue images is that they are such that we expect that areas that are consistent with the original image will be black and areas that are inconsistent will be white. These residue images are then powerful tools that can be used to disambiguate the directions of overlay manipulations or to clarify the differences between crops and insertions. We will discuss the specifics of these manipulations in the following sections.

3.2.1 *Cropping and Insertion*

In Figure 8a, we see an example of how a crop manipulation would appear in terms of the composite and residue images that drive our system. We see from the example that the image content in the larger-crop-area image is consistent with the composite image, which is reflected in the darkness of the residue image. This is consistent with a cropping operation. In Figure 8b, on the other hand, we see an example of an insertion operation. Here, the content of larger-crop-area image is different from the composite image, which is reflected in the many white areas of the residue image. This is consistent with an insertion operation. In summary, candidates for cropping and insertion are discovered by finding image pairs with differences in image area. Cropping and insertion can then be disambiguated by examining the properties of the residue image in the out-of-crop region.

3.2.2 *Overlay*

In Figure 8c, we see an example of the composite and residue images that would be seen with an overlay. We see that the overlay image has a region that is highly different from the original, which is reflected in white pixels in the residue image. We also see the relationship between the overlay and insertion operations. They both exhibit image regions with high dissimilarity to the composite image. The areas of difference for overlays are inside the image crop area shared by both images, while these areas are outside the main image crop area in the case of insertion.

4. EXPERIMENTS AND ANALYSIS

We have applied the above-described automatic manipulation detectors to several sets of image copies. After each

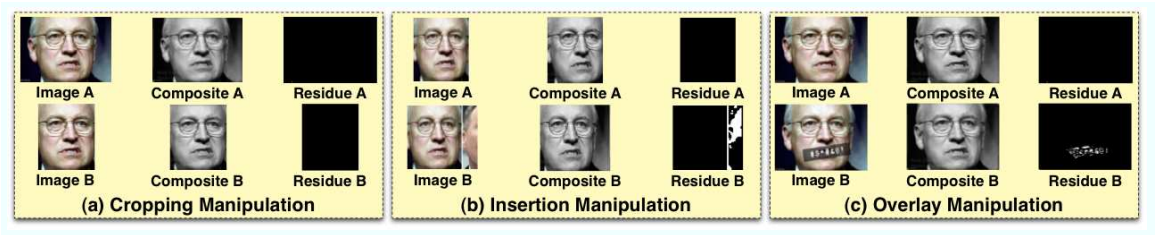

Figure 8: Examples of residue images in cropping, insertion, and overlay manipulations.

Figure 9: Images evaluated in our experiments. We use the proposed IIAS system to discover the plausible manipulation history for each of these iconic images on the Internet.

of the individual detectors have returned their results, we can use the consistency checking approaches and the graph construction techniques discussed in Sections 2.1.2 and 2.3 to use these automatically-detected cues to construct VMMs for each set of images and deliver summaries of their contents to end users. We evaluate our method in the context of web image search by taking the top results returned to the user from a web image search engine as our pool of images and extracting the manipulation histories of various highly-reused images contained within the set. Here, we will describe the queries that we have processed and the characteristics of the resulting data, along with the ground-truth manual annotations that we have generated about the manipulation operations associated with the image data. We will also discuss our intrinsic evaluations of the quality of the automatic manipulation detectors (from Section 3) and the VMMs that result by explicitly comparing against manually-generated manipulation labels and VMMs. Later, in Section 5, we will further evaluate the larger image archaeology system, extrinsically, in terms of its utility for applications in discovering "interesting" images from within the result pool.

4.1 Experimental Data

We evaluate the system against images from the web. To gather these images, we query the web with a set of queries culled from a variety of sources. Among these sources are the queries in the Google image search Zeitgeist[1], which lists popular queries entered into the engine in recent history, and the query topics for named persons used over the past seven years in the TRECVID video search evaluation[2]. In the end, we arrive at a list of approximately 100 image search queries, spanning a range of categories including named persons (such as politicians, public figures, and celebrities), locations (such as specific cities and tourist destinations), events in the news, films, and artists. From these queries, we manually generated a set of keywords that would be expected to return relevant photographs from a web image search. We then fed these keywords into the Yahoo! web image search engine and collected the top-1000 returned images (the maximum number available). Across this set of 1000 images, we then applied a copy detection approach (previously described in Section 3.1.1) to find all copy pairs within these sets. We form edges between images to construct a copy graph, which typically consists of many different connected components. For each result set, we then take the largest connected component (i.e. the most-copied image) as the set of images to be fed into our manipulation detection and VMM construction algorithms. But, first, we filter down our connected components to only those which contain interesting manipulation patterns. Most classes of queries, such as locations, films, and artists, do not exhibit perspective-changing manipulations. Some classes, such as political figures and celebrities do contain such manipulations. We do this filtering process manually, by visually skimming the contents of the connected components. This process might be automated by detecting the presence of manipulations using some adaptations of the methods that we have discussed. In the end, we evaluate the IIAS system against 22 unique queries, shown in Figure 9. For each query, the largest connected component (which we will process) typically contains several dozen copies of the image.

A single human annotator provides ground-truth labels for the manipulations that we wish to detect: copy, scaling, cropping, insertion, overlay, and grayscale. The annotator inspects each pair of images and individually labels whether any of these manipulations are present between the pair. If the manipulation is present, the annotator also labels the directionality of the manipulation (i.e. which image is implied to be derived from the other). Many of these manipulations can be very simple to observe visually. For example, a grayscale image is completely obvious. Overlaying external content and insertion within other images also tend to be quite observable. The degree to which other manipulations can be observed can be subject to the magnitude of the manipulation. In scaling, if one image is decimated by 50% compared to the other, then it should be obvious. A 1% relative scaling would be harder to accurately notice, however. So, as with any human-generated annotation, this data is subject to errors, but we contend that it is still helpful for comparing our automatic approach against manually-generated approaches. Given the individual pairwise manipulation labels, we can then apply the consistency-checking approach from Section 2.1.2 to form parent-child

[1]http://www.google.com/press/zeitgeist.html
[2]http://www-nlpir.nist.gov/projects/trecvid/

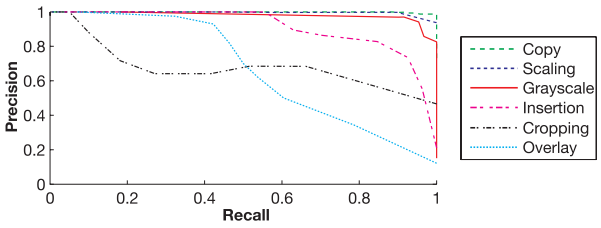

Figure 10: Performance of the manipulation detectors.

links between images, based on manual (instead of automatic) labels. These links across a set of image copies form a manually-generated VMM, against which we can compare our automatically-generated VMM.

The human annotator also annotates two properties of each individual image: its manipulation status and the viewpoint that it conveys. The first property, the manipulation status, simply reflects whether the image is one of the types of images shown in Figure 2 ("original" or "highly-manipulated"). These annotations are gathered by having the annotator scan all of the images within a connected component of images to gather an intuition about the appearance of the original image crop area and content. These classes are largely easy to observe and the annotations are quite reliable. The second property, the viewpoint conveyed by the image, is more subjective and we rely on the content of the original HTML page that referred to the image. We examine these pages and evaluate the viewpoint of the document as either positive (supportive), neutral, or negative (critical) of the subject of the image.

4.2 Image Manipulation Detector Performance

We evaluate the core image manipulation detectors by comparing their results against the ground-truth labels given by the human annotator. We evaluate in terms of precision and recall. Precision is defined as the percentage of the automatically detected manipulations returned by our system that are manually labeled as true manipulations in our ground-truth. Recall is defined as the percentage of manually-labeled ground-truth manipulations that are successfully detected by our automatic system. Each of the methods relies on some sort of threshold to make a binary decision. Examples of these thresholds might be the absolute magnitude of the detected manipulation, such as the percentage by which a scaling edit decreased the size of an image or the percentage of the image that is occupied by detected overlay pixels. We scan over different threshold levels and observe the relative shifts in precision and recall.

We see precision-recall curves for each of the detectors in Figure 10. All of the basic, context-free detectors (copy detection, scaling, and color removal) have nearly perfect performance, each is able to exceed a precision in the range of 95% with recall in the range of 90%. The context-dependent detectors still leave some room for improvement, however. The most successful among these detectors is the insertion detection method, which retains moderately high precision through most recall ranges. The overlay detection method provides near-perfect precision up to a certain recall level and then falls off precipitously. We find that the size and color contrast of an overlay is causing this effect: given overlays that are large enough and different enough from the original image, then the method performs well. Smaller, less-perceptible overlays still remain as a challenge. Cropping provides fair precision throughout all recall levels. Further observation of the errors typically leads us to mistrust the quality of our manual annotation of cropping effects. In many cases, where the crop is only a few pixels, close inspection would reveal that the machine was correct and the human was in error, so the computed precision-recall value may not reflect the strength of the detector.

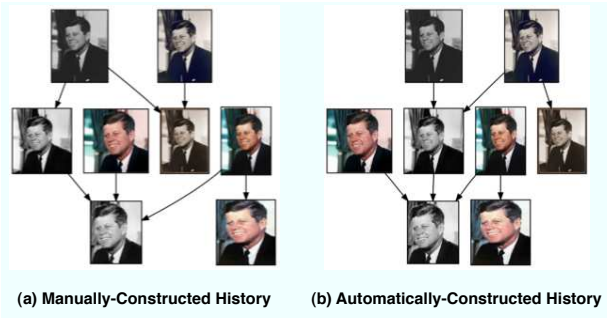

Figure 11: Comparison of automatically-produced and manually-produced VMMs. Note the significant agreement.

In our experiments, on an Intel Xeon 3.0 GHz machine, with our methods implemented in Matlab, it takes 2 hours to conduct copy detection and 15 minutes to compute the VMM. The speed might be increased by using recent fast copy detection approaches and optimized implementations.

4.3 Constructed Migration Maps

How useful are these manipulation detectors for constructing visual migration maps? To evaluate this, we construct VMMs using two methods: one using the manipulation labels that we have collected by manual annotation and another using manipulation labels that are automatically computed using the results coming from our detectors. In Figure 11, we can see an example comparison between the resulting VMMs. A careful comparison between the two graphs reveals that there is a great deal of agreement. Across all of the various image sets, we compare the automatic VMMs against the manual VMMs. Specifically, we do this by evaluating the pairwise relationship between each pair of images. Given a pair of images, we want to detect if a parent-child edge should exist between the two. We take the manually-determined edges as ground truth and use them to evaluate our automatically-determined ones. We take the correct detection of an edge between an image pair as a true positive and the incorrect detection as a false alarm and evaluate in terms of precision and recall. In our experiments, precision is 92% and recall is 71%, on average. The errors in these automatic VMMs are, intuitively, the result of errors in the detectors being propagated into errors in the detection of edges. It is also interesting to note that manual labels of image manipulation relations may not be completely reliable. Sometimes, automatic detection may be more accurate than human judgments in detecting subtle changes that cannot be easily perceived by humans (such as cropping of only one or two lines of pixels at the image boundary).

5. APPLICATION SCENARIOS

The resulting VMMs emerging from this analysis can give us a great deal of information about the qualities of the individual images, which can be used in an Internet Im-

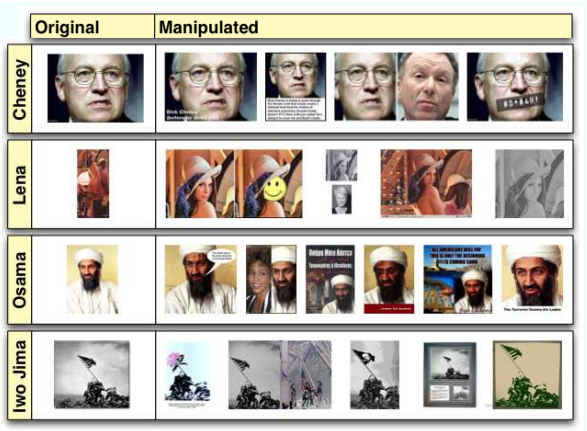

Figure 12: Examples of automatically discovered "original" and "manipulated" summaries for several images in our set.

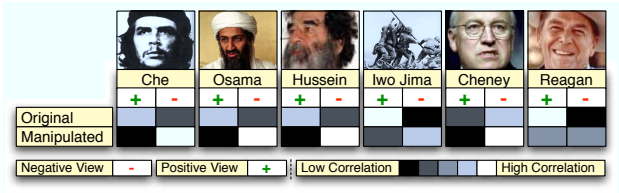

Figure 13: Observed correlations between image types and document perspectives.

age Archaeology System to help navigate and summarize the contents of a pool of related images in search or exploration tasks. Most simply, the "original"-type images that we are seeking will be the ones corresponding to source nodes (those with no incoming edges) in the graph structure, while the "highly-manipulated"-type images that we are seeking will be the ones corresponding to the sink nodes (those with no outgoing edges). As we have stated earlier, the types of "highly-manipulated" images that we are most interested in are the ones whose histories include a relatively large amount of information addition, which leads to changes in meaning and context. We can disambiguate between these "information-added" types of images and the less-desirable "information-subtracted" types by tracing the history from a source node, evaluating the types of manipulations experienced, and determining the relative number of additive versus subtractive operations that have taken place. Given these tools for analyzing the history of the images in the collection, the exact mechanisms for feeding the results back to the users can be left to be adapted for the specific tasks at hand. In a search task, where authentic relevant images might be preferred, perhaps the "original"-type images will be most useful to the user and other copies can be suppressed and removed from the results. In an exploration task, where the user may be interested in exploring the different perspectives surrounding a person or issue, the "interesting highly-manipulated" types might be most useful for comparative purposes. These considerations can be left for specific system designs, but they do rely on the ability of the image manipulation history to surface these various kinds of images. In Figure 12, we show examples of automatically discovered "original" and "manipulated" summaries which are indeed quite accurate.

5.1 Implications Toward Perspective

A key claim in this work is that the manipulations conducted against a particular instance of an image can change the image's meaning and reflect the opinion being conveyed by the author. An Internet Image Archaeology System might enable users to browse varying perspectives within sets of images. In Figure 13, we present a summary of the correlations between image types ("original" or "manipulated") and the viewpoints represented by the web pages upon which they appear. Here, we boil the viewpoints down to simply "positive" (for cases in which the author takes a position specifically in favor of image subject, or the author is neutral and takes no position) or "negative" (for cases in which the author takes a position specifically opposed to the image subject). Through many of the examples, including "Che," "Osama," "Hussein," and "Iwo Jima," we can see that there is, indeed, a correlation between the type of image and the viewpoint of the document. In these cases, the original-version images are highly associated with positive and neutral documents, while the manipulated images are associated with negative documents. This lends some credence to our assertion that the status of the manipulation history of an image can indicate the meaning or perspective that it conveys. With some other images, manipulations are not observed to change the meaning as much. For the "Cheney" image, the original version is already unflattering and is frequently used as-is to convey negative viewpoints. Though, in some cases, it is manipulated, and the resulting images are still associated with negative viewpoints. The "Reagan" image is rarely manipulated in our set, so there is little chance to discover such a correlation.

6. RELATED WORK

To the best of our knowledge, this is the first work that has dealt with the issue of automatically detecting the manipulation history of an image and utilizing this information to discover important instances of images and perspectives on the internet. There are, however, several works in related fields upon which this work draws some influence.

The challenge of identifying whether or not two images are copies of each other has been addressed in recent research efforts in image near-duplicate detection [9]. As the term "near" would imply, these methods aim to detect pairs where the duplication may not be exact. As a necessity, these methods must be robust against a variety of distortions, such as cropping, scale, and overlay. A similar problem is video copy detection [3], which aims to find repeated occurrences of the same video clip in various streams and locations, without requiring the exibility of near-duplicate detection. This is necessary for applications, such as copyright protection on video sharing web services. In our work, we do not seek to re-invent near-duplicate or copy detection. Instead, we stand on top of existing work and incorporate it as a component in our proposed system. We extend beyond copy detection by turning attention specifically towards the ways in which two duplicate images differ and detect the manipulations that the image has been subjected to.

In the field of image forensics, the objective is typically to take a single image and identify whether or not any manipulations have taken place [4, 1]. Approaches in this field may involve checking the consistency of various regions of the image for artifacts induced by the physics or the pecu-

liarities of the acquisition and compression processes. Such cues can be used to identify the camera used to take a photograph or if two regions of a photograph are derived from separate sources. Our work differs in that we do not consider single images, instead we evaluate the manipulations of images in the context of a plurality of various instances of the same image, which makes the task easier. We further aim to not only detect the presence of a manipulation, but to also characterize the types and history of manipulations and use that information to enhance browsing and search.

One of the goals of our work is to extract cues about the message conveyed by the photograph and how that may have been subverted through manipulation. Some works have looked at the messages and perspectives inherent in multimedia documents in a more general sense. In [6], the authors have investigated the the differences between documents dealing with the same subject from different ideological standpoints. In text documents, it is shown that documents reflecting different sides of an issue have divergent distributions of divisive keywords. Similarly, in visual documents, the use of differing types of imagery (such as tanks and explosions versus peaceful scenes) can express differing viewpoints [7]. These effects, of course, are exhibited by completely separate documents. In this work, we examine how ideological differences are expressed through manipulation and re-distribution of the same original document.

7. CONCLUSIONS AND FUTURE WORK

We have proposed an Internet image archaeology system, which can be used to process a given set of related images and find interesting instances of images or perspectives. A key component of this IIAS framework is the visual manipulation map, which assumes that images on the web are frequently copied, manipulated, and re-used, and that this behavior can lead to a plurality of instances of the image across many sources. The VMM can acquire knowledge about the shared history among these photographs and the lineage of each instance, leading to intelligent approaches for browsing and summarizing the collection of copied photos. In particular, we believe that specific instances of images (such as the one closest to the original photograph or the versions with the most added external information) will be of the most interest to users. To find these images, we aim to detect parent-child derivation relationships between pairs of images and then construct a plausible VMM.

We have suggested a novel approach of considering pairs of near-duplicate images in the context of the plausible combinations of manipulations that could have resulted in one image being derived from the other. We propose that many manipulation operations performed on images are directional: they either remove information from the image or inject external information into the image. So, there are clues about the parent-child relationship between images encoded in the image content. We decompose the parent-child relationship detection problem into the task of individually detecting each type of manipulation and its directionality. Given these results, we can then check whether the directions of manipulations are in agreement or not. We show that the directionality of many types of editing operations can be detected automatically and that we can construct plausible visual migration maps for images and use these cues to conduct archaeological digs against internet image collections to discover important images from within the set.

We have observed that, out of 100 queries submitted to a image search engine, only 22 queries (about 20%) returned images exhibiting interesting visual migration maps. Most of these interesting images are related to political figures. It is still unclear how frequently users will find images with interesting visual manipulation maps. In general usage cases, do 20% (or a different percentage) of image queries exhibit interesting patterns? Future work might explore the reach of this work by leveraging collections of real queries from actual users to understand how well the approach generalizes.

Finally, we also note that one run of the proposed IIAS framework only presents a snapshot of the evolution of the meanings and re-use patterns in image collections. So, by running the system periodically over time and tracking the emerging structures, we may uncover temporal aspects of visual migration maps and how they grow over time. Similarly, image manipulations may spread spatially over different geographical and cultural locations or topologically across internet-based sub-cultures. In future work, we might take the initial results that we have obtained so far for the initial 100 queries as a seed for capturing the visual migration map maps on the Web over time and space. Utilizing the content of the web pages encompassing the images within the migration map, we may extract text patterns or hyperlink structures to further probe the web and expand the utility of the novel visual migration map structures.

8. ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0716203 and the US Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and the US Government.

9. REFERENCES

[1] H. Farid. Detecting Digital Forgeries Using Bispectral Analysis. Technical Report AIM-1657, MIT, 1999.

[2] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[3] A. Hampapur, K. Hyun, and R. Bolle. Comparison of sequence matching techniques for video copy detection. In *Conference on Storage and Retrieval for Media Databases*, pages 194–201, 2002.

[4] J. He, Z. Lin, L. Wang, and X. Tang. Detecting Doctored JPEG Images Via DCT Coefficient Analysis. *European Conference on Computer Vision*, 2006.

[5] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *World Wide Web*, 2008.

[6] W.-H. Lin and A. Hauptmann. Are These Documents Written from Different Perspectives? A Test of Different Perspectives Based On Statistical Distribution Divergence. In *Proceedings of the International Conference on Computational Linguistics*, 2006.

[7] W.-H. Lin and A. Hauptmann. Do these news videos portray a news event from different ideological perspectives? In *International Conference on Semantic Computing*, 2008.

[8] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] D.-Q. Zhang and S.-F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM Multimedia*, 2004.