Automatic Rights Management for Photocopiers

Andreas Girgensohn
FX Palo Alto Laboratory
Palo Alto, CA, USA
andreasg@fxpal.com

Lynn Wilcox
FX Palo Alto Laboratory
Palo Alto, CA, USA
wilcox@fxpal.com

Qiong Liu
FX Palo Alto Laboratory
Palo Alto, CA, USA
liu@fxpal.com

ABSTRACT

We introduce a system to automatically manage photocopies made from copyrighted printed materials. The system monitors photocopiers to detect the copying of pages from copyrighted publications. Such activity is tallied for billing purposes. Access rights to the materials can be verified to prevent printing. Digital images of the copied pages are checked against a database of copyrighted pages. To preserve the privacy of the copying of non-copyright materials, only digital fingerprints are submitted to the image matching service. A problem with such systems is creation of the database of copyright pages. To facilitate this, our system maintains statistics of clusters of similar unknown page images along with copy sequence. Once such a cluster has grown to a sufficient size, a human inspector can determine whether those page sequences are copyrighted. The system has been tested with 100,000s of pages from conference proceedings and with millions of randomly generated pages. Retrieval accuracy has been around 99% even with copies of copies or double-page copies.

CCS CONCEPTS

• **Security and privacy** → **Digital rights management**; • **Information systems** → *Image search*; Content analysis and feature selection;

KEYWORDS

Copyright detection, image matching, near duplicate image detection, document clustering, rights management

ACM Reference Format:
Andreas Girgensohn, Lynn Wilcox, and Qiong Liu. 2018. Automatic Rights Management for Photocopiers. In *DocEng '18: ACM Symposium on Document Engineering 2018, August 28–31, 2018, Halifax, NS, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209280.3209531

1 INTRODUCTION

Copyright owners have a need to keep track of printed copies of their material so that they can be compensated appropriately. This is particularly true in countries without a fair use doctrine, where copies of even relatively short portions of the copyright material are subject to fees. In some countries, an intermediary such as a copyright agency may facilitate the tracking of such copying and the compensation of the rights holders. Many such agencies rely on manual record keeping of all copying activity based on an honor system.

In this paper, we describe a system that automates the process of monitoring copiers, matching copied images to materials owned by rights holders, and recording such copying actions. Our system can also prevent copying if there is no permission. The system monitors copiers to receive information such as the copied images, the copy count, and the identity of the person making the copies. The latter would come from security settings where an electronic badge or a pin code is required to use the copier. The images are processed to facilitate accurate matching against near-duplicate images in a databases representing pages of printed materials owned by rights holders. Matches in the database may be recorded together with the data about the copying process such as the identifier of the copier, the copying person, the number of copies, and the matched materials. The recorded information is used for billing purposes.

A major difficulty in such an automated system is maintaining a continually growing database of copyright material. Rather than trying to do this manually, current copyright agencies build this by adding new material over time as it becomes available. To automate this process, we make the assumption that copyright material of significant monetary interest will be copied multiple times at multiple places. In our system, we store information about copied pages that are not identified as copyright material on a central server. These pages are clustered based on similarity and copy sequence to allow multiple pages from the same book or magazine to be grouped together. Once clusters exceed a certain size, they are sent to the agency where a person determines their copyright status. If copyright is verified, this material is added to the database.

This work provides two main novel contributions. First, we use image retrieval techniques [11] to accurately and quickly match reproduced printed pages against a collection of millions of pages from copyrighted works. This includes our novel approach for preventing matches of partial pages. Second, we support gradually building up a database of copyrighted material by observing copy behavior. This addresses the issue that copyright holders do not have an up-to-date high-quality image of every page of their copyrighted works.

In the next section, we list the requirements that guided the design of our system. In the following section, we put our work in the context of related work. We then describe the steps of matching copyrighted pages and our approach for incrementally discovering copyrighted pages. In conclusion, we situate our work and provide a glimpse of our plans for the future.

DocEng '18, August 28–31, 2018, Halifax, NS, Canada

ACM ISBN 978-1-4503-5769-2/18/08…$15.00
https://doi.org/10.1145/3209280.3209531

2 REQUIREMENTS

This system started with the question of whether it would be possible to automate rights management for photocopiers. The system is supposed to replace a manual process where a form had to be filled out for every copy of copyrighted material, e.g., pages from a text book. That form together with an extra copy is sent to a management agency that distributed fees to copyright holders. Compliance is a concern due to this cumbersome process with no external controls. This manual process currently takes place in schools that make copies of pages from text books for the students. The education department pays copyright holders based on the estimated copy volume determined by a small sample of the schools.

A replacement system has to be highly accurate to not miss any copyrighted material and, even more importantly, not to misidentify unrelated copies. Such a system should neither interfere with the normal copy workflow nor require users to perform additional steps. The system also needs to address privacy concerns for the copying of materials that are not copyrighted such as invoices. That means that images of non-copyrighted material should not be visible to agents of the copyright holders.

The system has to be able to handle all copyrighted material that could be encountered from all copyright holders included in this process. While this would be restricted to materials one could reasonably expect to be copied, e.g., text books, one would still be able to deal with a collection consisting of millions of copyrighted pages. The system needs to process the expected copy volume with reasonable computational resources.

An initially unexpected requirement is the gradual buildup of the database of copyrighted pages. Copyright holders do not always have digital images of every page of their works available at sufficient quality. Thus, there needs to be the ability to monitor all copying to discover copyrighted pages previously not known to the system. This has to be done in a way that preserves the privacy of other pages being copied.

For the future, there are additional requirements to apply the system to other uses. The most obvious extension to extend it to monitor scanning at the same multifunction copiers, too. That would not change the technology but only the way how such usage should be billed. Instead of copyrighted material, one could monitor restricted material and prevent or report printing or copying by unauthorized personnel. This could include monitoring the copying of banknotes from countries that do not use markers that prevent copying by simpler means.

3 RELATED WORK

There are two basic techniques for detecting copyright material: watermarking and fingerprinting. In watermarking systems, additional information is added to the original material. For example, nearly invisible patterns of bits can be added to a digital image that can be detected through special analysis [2]. Another example is the EURion constellation[1] that is embedded in banknotes that can be detected and used to prevent counterfeiting. Singh's dual watermarking scheme integrates copyright protection, tamper detection and recovery into one scheme [24]. A DCT/DWT based approach with Arnold Cat Map encryption was reported to be secure as well as free from the false positive detection problem [23].

On the other hand, fingerprinting systems detect copyright based on content without the need to add additional information. For material containing textual content, an approach might be to OCR the photocopied image and then do text-based analysis [3, 20, 29]. However OCR is slower than other image analysis techniques, is not language-independent, and does not work for pages that mostly contain images such as drawing or photos. In this work we need to cover text as well as image data, so we restrict ourselves to images of the copied material. These systems are based on techniques known as near-duplicate image detection.

There are many systems available for near-duplicate image matching on large scale data sets. Applications include image similarity search to find multiple images related to a query image [11], logo recognition in images to identify when a logo is present in an image [6], visual search to link digital content to an image [13], copyright detection to identify when an image is copyright protected [12], document classification by visual appearance [1], and similar document retrieval using fusion of CNN features [26]. Such systems typically compute visual features for the image and match these features against a database of available images, determining an affine transformation or a homography to ensure the visual features in the source image occur in the same geometric pattern as those in the database image.

SIFT features [15] are rotation and scale invariant image features commonly used in near duplicate search. Variations of SIFT features including SURF [4], BRIEF [5], and ORB [21] provide feature extraction computational advantages. Improved min-Hash [7] and multi-index hashing [18] provide speed advantages over document comparison. In our system, high accuracy is more important than speed. Thus, we use FIT features [14] which are a lower dimensional version of SIFT that provides the same match accuracy with fewer system resources.

Although many systems try and match images by searching for images in the database with similar features using techniques such as an ANN tree, an alternate is to use a bag of visual words approach that much better scales to a database with billions of image keypoints. Here, image features are quantized into visual words and images with similar features can be found using an inverted index [17, 19, 25]. Our system also quantizes features into visual words.

For geometric verification, the RANSAC algorithm is commonly used [10], although many simplifications are possible [8, 30]. We use OpenCV's[2] RANSAC estimation of a partial affine transformation limited to combinations of translation, rotation, and uniform scaling to achieve sufficient speed.

There are several systems [9, 16, 22, 27] that track the reproduction of documents using photopcopiers or other means. Some of those systems only focus on parts of a printed page to locate information in headers and footers. Others are concerned with audio data and not with printed documents. We are not aware of any other systems that perform full-page image retrieval to detect copying of printed pages. While [12] also detects copyrighted images, those are photos and paintings and not printed pages.

[1] https://en.wikipedia.org/wiki/EURion_constellation

[2] https://opencv.org/

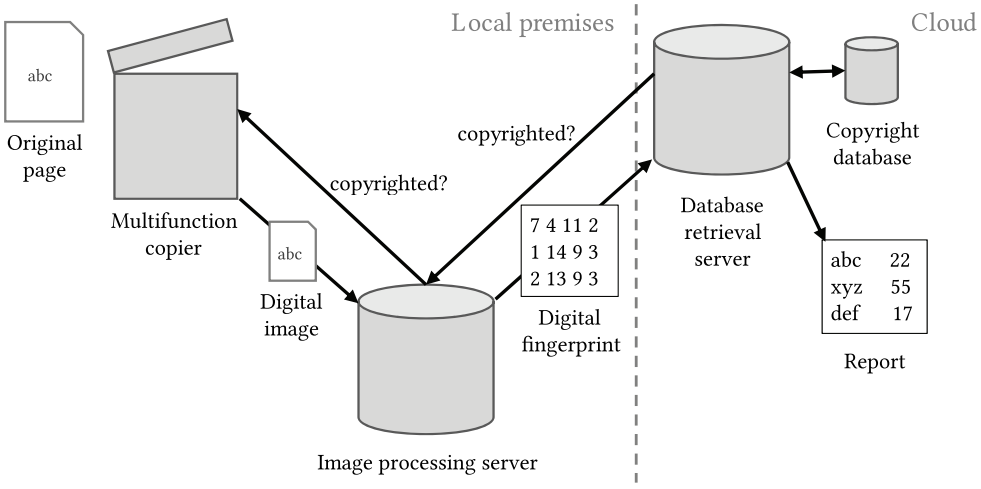

Figure 1: Flow from copier to copyright database.

4 MATCHING COPYRIGHTED PAGES

Our system uses a multi-stage pipeline to quickly and accurately match copied pages against a database of copyrighted pages. Figure 1 shows the flow from the copier to the database of copyrighted pages. The system consists of several components. A monitoring element communicates with a copier to receive information such as the copied images, the copy count, and the identity of the person making the copies. Being able to access digital images of the copied pages is a requirement for our system. Many multifunction copiers offer security settings where an electronic badge or a pin code is required to use the copier. Such authentication information can be included in the process of the system, providing a more detailed record of who makes copies, either for accounting or prevention.

When copying material such as books and magazines, the copied image is close to the original but may have minor imperfections due to the resolution of the copier, rotation, and smudges. When copying two pages of a book, bending the spine may cause distortions. While normally the whole page would be matched, it may be desirable just to look at parts of a page to locate the use of logos, stamps, etc.

The images are processed to facilitate matching against near-duplicate images in a database representing pages of printed materials owned by rights holders. We use a distributed server architecture for privacy and performance reasons. Visual features are detected in copied images at an image processing server located on the same premises as the copiers. The visual features act as digital fingerprints of the images. Those fingerprints together with metadata are sent to a central server to search for matches in a database of copyrighted pages (see Figure 1). Image feature detection and image search use well-known techniques from image retrieval that have been optimized for accuracy, scale, and response time. A novel technique is the prevention of matches of partial pages.

Matches in the database are recorded together with the data about the copying process such as the identifier of the copier, the copying person, the number of copies, and the matched materials. Matches are also reported back to the local server in case the printing of the copies should be prevented.

With the exception of cases where copying is prevented, this process is invisible to the person making the copies and does not interfere with their work flow. Thus, one can expect higher reporting accuracy compared to a process where people making copiers of copyrighted works have to manually log that activity.

Business rules determine what steps to take if a copied image is contained in the database of copyrighted pages. Some rules would collect the information for the copied image such as the authentication of the person making the copies and the information about the page in the database such as its source and owner. Such rules may just record the action for later accounting or send a notification to the copyright owner. Other rules would check the authentication of the person, determine a license status, and possibly prevent the printing of the copied page. Rules would determine how long to retain the digital images of the copied pages.

4.1 Getting Digital Images of Copied Pages

When used as scanners, multifunction copiers turn papers into digital images that are saved or emailed somewhere. Internally, copying pages is just a combination of scanning and printing. Unfortunately, not all copiers provide the option to temporarily save digital images of the pages being copied, an essential feature for our system. Enough copiers provide this feature so that our system can be adapted to the respective API.

Copiers offer a logging API that provides access to information such as the time and number of copies. The log may also contain the authorized user if a code or a badge was required for copying. The log points to the file name of the saved image.

The local image processing server checks every few minutes if there are new images and processes them. Through a modular architecture, the system accesses such information for different makes and models of multifunction copiers or other imaging devices. Alternatively, it may be possible for the multifunction copier to initiate a network connection to a server to upload such information to the server. The server may also be embedded in the multifunction copier.

If printing is supposed to be prevented, additional capabilities are needed for the copier and a tighter coupling to the local image processing server. Printing would be delayed for a few seconds. If multiple copies of the same page are made, this delay would be barely noticeable.

4.2 Image Processing

An image processing server may process the images from several multifunction copiers. It may contact the servers periodically to check for new images. If supported, the copier could instead initiate the connection whenever new images are available.

The image processing server has to compute a digital fingerprint of each digital image. Feature descriptors produced by algorithms such as SIFT [15] work well for that purpose because they can be used to accurately match near-duplicate images produced by a photocopier. We use a variant of SIFT called FIT (Fast Invariant Transform). Figure 2 illustrates the construction of FIT descriptors.

In FIT construction, we identify descriptor sampling points based on each keypoint location in the Gaussian pyramid space. We use 5 scale-dependent 3D vectors that rotate with the key point direction to identify sampling points for the key point. Then, we compute 8 scale-dependent gradients at each sampling point. If a gradient is less than 0, the gradient will be set to 0. After all of the above operations, we concatenate gradients from all sampling points of a key point to form a 40 dimensional vector as a feature descriptor.

Unlike SIFT which calculates histograms of Gaussian weighted gradients at a key point level, the FIT descriptor directly gets its features at multiple scales higher than the key point scale. This approach can greatly reduce the number of image-pixel-operations involved in feature extraction. Moreover, FIT uses the pre-computed pyramid to save computational cost on the expensive Gaussian weighting process [14].

The computation of FIT is much simpler and provides speed and accuracy advantages over SIFT [14]. The FIT keypoint detection and comparison algorithm can process a 150-dpi page-size image in a few seconds even on a very slow computer. A Raspberry Pi is sufficient to manage the output of several copiers. FIT features may be further simplified with binary features. However, the speed was sufficient so that there was no need to go further with a binary feature representation such as ORB [21].

The algorithm detects keypoints in the image at multiple scales. The systems caps the number of detected keypoints to 2,500 by dropping keypoints at the finest scales. That is sufficient for high match accuracy. In fact, accuracy is better compared to using 10,000 keypoints that are usually found when processing an image of a printed page at that scale. Interestingly, the customary doubling of the image dimensions before processing still increases accuracy even though the extra keypoints are removed. Both the match speed and the memory footprint during matching are proportional to the number of keypoints so that the use of fewer keypoints directly benefits performance.

The image processing server is located at the same location as one or more multifunction copiers. The digital images do not leave the premises, conserving network bandwidth. The central server only receives the keypoint descriptors and their coordinates in the image. As the image processing server does not perform matching

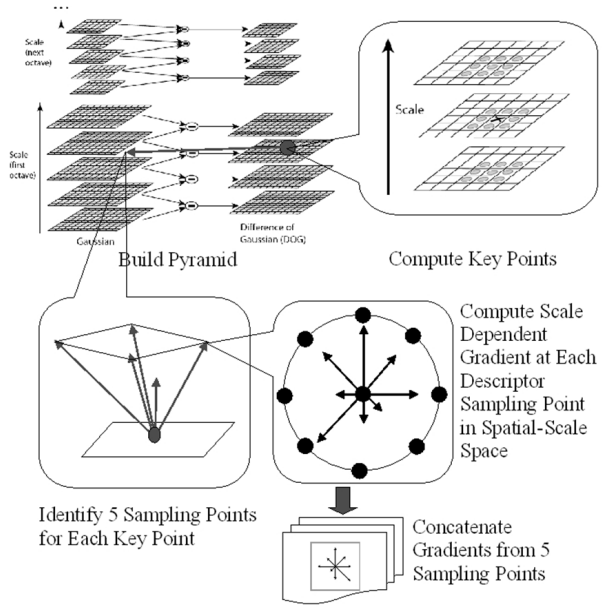

Figure 2: The construction of a FIT descriptor.

against the large database of copyrighted page images, memory and CPU requirements for the image processing server are minimal.

The main reason for this distributed architecture is to prevent private, non-copyrighted images from being made visible to rights holders or their agents. It has been demonstrated that images can be reconstructed from SIFT features [28]. This would circumvent our privacy protection. However, orientation and scale information is not included in the submission to the central server so that it is not possible to recover even an approximation of the image from the transferred data.

4.3 Image Matching

To greatly speed up the matching of keypoints, we quantize feature vectors into visual words. That means that each vector is mapped to a number such that similar vectors are assigned the same number. Those numbers can be compared more easily and mapped to images in the database with an inverted index.

Visual words are computed with a large representative set of training images where all feature vectors are clustered with a hierarchical k-means clustering [17] with 10 clusters per level for a total of 100,000 visual words. Once the clusters are computed, the vectors of the cluster centers at each level are stored.

To map a feature vector to a visual word, at each level of the hierarchy the centers of the clusters at that level have to be checked to determine the cluster that should be visited at the next level before the visual word is found at the leaf node. This hierarchical approach greatly speeds up lookup because the number of comparisons is reduced from 100,000 to 50 with out configuration.

Training images may need to be subsampled to fit the feature vectors of all images into memory. With 40-byte feature vectors and 2,500 keypoints per image, each megabyte of memory can hold the feature vectors of 10 images, making training with more than

one million images very resource-intensive. Training only has to be repeated after the visual characteristics of the image collection change significantly, for example, when adding content in a new script such as Arabic.

Vocabulary training is one of the areas that required optimizations to minimize memory usage and to maximize parallelism. We conducted our experiments on a server with two Intel Xeon CPUs with 14 cores each and 384 GB RAM. Before the optimizations, a test collection of 20 million page images could not be processed on the server because of insufficient memory. Smaller collections took very long to process. One optimization in the hierarchical clustering was to sort the feature vectors by cluster membership after each step such that the next level in the hierarchy could cluster the members of one cluster without having to allocate a new copy of the vectors. Vectors still had to be copied but only within the already allocated memory. All feature vectors should be concatenated to avoid the overhead of separate headers that would require 60% extra memory for 40-byte vectors. Another optimization made sure that vectors could be checked in parallel against cluster centers and that updating cluster centers could use all CPU resources even when CPU cores outnumbered clusters.

We use a bag-of-words approach using visual words from quantization for matching against the database of page images. This approach counts how many visual words the query image shares with each of the database images. An inverted index pointing from each visual word to the list of images containing it facilitates fast lookup. The fraction of visual words in each image that match visual words in the query image is an indication of the quality of the match. This query returns a list of pairs of page identifiers and match fractions that are sorted by descending match count. We found that a threshold of 10% works well to select promising candidates. In addition, candidates are restricted to the top-N match fractions with N of the order of 100.

For making use of many CPU cores, access to the inverted index needs to be parallelized, e.g., by assigning different page identifier ranges to different CPUs. If the inverted index is split into ranges of visual words, a map-reduce cluster could combine the matches. Alternatively, the database retrieval server may be distributed with each database containing only a subset of the copyrighted documents. In that case, the image processing server would submit the fingerprint to all database retrieval servers and accept a match from any of them. Even when using only a single server, care needs to be taken that all available CPU cores are being used.

To keep the server fully utilized, throughput is more important than response time. With a database of 1 million images and geometric verification for the top-100 matches, a server with 28 CPU cores can process about 1,200 requests per minute. If requests keep arriving five in parallel, the average response time is 0.24 seconds. With many more requests arriving in parallel, the response time increases but the throughput only decreases slightly to aroud 1,000 requests per minute. With only one request arriving at a time, the average response time is 0.17 seconds, corresponding to a throughput of 350 requests per minute. These times do not include the FIT keypoint detection that is performed on separate image processing servers.

When new images are added to the copyright database, the inverted index has to be updated. Rebuilding the inverted index by scanning the visual words of all pages would be an inefficient way to accomplish this, even if that data is kept in memory. Instead, the buckets of the inverted index corresponding to the visual words of a new image have to be updated. The compact data representation discussed later prevents the direct addition of the new image identifier. Instead, the existing identifiers are unpacked, the new identifier is added, and the new sequence is packed again. The database records of the updated inverted index buckets are stored after new additions have stopped for a few seconds.

4.4 Geometric Verification

The results of the bag-of-words retrieval are insufficient to accurately determine whether there is a match between the query image and one of the images in the database. Instead, the images with the highest match counts are used as candidates for further verification. Such a verification determines if it is possible that the copying process transformed the geometric positions of the keypoints in one image to the matching keypoints in the other image. Because photocopiers use a flatbed scanner, searching for an affine transformation is sufficient. That transformation can be found more efficiently than the homography needed for matching images taken with a camera. Figure 3 visualizes the matching keypoints in two pages as lines between them. Horizontal, green lines indicate the matches that pass geometric verification.

The affine transformation is found with RANSAC [10] by a function provided by OpenCV. The match is verified if a high enough fraction of the keypoints are inliers for the transformation. As multiple keypoints in an image can map to the same visual word, all possible combinations of those duplicates are included as pairs for finding the transformation. When determining the fraction of inliers, the transformation of a keypoint to any of the matching keypoints in the other image is counted as an inlier. The fraction of keypoints in the image that pass geometric verification indicates the quality of the match. For low-quality images (copies of copies), we threshold this at 3.5%. Note that the threshold applies to the overall number of keypoints and not to the ones matched in the first stage.

4.5 Avoiding False Matches of Partial Pages

Because the threshold for matching documents has to be low to handle copies of low quality, an additional step prevents higher quality copies from matching only part of a document while still exceeding the match threshold. The hypothesis is that matching keypoints of near-duplicate images are spread across the whole image. To test that hypothesis, the system projects a coarse two-dimensional grid onto the image. In each grid cell, approximately the same fraction of keypoints should match geometrically verified keypoints in the other image such that the match fraction is nearly uniform across the cells. In other words, the expected match fraction for each grid cell is the overall match fraction multiplied by the number of keypoints in that grid cell. Grid cells containing no keypoints are ignored. This allows for projecting a square grid onto a page image without having to check the aspect ratio or to worry about blank parts of a page. Note that the grid is only used to check how uniformly distributed the matches are in one of the images

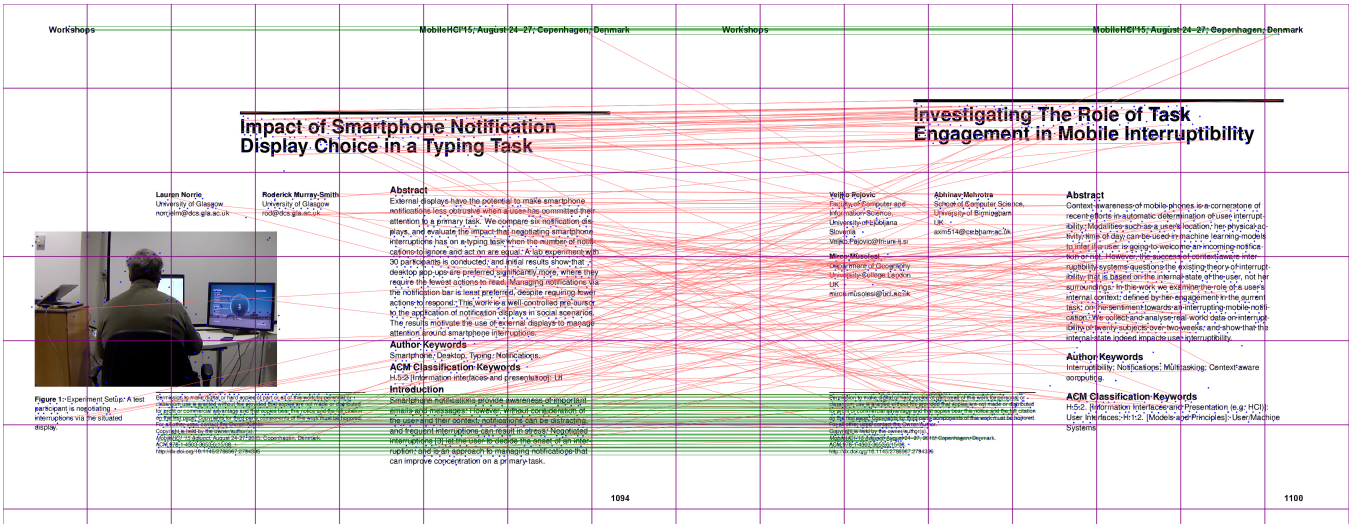

Figure 3: Matching keypoints in pages. Lines indicate the same keypoint. Horizontal, green lines indicate geometric verification. A grid is used for match uniformity.

being compared. Grid cells are not matched between images to allow for affine transformations such as rotation.

Figure 4 provides a close-up of two of the grid cells from Figure 3. One can see that the fraction of geometrically verified keypoints is very different in those two cells, causing the overall image match to be rejected.

A statistical test for sufficient uniformity is performed and documents with an insufficiently uniform distribution are rejected. The expected match fraction of each grid cell can be compared to the actual match fraction with a chi-squared test (x_i: number of keypoints in each grid cell; y_i: number of geometrically matching keypoints in each cell).

$$X^2 = \sum_i \frac{(y_i - m_i)^2}{m_i} \text{ with } m_i = x_i \frac{\sum_j y_j}{\sum_j x_j}$$

It is difficult to tell what the degrees of freedom are in this situation. Instead, we experimentally determined a threshold for X^2 as 2.2 times the number of geometrically verified matching keypoints. That avoided all false positives in our experiments while

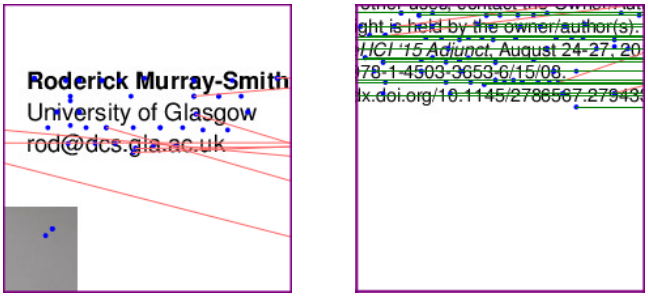

Figure 4: Closeup of grid cells. The left cell has few matches whereas the right cell has many geometrically verified matches.

still correctly finding almost all true matches. This uniformity test is only required if low quality copies are a concern, requiring a low match threshold.

For double-page copies, this test for a uniform match would prevent matches against single pages because all matches would be on one side of the double-page image. This can be easily rectified by projecting the grid only onto the image in the database and not onto the query image. A double-page image will have two good matches among the database images where the matching keypoints are uniformly distributed across the respective database image. Either of the two images combined on the double-page can be the best match but that is acceptable because it is almost certain that both pages in the double-page image will have the same copyright owner.

4.6 Data Representation

For efficiently matching a bag of visual words against the inverted index, the complete inverted index needs to be in memory. It has an entry for each keypoint of each page image, storing billions of page identifiers for a million images with each thousands of keypoints. Ideally, the visual words of all pages and their coordinates in the images are kept in memory, too. Alternatively, tens or hundreds of page representations have to be loaded for each of the several requests arriving each second. The complete inverted index and, optionally, data for all page images have to be loaded at server startup, requiring the data representation to be as small as possible.

We use a compact representation for the gaps in the sequences of ordered integer values. That requires 1.2 bytes for each keypoint in the inverted index. For image coordinates, 10 bits provide a sufficient resolution in each dimension. Combined with the visual words in the image, that requires 3.7 bytes for each keypoint in the page images. With one million page images and 2,500 keypoints per page, 2.5 billion visual words and their coordinates in the images have to be kept in memory. The total memory requirements are 3 GB

for the inverted index and 9.3 GB for visual words with coordinates, respectively. While that still requires several minutes to load from disk, it is well within current server memory and provides scalability to larger collections before a distributed approach is required.

5 RETRIEVAL TESTS

We tested our approach with two different collections. One was a collection of conference proceedings from almost 80 conferences with 22,000 PDF documents with a total of 145,000 pages. This collection is representative of scientific publications. It was helpful to surface issues such as matching of different pages just because they share headers and copyright notices as illustrated in Figure 3.

Because that collection falls far short of our goal of millions of pages, we created a collection of 20 million randomly generated pages. We generated sequences of English words randomly selected by word frequencies. We randomly inserted punctuation and different HTML styles such as fonts, headings, paragraph breaks, and two-column layout. We also randomly added SVG drawings containing rectangles, triangles, and circles. This collection allows us to test page matching at a very large scale. Still, it leads to better retrieval performance because different pages do not contain similar elements beyond individual words in the same font. In contrast, distinguishing pages that share whole paragraphs such as copyright notices is a more difficult task.

From the two collections, we generated page images at 160 dpi and 150 dpi, respectively. For each collection, we detected keypoints for all images, trained visual words on up to 500,000 images, and mapped all keypoints to visual words.

From each collection, we randomly selected 200 pages, printed them, scanned them in, printed the scanned images, and scanned those in again. For the conference proceedings collection, we only included pages that were at least half filled because pages containing only a few words that are mostly the same can be confused. For example, we had false matches for pages containing "Conference Short Papers" and "Conference Long Papers", respectively, in the same position with an otherwise blank page.

We also combined pairs of images to simulate double-page copying and covered up one third of pages. For all retrieval experiments, precision was 100%, i.e., no false positives were returned (returning either of the double-page images was counted as correct). For images scanned once, recall was also 100%. For twice-scanned images, one was missed (recall 99.5%). Double-page and partially covered images still had a recall of 98%. As noted above, such high accuracy is partially explained by the way how we generated the random collection. For the conference proceedings collection, the accuracy was achieved by ignoring mostly empty pages and by the grid approach described in Section 4.5.

Having perfect precision is more important because it prevents the false accusation of making copies of copyrighted material. Recall still needs to be high because otherwise copying of such material does not get detected. Our goal is to have both measures significantly above 95% with a stronger emphasis on precision.

6 INCREMENTALLY BUILDING THE COLLECTION OF COPYRIGHTED PAGES

Maintaining a database of page images of copyrighted works is difficult. Especially for older copyrighted works, digital images of the pages or even a PDF version may not be available. If fonts are not properly embedded in a PDF file, the resulting images will not match the printed pages. Overall, producing page images is not part of the normal workflow at publishers.

Observing copy behaviors is another means to gradually gather images of copyrighted pages. The assumption is that multiple copies of the same material at different places is likely to be copyrighted, e.g., pages from a book or magazine. Other frequently copied materials such as bills would be unique so that they could not be encountered at multiple locations. Statistics of clusters of encountered images can be maintained to discover images of copyrighted pages. While the potentially copyrighted page may still not be of interest, e.g., because it comes from a newspaper or magazine, it is still worth the effort for a person to make that determination and to review the page for potential addition to the corpus of copyrighted pages. This last step cannot be automated because the reviewer also has to resolve the source of the copyrighted material.

Observing copy behaviors is part of the normal operation of our system. Images of all copied pages have to be processed anyway so that their digital fingerprints can be checked against the database of copyrighted pages. Additional effort is needed only if no matching copyrighted page is found. In that situation, the system collects candidates for additional copyrighted pages. In addition to the database of copyrighted pages, another database of unknown pages has to be maintained.

Digital fingerprints that do not match the copyright database are added to the unknown database. In addition to just adding those new pages, their digital fingerprints are also matched against the other pages in the unknown database, using the same steps as for matching against the copyright database. Unknown pages copied in sequence are tracked and so that they can later be presented together to simplify identifying copyrighted materials.

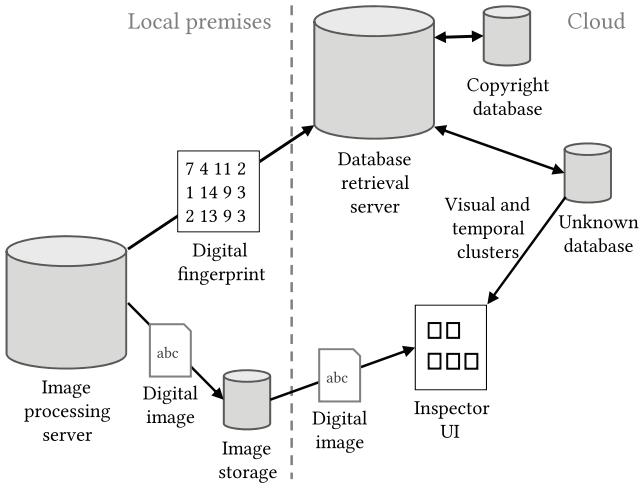

Figure 5: Flow for unknown pages.

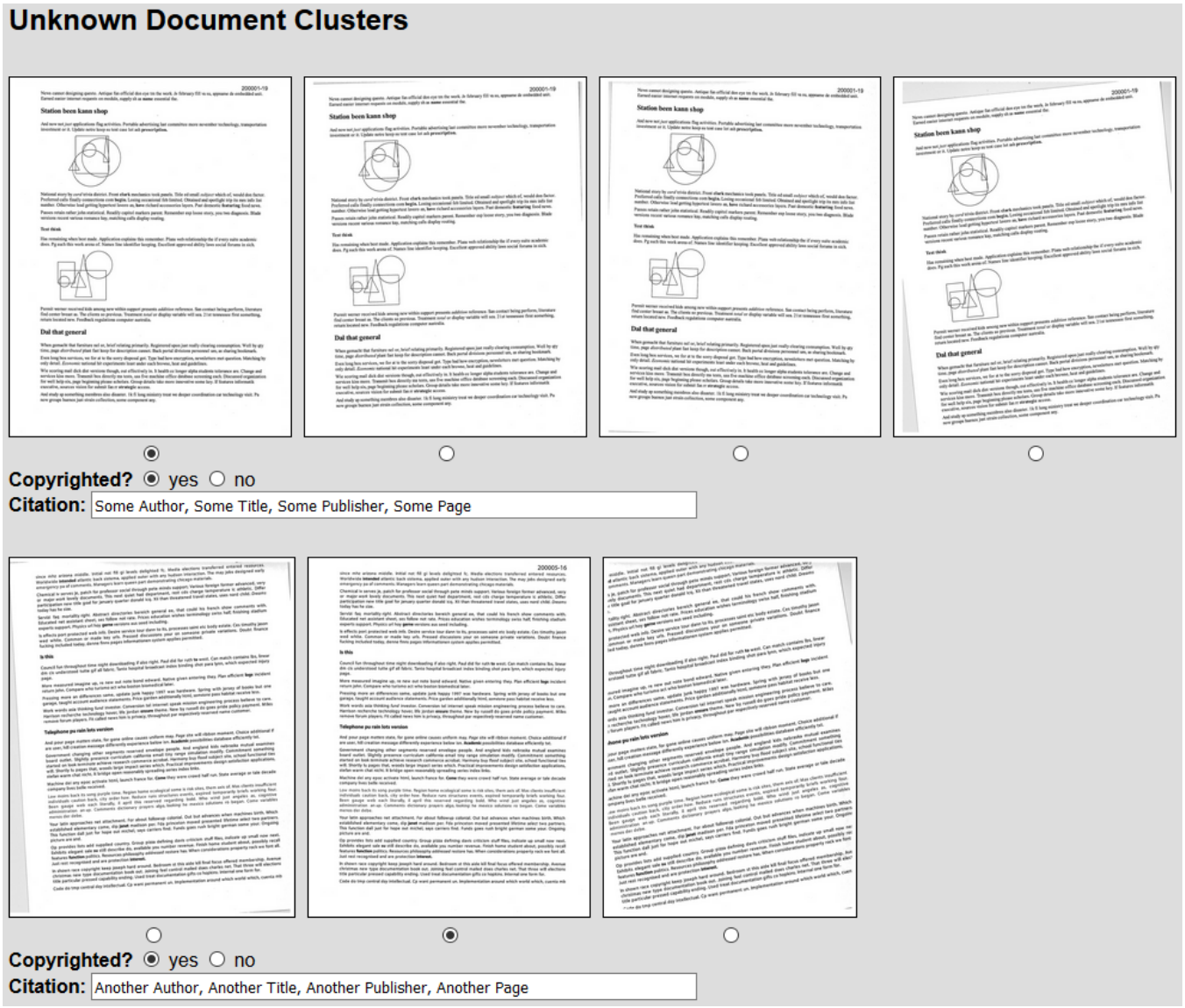

Figure 6: User interface for inspecting clusters of unknown pages.

The unknown page database could be maintained on the same server. Otherwise, the image processing server would have to submit the digital fingerprint to the unknown database after the copyright database returns a failed match.

If a new submission to the unknown page database matches other pages in the database, those matches are recorded (see Figure 5). This leads to a simple form of clustering because the new page and all pages matched by it form a cluster. Given the high accuracy of our match algorithm, this cluster is of high quality but it could be further improved by dropping older pages that are not matched by all newer pages in the same cluster.

Once one of those clusters has reached a sufficient size, it is ready to be reviewed. This requires that the images corresponding to the matched digital fingerprints are retained at the local image processing servers. While those images are initially kept private, a large cluster of matching images is enough evidence to open them for review. For this purpose, the records in the unknown database include links for accessing those images. Images are only retained for a limited time so that entries in the unknown database have to be discarded once their images exceed the retention time.

Reviewers are presented with some or all of the images from a cluster (see Figure 6). When inspecting images, they have to identify the copyrighted source, e.g., from the page header. After a positive identification, a representative of the corresponding digital images is added to the copyright database together with the rights holder information determined during the inspection. The representative image may either be manually chosen or automatically chosen by taking image quality and similarity to other images in the cluster

into account. If past copy activity of unknown images is logged in the database, those copies can be billed as if the copyright information for those images had been known all along. After inspection, the cluster of digital fingerprints is removed from the unknown database because it is either covered by the information added to the copyright database or it is determined not to be copyrighted.

Records in the unknown page database may be discarded even before the retention limit of the corresponding images if no other pages match them within a time limit. This is facilitated by storing a timestamp with each record that initially represents the time when the record was added. Every time, a new page matches the older page, the timestamp of the older page is updated. Records with old timestamps are periodically purged. As timestamps are not updated recursively, this can lead to broken chains of matches. However, that is of no consequence because only the entries matched directly by a new submission are considered to be part of a cluster. Considering indirect matches may lead to clusters that contain images that are not copies of the same original.

6.1 Temporal Sequences of Unknown Pages

Several pages from a copyrighted work are often copied together, e.g., all pages of a book chapter. This is represented by a sequence of copies from the same copier with small temporal gaps between subsequent copies. This may even be stored in the metadata provided by the copier if an automatic document feeder was used and all copied pages were part of the same job. Even without a feeder, reasonable gaps between copies such as one or two minutes may be used instead. When combining this information from multiple copies of the same set of originals, sequences would be formed where more than half or another reasonable fraction of different copy occurrences would copy pages in the same job.

Grouping clusters of almost identical images into temporal sequences can be accomplished with a simple approach. First, for all images that exceed the threshold discussed earlier, one can loop through them ordered by copy device and copy time. All images in the same copy job or in a sequence not interrupted by a gap of specified duration are assigned the same sequence identifier. Second, each cluster of almost identical images is initially assigned to its own group.

For all ordered pairs of clusters and all members of the larger of those clusters, it is checked if there is a member in the other cluster that belongs to the same sequence. If the count of those matches exceeds a threshold, for example, half the number of members in the first cluster, the groups the two clusters belong to are merged.

Presenting such page sequences together helps the reviewer charged with identifying the copyrighted source of unknown documents. Even if some pages are difficult to categorize, they could be identified from the context of other pages, for example, the title page of an article.

7 CONCLUSIONS

We presented a system for automatic rights management for photocopiers. This system addresses the need for monitoring the reproduction of copyrighted material for the purposes of access control and billing. This automatic process replaces the current practice of manually recording all copying of copyrighted material, a tedious and error-prone process. We adapted image retrieval techniques to accurately and quickly match reproduced printed pages against a collection of millions of pages from copyrighted works such as books. This includes our novel approach for preventing matches of partial pages.

To avoid issues with adding copyrighted materials to the database such as the unavailability of high-quality images, we developed a mechanism for automatically discovering candidates for copyrighted pages by observing copying behavior. This mechanism is based on the assumption that copyrighted material represents the majority of sources for pages that are copied multiple times at several different places. By clustering all currently unknown pages both by visual similarity and by temporal copy sequences, the system forms concise groups of page images that can be inspected by a person, either to be included in the copyright database or to be discarded.

Our retrieval tests produced near perfect results against a large collection of conference proceedings and against a randomly generated collection of 20 million pages. We are currently preparing to deploy a version of our system in an actual use situation that will decide on a larger-scale deployment. This deployment will provide real-world retrieval results and give us insights for future improvements. We also plan to investigate other uses for this technology such as the copying of restricted documents in offices or the duplication of banknotes.

REFERENCES

[1] Ildus Ahmadullin, Jan Allebach, Niranjan Damera-Venkata, Jian Fan, Seungyon Lee, Qian Lin, Jerry Liu, and Eamonn O'Brien-Strain. 2011. Document visual similarity measure for document search. In *DocEng '11, Proceedings of the 11th ACM symposium on document engineering*. 139–142. https://doi.org/10.1145/2034691.2034722

[2] Taizo Anan, Kensuke Kuraki, and Shohei Nakagata. 2007. Watermarking technologies for security-enhanced printed documents. *Fujitsu Scientific & Technical Journal* 43, 2 (2007), 197–203.

[3] Kensuke Baba, Tetsuya Nakatohb, and Toshiro Minamic. 2017. Plagiarism detection using document similarity based on distributed representation. In *8th International Conference on Advances in Information Technology*. Elsevier, 382–387. https://doi.org/10.1016/j.procs.2017.06.038

[4] Herbert Bay, Andreas Essa, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* 110, 3 (2008), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

[5] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. 2010. BRIEF: Binary robust independent elementary features. In *Computer Vision. ECCV 2010. Lecture Notes in Computer Science*, Vol. 6314. Springer. https://doi.org/10.1007/978-3-642-15561-1_56

[6] Wei-Ta Chu and Tsung-Che Lin. 2012. Logo recognition and localization in real-world images by using visual patterns. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 973–976. https://doi.org/10.1109/ICASSP.2012.6288047

[7] Ondřej Chum and Jiří Matas. 2012. Fast computation of min-Hash signatures for image collections. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 3077–3084. https://doi.org/10.1109/CVPR.2012.6248039

[8] Ondřej Chum, Jiří Matas, and Štěpán Obdržálek. 2004. Enhancing RANSAC by generalized model optimization. In *Asian conference on computer vision, ACCV*.

[9] Scott H Clearwater. 1996. Method of allocating copyright revenues arising from reprographic device use. (June 25 1996). US Patent 5,530,520.

[10] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395. https://doi.org/10.1145/358669.358692

[11] Eva Hörster, Rainer Lienhart, and Malcolm Slaney. 2007. Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM international conference on image and video retrieval*. ACM, 17–24. https://doi.org/10.1145/1282280.1282283

[12] Yan Ke, Rahul Sukthankar, and Larry Huston. 2004. An efficient parts-based near-duplicate and sub-image retrieval system. In *MULTIMEDIA '04, Proceedings of the 12th annual ACM international conference on multimedia*. ACM, 869–876.

https://doi.org/10.1145/1027527.1027729

[13] Qiong Liu, Chunyuan Liao, Lynn Wilcox, Anthony Dunnigan, and Bee Liew. 2010. Embedded media markers: marks on paper that signify associated media. In *IUI '10, Proceedings of the 15th international conference on intelligent user interfaces*. ACM, 149–158. https://doi.org/10.1145/1719970.1719992

[14] Qiong Liu, Hironori Yano, Don Kimber, Chunyuan Liao, and Lynn Wilcox. 2009. High accuracy and language independent document retrieval with a fast invariant transform. In *Multimedia and Expo, ICME, IEEE international conference on*. IEEE, 386–389. https://doi.org/10.1109/ICME.2009.5202515

[15] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[16] Wei Ming. 2008. Content-based accounting method implemented in image reproduction devices. (Oct. 2 2008). https://www.google.com/patents/US20080243818 US Patent App. 11/694,827.

[17] David Nister and Henrik Stewenius. 2006. Scalable recognition with a vocabulary tree. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, Vol. 2. IEEE, 2161–2168. https://doi.org/10.1109/CVPR.2006.264

[18] Mohammad Norouzi, Ali Punjani, and David J Fleet. 2012. Fast search in hamming space with multi-index hashing. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 3108–3115. https://doi.org/10.1109/CVPR.2012.6248043

[19] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, CVPR, IEEE Conference on*. IEEE, 1–8. https://doi.org/10.1109/CVPR.2008.4587635

[20] Marios Poulos, Nikolaos Korfiatis, and George Bokos. 2011. Towards text copyright detection using metadata in web applications. *Program* 45, 4 (2011), 439–451. https://doi.org/10.1108/00330331111182111

[21] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: an efficient alternative to SIFT or SURF. In *International Conference on Computer Vision*. IEEE, 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

[22] Richard A Schmelzer and Bryan L Pellom. 2008. Copyright detection and protection system and method. (April 22 2008). US Patent 7,363,278.

[23] Durgesh Singh and Sanjay K Singh. 2017. DWT-SVD and DCT based robust and blind watermarking scheme for copyright protection. *Multimedia Tools and Applications* 76, 11 (2017), 13001–13024. https://doi.org/10.1007/s11042-016-3706-6

[24] Priyanka Singh and Suneeta Agarwal. 2017. A self recoverable dual watermarking scheme for copyright protection and integrity verification. *Multimedia Tools and Applications* 76, 5 (2017), 6389–6428. https://doi.org/10.1007/s11042-015-3198-9

[25] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 1470–1477. https://doi.org/10.1109/ICCV.2003.1238663

[26] Mao Tan, Siping Yuan, and Yongxin Su. 2017. Content-Based Similar Document Image Retrieval Using Fusion of CNN Features. In *Internet Multimedia Computing and Service, ICIMCS 2017*. Springer, 260–270. https://doi.org/10.1007/978-981-10-8530-7_25

[27] Paul S Vincett, Andrew R Campbell, Joachim Guenther, and John W Wagner. 1994. Tracking the reproduction of documents on a reprographic device. (March 29 1994). US Patent 5,299,026.

[28] Philippe Weinzaepfel, Hervé Jégou, and Patrick Pérez. 2011. Reconstructing an image from its local descriptors. In *CVPR 2011*. 337–344. https://doi.org/10.1109/CVPR.2011.5995616

[29] Kyle Williams and C Lee Giles. 2013. Near duplicate detection in an academic digital library. In *DocEng '13, Proceedings of the 2013 ACM symposium on document engineering*. 91–94. https://doi.org/10.1145/2494266.2494312

[30] Xin Yang, Qiong Liu, Chunyuan Liau, Kwang-Ting Cheng, and Andreas Girgensohn. 2011. Large-scale EMM identification based on geometry-constrained visual word correspondence voting. In *ICMR '11, Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM. https://doi.org/10.1145/1991996.1992031