Large-Scale EMM Identification Based on Geometry-Constrained Visual Word Correspondence Voting

Xin Yang[1], Qiong Liu[2], Chunyuan Liao[2], Kwang-Ting Cheng[1], Andreas Girgensohn[2]
[1] Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, USA
[2] FX Palo Alto Laboratory, 3400 Hillview Avenue, Bldg.4, Palo Alto, CA 94304, USA
xinyang@umail.ucsb.edu, {liu, liao}@fxpal.com, timcheng@ece.ucsb.edu, andreasg@fxpal.com

ABSTRACT
We present a large-scale Embedded Media Marker (EMM) identification system which allows users to retrieve relevant dynamic media associated with a static paper document via camera-phones. The user supplies a query image by capturing an EMM-signified patch of a paper document through a camera phone. The system recognizes the query and in turn retrieves and plays the corresponding media on the phone.

Accurate image matching is crucial for positive user experience in this application. To address the challenges posed by large datasets and variation in camera-phone-captured query images, we introduce a novel image matching scheme based on geometrically consistent correspondences. A hierarchical scheme, combined with two constraining methods, is designed to detect geometric constrained correspondences between images. A spatial neighborhood search approach is further proposed to address challenging cases of query images with a large translational shift. Experimental results on a 200k+ dataset show that our solution achieves high accuracy with low memory and time complexity and outperforms the baseline bag-of-words approach.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, search process*.
I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision – *Applications*.

General Terms
Algorithms, Design, Experimentation.

Keywords
camera-phone applications, EMM identification, image matching, hierarchical gridding, approximate geometric verification, translation compensation.

1. INTRODUCTION
Techniques of linking dynamic media with a static paper document via camera phones have many interesting applications, such as multimedia enhanced books and multimedia advertisement on paper. Techniques based on 2D barcodes [10] are commonly used, which can be easily recognized via modern camera-phones. However, barcodes, when printed on a document for association with specific document contents, are visually obtrusive and could interfere with the document layout. Thus several current systems [11,12,13,17] rely on the document content itself for identification of media association. For example, from the camera-phone captured query image, SnapTell [12] and Kooba [13] extract visual features, and HOTPAPER [11] utilizes the layout of word boxes to determine the source page, and the location on the page for linkage to the corresponding media data. However, these methods can hardly achieve good accuracy and scalability without clear specifications of which contents/locations on a paper document are linked to media data. A query image captured aimlessly on a document without specific guidance may result in various distortions and thus lead to low accuracy. In addition, as the camera-phone may capture any part of the page, the system would need to characterize and index the entire document pages [11], resulting in high time/memory cost for large datasets. To alleviate these problems, an EMM identification system has recently been introduced [7,8], which utilizes meaningful awareness-markers overlaid on the original paper document to guide image capture and limit processing cost. However, the current EMM identification system [7, 8] relies on

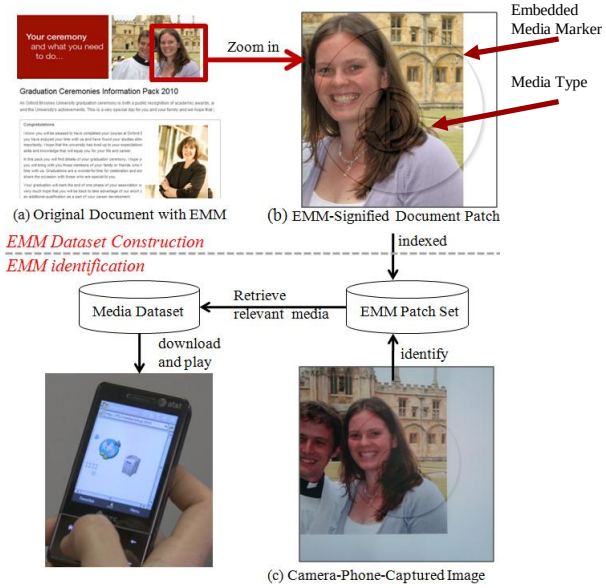


Figure 1: EMM dataset construction and identification

ICMR'11, April 17–20, 2011, Trento, Italy.

general local-feature-based matching approaches without taking into account any application-specific matching constraints for this application. As a result, it suffers low accuracy and high memory/time complexity. To address this problem, this paper analyzes the unique characteristics of image matching for EMM identification and proposes a novel image matching scheme which utilizes such characteristics to improve accuracy and to reduce memory and time complexity.

1.1 EMM identification

Embedded Media Markers (EMMs) [7,8] are nearly transparent markers printed on paper documents at certain locations which are linked with additional media information. Analogous to hyperlinks, EMMs indicate the existence of links to digital media. Users can capture the EMM-signified document patch with a camera-phone to retrieve the associated digital media and view it on the phone. Figs. 1(a) and (b) show the original document with an EMM overlaid at the top-right corner and a close-up of an EMM-signified patch, respectively. For EMM-enriched documents, only the EMM-signified patches need to be characterized and indexed, which can greatly reduce the time and memory usage for feature extraction and enhance the accuracy because of the exclusion of noisy features which correspond to contents outside an EMM region. The EMMs can guide users to capture an EMM-signified region, yielding a query image with little distortion (as shown in Fig.1(c)). The task of EMM identification is therefore to match the camera-phone-captured query image to an original EMM-signified patch indexed in the dataset.

1.2 Image Matching for EMM Identification

Accuracy and scalability of image matching are crucial for large scale EMM applications, e.g. linking media for cyclopedia or daily newspapers over a period of multiple years. Several image matching approaches [1,2,3,4,9] have been proposed and successfully employed in similar applications, such as image-based object recognition and near-/partial-duplicate detection. However, these generic methods could not utilize two particular matching constraints, namely *"injection"* and *"approximate global geometric consistency"* (*AGGC* for short), which are unique for the EMM identification. As a result, these methods unnecessarily cost more memory and time in order to achieve a satisfactory accuracy for this application.

The injection constraint is enforced by the way of generating a query image in EMM identification, where a query image is a camera-captured version of an original EMM-signified patch, as shown in Fig. 2(a). This property implies that each detected "salient" region of a query image can be mapped to by at most one common region of the target image, i.e. *"injective mapping"*. Such constraint may not hold in near-/partial-duplicate image detection, where a query image could be generated by extensive digital editing of an original image. Fig. 2(b) illustrates an exemplary case that violates this constraint while needs to be targeted by partial-duplicate detection.

The AGGC constraint is enforced by EMMs, which confine the geometric changes between a query image and its target within a small predictable range, so that the spatial layout of a query image should be globally consistent with that of its target image with high fidelity. Such constraint does not always hold in other similar applications. Fig. 2(c) illustrates an example of two images containing the same object of very different scale. Matching them

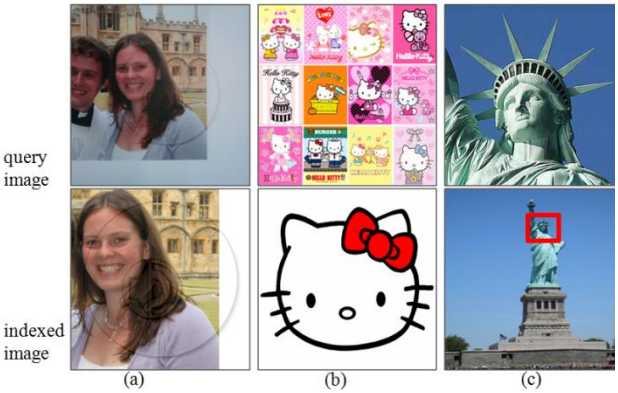


Figure 2: query images and indexed images of (a) EMM identification; (b) partial duplicate detection; (c) image-based object recognition

is required for object recognition applications, but is not expected for EMM identification.

Limiting the scope of the search by taking into account these two constraints can help further increase accuracy and reduce memory and time complexity. In this paper, we propose a novel image matching scheme, which takes advantage of these two constraints. To achieve high accuracy, two constraining methods are designed based on multi-resolution gridding information to detect *"injective"* and *"AGGC"* correspondences and uses them for measuring image similarity. A spatial neighborhood search approach is further proposed to address challenging cases for which the query image has a large translational shift. To achieve scalability, a hierarchical strategy is proposed to compact the memory and limit the processing time. Experiments based on a 100k+ dataset for EMM identification show that our matching scheme can achieve 96.7% accuracy for document images and 94.5% accuracy for natural images while cost less than 2GB memory usage and ~500ms average indexing time.

The rest of the paper is organized as follows: Section 2 overviews existing image matching techniques. Section 3 presents the goal, details, and special features of our matching scheme. In Section 4, we report the experimental results, followed by the conclusion in Section 5.

2. EXISTING IMAGE MATCHING TECHNIQUES

Most state-of-the-art image matching approaches rely on local feature representations [5] to achieve high accuracy. The current EMM identification system [7,8] matches each query local feature to an indexed feature by an approximate nearest neighbor search scheme, K-D tree, based on L2 distance. And then the number of matches, whose distances are within a predefined threshold, is used for ranking database images. Grauman et al. propose a pyramid matching scheme [14] which works by placing a sequence of increasingly coarser grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution. These schemes provide an accurate image similarity measure but the high memory and time complexity for storing and processing all the local features prohibit its use for large-scale databases.

Bag-of-Words (BoW) matching [9] is an effective strategy to reduce memory usage and support fast matching via a scalable

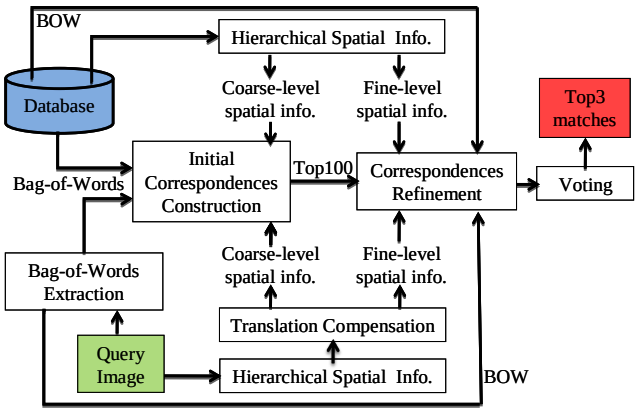


Figure 3: EMM identification with Geometry-Constrained Visual Word Correspondence Voting.

indexing scheme, e.g. an inverted file [15]. Typically, BoW matching quantizes local image descriptors into visual words and then computes the image similarity by counting the frequency of words co-occurrences. However, it completely ignores the spatial information; hence it may greatly degrade the accuracy. In order to enhance the accuracy for BoW matching, several approaches have been proposed to compensate the loss of spatial information. For example, geometric verification [5, 6], which is designed for general image-matching applications, is a popular scheme which verifies local correspondences by checking their homography consistency. Wu et al. present a bundling feature matching scheme [1] for partial-duplicate image detection. In their approach, sets of local features are bundled into groups by MSER [16] detected regions, and robust geometric constraints are then enforced within each group. All these schemes yield more reliable local-region matches by enforcing various geometric constraints. However, these schemes are either too computationally expensive or designed to meet unique requirements for specific applications, thus cannot meet the needs for EMM identification.

To some extent, spatial pyramid matching [2], which considers approximate global geometric correspondences, is suitable for EMM identification. The scheme partitions the image into increasingly finer sub-regions and computes histograms of local features found within each sub-region. To compute the similarity between two images the distance between histograms at each spatial level is weighted and summed together. However, this scheme cannot enforce *"injection matching"*, therefore cost unnecessary time to process lots of unqualified matches. In addition, without an efficient spatial information encoding/decoding strategy, the scheme needs to store all the histograms at every spatial level in memory; resulting in significant memory and time overheads.

3. Geometry-Constrained Visual Word Correspondence Voting

Fig. 3 illustrates the workflow of the proposed matching scheme - "**G**eometry-**C**onstrained Visual Word **C**orrespondence **V**oting" (*GCCV* for short). It consists of three major steps: (1) Initial "*AGGC*" correspondences construction places coarse-level grids over the image space and matches only those visual words residing in the same grids to one another. All the indexed images are then ranked based on the "*AGGC*" correspondences; (2) Correspondence refinement partitions each top-ranked image into fine-resolution grids, and verifies the initial correspondences using the *"injection"* constraint at fine granularity; (3) Finally, the qualified correspondences are used for ranking database images and the top3–ranked images are returned to users for a final confirmation. To further reduce errors caused by large translational shifts, we propose a "translation compensation" algorithm which estimates the translation changes and roughly aligns images before searching for the qualified correspondences. In addition, a hierarchical encoding/decoding strategy is incorporated for efficiently storing and utilizing the multi-resolution grid information.

In the following, we describe the motivations, techniques and strengths of each step in details.

3.1 Initial Correspondence Construction at Coarse Granularity

The "*AGGC*" constraint implies that spatial layout of a query image should be globally consistent with that of its target image with high fidelity. Therefore, we can assume that the corresponding features should locate at similar locations of two respective images. Based on this assumption [1], we propose a matching scheme called Grid-Bag-of-Words (G-BOW) matching for finding initial correspondents which satisfy the "*AGGC*" constraint. G-BoW matching partitions an image into n equal-sized grids and then matches a local feature f_q of a query image to a local feature f_{idx} of an indexed image if f_q and f_{idx} are quantized into the same visual word by the quantizer $q(.)$ and have the same grid-id. That is,

$$F_{G-BOW}(f_q, f_{idx}) = \begin{cases} 1 & \text{if } q(f_q) = q(f_{idx}) \,\&\, \\ & \quad grid\text{-}id(f_q) = grid\text{-}id(f_{idx}) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We can estimate the similarity for grid i by calculating the normalized sum of the G-BoW matching value for every query feature within a grid i,

$$sim(I_{qi}, I_{idxi}) = \frac{\sum_{f_q \in I_{qi}} F_{G-BOW}(f_q, f_{idx})}{|I_{qi}| \times |I_{idxi}|} \quad (2)$$

where $|I_{qi}|$ and $|I_{idxi}|$ are the total number of visual words within grid i of a query image and an indexed image, respectively.

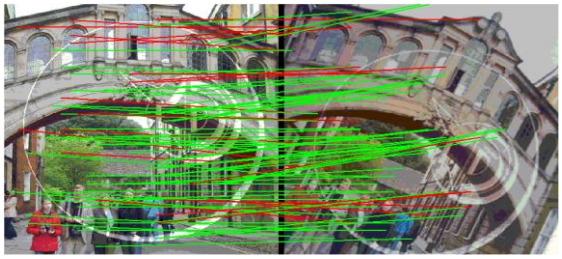
We then calculate the sum of the similarity scores of all grids, yielding the final image similarity between a query image I_q and an index image I_{idx}

$$sim(I_q, I_{idx}) = \sum_{i=0}^{n} sim(I_{qi}, I_{idxi}) \quad (3)$$

To illustrate the effectiveness of G-BoW matching, we compare the correspondences obtained by BoW matching with those from G-BoW matching using 4×4 grids. We estimate the homography using RANSAC [6] and then verify the obtained correspondences by geometric consistency. Fig. 4 displays the first 100 correspondences from the two correspondence sets. In this

[1] This assumption doesn't hold if the camera-phone is rotated by 90°, 180° or 270° when capturing a query image. But most modern smart-phones have an accelerometer to determine the orientation of the phone. In addition, an EMM should always be upright for normal reading. Therefore, in the rest of the paper, the algorithm focuses only on cases with a rotation respect to the 0° axis.

(a) Bag-of-Words overlapping (codebook size = 100K)
Qualified Correspondences Ratio = 47%



(b) Grid-Bag-of-Words matching(codebook size = 100K)
Qualified Correspondences Ratio = 75%

Figure 4: Top 100 correspondences between a query image (right) and the target image (left) obtained by (a) BoW matching and (b) 4x4 G-BoW matching. We estimate the homography using RANSAC [6] and verified all the correspondences by geometric consistency. Red lines denote unqualified matches which are inconsistent with the homography between the two images, while green lines indicate the qualified matches.

example, 75% of the matches obtained by 4×4 G-BoW satisfy the "*AGGC*" constraint; while only 47% qualified matches are obtained by BoW matching.

Memory Overhead. To implement G-BoW matching efficiently, we can offline compute the grid id of indexed local features and record them in an indexing file. This solution costs only slightly more memory space for an indexing file than that produced by BoW. For example, to record a grid id of 4×4 grids, we only need to add 4 more bits for each local feature.

Time Complexity. The time overhead for the matching includes: 1) online grid-id computation for features of a query image; and 2) fetching the grid-ids of indexed features from memory and comparing them with those of a query image. However, the experimental results in Section 4.2 show that, involving grid matching does not increase the overall computation time. On the contrary, it slightly decreases the time due to the time saved from unnecessary matching of unqualified features and updating their matching scores.

3.2 Correspondences Refinement at Fine Granularity

G-BoW matching provides us initial correspondences satisfying the *"AGGC"* constraint. Whereas, such scheme cannot guarantee the *"injective"* condition when *M* features, which are quantized into the same grid, match to *N (M≠N)* features quantized into a common grid. Increasing the number of grids, i.e. enforcing a stricter spatial constraint, may help exclude unqualified correspondences but may also decrease the robustness to geometric changes. To solve this problem, we can employ homography verification [6] which can preserve the property of *"injection"* when the perspective differences between two images are small (such a condition is satisfied in EMM identification). In a traditional procedure, a hypothesized homography is first estimated based on candidate correspondences at the pixel level, and each correspondence is then verified by checking the homography consistency. Finally, the matching score is updated according to the number of the homography consistent correspondences.

However, the traditional homography estimation and verification is not ideal due to the following reasons: 1) loading the pixel-level coordinates from the disks takes a lot of time; 2) homography estimation and verification using pixel-level spatial information is sensitive to keypoint location changes; 3) highly noisy match candidates obtained from BoW matching would significantly increase the time for deriving a transformation matrix and also decrease the accuracy of the estimated transformation parameters.

To address these limitations, we propose a more efficient verification procedure at the grid level, called Approximate Geometric Verification (AGV). Fine-level grids information of the initial "*AGGC*" correspondences is used for estimating the homography matrix. All match candidates are then verified based on the homography consistency, where a match candidate is defined as a pair of features assigned to the same visual word. Algorithm1 summarizes our approximate geometric verification process. It's worth mentioning that hard quantization for constructing *"AGGC"* correspondences may cause loss of qualified matches. Therefore, in this process we verify all match candidates to partially make up such loss. Regarding the choice of the number of grids for AGV, there is a tradeoff between distinguishability and space complexity: the more grids we use, the more precise the coordinates of correspondences, but the more bits needed to store the grid information. Empirically, we tested several choices including 16×16, 32×32 and 64×64, and concluded that 32×32 is the best choice.

Algorithm 1. Approximate Geometric Verification (AGV)

Definition: $M_{initial}$: initial correspondences; M_{all}: all match candidates based on BoW overlapping; $H_{homography}$: homography matrix; M_{final}: final correspondences after verification

for {TopK candidate images}

- G-BOW matching → $M_{initial}$;
- Homography estimation ($M_{initial}$)→ matrix $H_{homography}$;
- Align query image I_q with candidate image $I_{candidate}$ using $H_{homography}$.
- Verify all match candidates M_{all} and obtain qualified matches M_{final}.
- Update matching score S_{new} according to formula (2) and (3)

end

Re-rank TopK candidate images based on S_{new}.

AGV vs. traditional geometric verification. Regarding the runtime, AGV outperforms the traditional geometric verification due to two reasons. First, quantized location information is more compact (e.g. a 32×32 grid id takes only 10bits per feature) and can more likely be stored in memory, which helps reduce or even eliminate the time for disk IO during the refinement step. Second, correspondences obtained by G-BoW matching are much less noisy than those from BoW matching, thus using them can greatly reduce the estimation time. Experimental results also show that, using correspondences from G-BoW matching achieves much higher identification accuracy than using those from BoW.

3.3 Translation Compensation

For challenging cases which incur significant geometric changes, a hard quantization may inevitably discard many qualified *"AGGC"* correspondences and consequently degrade the homography estimation accuracy or even completely miss the target image if the target image fails to be ranked within the top candidate list. For example, a translational shift that is larger than $image\_size/n^{1/2}$ (n is the number of grids used in the "*AGGC*" correspondence construction step) will completely misalign all the grids. As a result, none of the *"AGGC"* correspondences can be detected for the target image. Therefore, adjustment for compensating the errors caused by misalignment is crucial for achieving good identification accuracy.

Before tackling this problem, we first evaluated the top1, top3 [2] and top100 accuracy with respect to the four observed geometric transformations and Table 1 summarizes the results. As shown in Table 1(a), for query images with large rotation and scale changes, the top1 and top3 accuracy numbers are close to that of top100 and are sufficiently high for satisfactory user experience. However, for query images with large translational shifts, the top1 and top3 accuracy numbers dropped dramatically. In addition, there exists a large gap (19%) between the top1 accuracy and the top100 accuracy. This result confirms that significant loss of *"AGGC"* matches would hurt the effectiveness of approximate geometric verification. Therefore, in the following, we propose a translation compensation algorithm to address the translation-caused errors.

A straightforward solution to solving the misalignment problem caused by hard-quantized grids is using a soft spatial assignment. In this simple solution, we assign a point to each of the eight neighboring grids in addition to the grid where the point falls in. However, this simple strategy also introduces extra noises and consequently decreases the accuracy and increases the runtime. In most cases, out of nine quantized directions, there is only one direction which can best approximate the real translation change. Thus, most points assigning to the other wrong grids become noise.

To minimize translation-caused errors, we propose a better solution, which estimates the best direction for translation compensation between two images and then assigns all the points to this estimated direction. The idea is based on the fact that if all points are shifted towards the best direction for translation compensation, the majority of grids should achieve the maximum similarity score (based on Formula 2). In other words, the direction, among all nine directions, which results in the maximum matching scores over all the grids would the best direction for translation compensation. After obtaining the best translation direction, each point is then assigned to this direction for finding *"AGGC"* correspondences. Therefore, we can derive set M_{best} which contains correspondences between words of the current grid and words of the best neighboring grid. To compensate the errors caused by translation changes, we compute the matching score and estimate the homography based on M_{best}. Table 1(b) shows the accuracy improvement after using translation compensation. The accuracy for translated testing images is greatly enhanced for all three settings. At the same time the accuracy remains the same, or even becomes better, for the test cases with other transformations.

Table 1. Accuracy for test images with different transformations

(a)

Transforms / Accuracy	Top 1	Top 3	Top 100
Translation	**0.66**	**0.69**	**0.85**
Rotate	0.89	0.90	0.91
Up-Scale	0.95	0.95	0.96
Down-Scale	0.92	0.95	0.99

(b)

Transforms / Accuracy	Top 1	Top 3	Top 100
Translation	**0.85**	**0.88**	**0.96**
Rotate	0.89	0.92	0.93
Up-Scale	0.96	0.97	0.97
Down-Scale	0.93	0.96	1.00

(Examining the effect of different transformations on accuracy for the testing set-540 (which will be described in Section 4.1). (a) Using G-BoW matching + AGV. (b) Using G-BoW matching + TC + AGV. The vocabulary is trained based on 2k+ document images and a size of 100k. The size of the dataset is 100k+.)

3.4 Hierarchical Encoding / Decoding

An efficient strategy for storing and decoding the multi-resolution spatial information should meet the following three objectives: 1) taking as little memory space as possible; 2) efficiently computing the desired information, including the coarse-level grid id, the neighboring grid id, and the fine-level grid id; 3) easy to adjust the parameters, such as number of coarse-level grids. In this section, we present a hierarchical encoding and decoding strategy which is designed to meet these objectives. Each image is hierarchically quantized into $2^k \times 2^k$ grids: an image is firstly partitioned into 2×2 grids and then each grid is iteratively subdivided into 2×2 grids, yielding $2^k \times 2^k$ grids at level k, as shown in Fig. 5. Then each grid at level k is encoded by coordinates (x_i, y_i), $(1 \le i \le k)$, uniquely denoting one of the 4 positions in the upper level grid (x_{i-1}, y_{i-1}). Finally the coordinates at all levels are concatenated together to form a bit string, as shown in Fig. 5.

Memory complexity: Given the number of the finest-level grids, this scheme takes least amount of memory space by embedding all the coarser-level information into the corresponding finest-level grid id. In addition, such information can be bundled with image id of each local feature and stored in the inverted file for fast access. Fig. 6 shows the structure of our index. Each visual word has an entry in the index that contains a list of images in which the visual word appears. We use an integer to record the image and geometric information: the left most 22 bits is utilized to

[2] Before retrieving the relevant media, the system returns the top-ranked results to the user for a final confirmation. In order to trade off between visibility and the number of displayed results, we usually show the top3 images to the user on the cell-phone. Therefore, we are interested in Top3 accuracy.

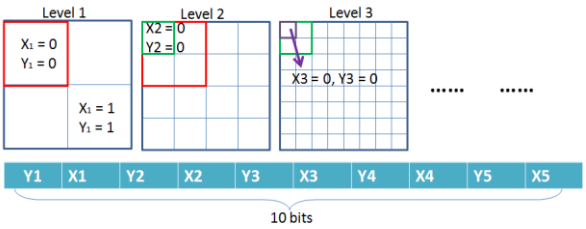


Figure 5. Hierarchical Quantization and Encoding

record the image ID, the remaining 10 bits is employed to record the hierarchical geometric information. This format supports at most 4 million indexed images, and the finest spatial resolution is 32×32.

Time complexity: This hierarchal strategy can parse all the required information using a small number of bit/add/subtract operations, which is very efficient in practical use.

4. EXPERIMENTS

In this section, we first introduce our database and two manually-captured testing sets for EMM identification. Then we report the performance comparison of the proposed matching scheme with the baseline bag-of-words approach [15]. Finally, we provide experimental results to exam the effects of different parameter settings and training images on the identification accuracy.

4.1 Database Description

4.1.1 Database

Our database contains three datasets: 1) EMM-ICME2K - generated from the ICME06 proceedings, which has 2188 letter-size (8.5"x11") document pages with text, images, and figures; 2) EMM-Oxford5K - constructed from Oxford5K dataset consisting of 5062 natural images including oxford buildings, groups of people, etc; 3) EMM-Flickr200K – generated from 200K distracting images arbitrarily retrieved from Flickr. In order to evaluate the performance with respect to different dataset sizes, we also built four smaller datasets - 10K, 20K, 50K, and 100K respectively.

All the database images are pre-processed using the following procedures: we overlay one EMM via the authoring tool [8] on each image and then crop a square region whose center is aligned with the center of the EMM with its side-length equal to 1.42 times the boundary-circle diameter of the EMM feature (refer to [8] for more details). After that, all the cropped images are normalized to 707×707 for the same reason presented in [8]. Figs. 7 (a) and (g) show two exemplar database images from EMM–Oxford5K and EMM-ICME2K respectively.

4.1.2 Testing Set

To evaluate the performance on different types of images, we constructed two testing sets: 1) testing set-540 for document images and (2) testing set-595 for natural images.

To establish the "ground truth" for these two sets, we randomly select 108 pages from the ICME06-corpus, 59 images containing buildings and 60 images containing groups of people from Oxford5k. For each page or image, we manually took five pictures of the EMM-signified patch: one image was taken with little geometric change and the other four images were taken by 1) shifting the EMM to the edge of the camera screen; 2) rotating the camera upto ±45°; 3) moving the camera closer to the document so that the boundary of the EMM hits the screen edge of the

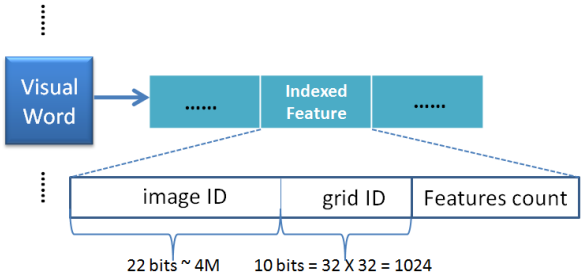


Figure 6. Inverted file structure. "image ID" and "grid ID" are encoded together into an integer. "Features count" is the number of feature records of current grid in one image.

camera, i.e. up-scaling; and 4) moving the camera farther from the document, i.e. down-scaling, respectively. We cropped the center square region of each captured image and normalized it to the size of 707×707. Figs. 7 (b)-(f),and (h)-(l) illustrate some exemplar query images from testing set-595 and testing set 540 respectively.

In the following evaluation, we use the top3 accuracy, which indicates the rate that a query image is ranked within the top three positions, as our identification accuracy.

4.2 Evaluation of our method

Baseline. We used the bag-of-words approach as the "baseline" approach. We trained a vocabulary of 100k visual words using hierarchical k-means [15]. The training images consist of all the images from EMM-ICME2K and EMM-Oxford5K.

Comparison. We then enhanced the baseline method with our approximate geometric verification. We tried four different configurations: 1) "G-BoW", in which we only use the *"AGGC"* constraint; 2) "G-BoW + TC", which compensates for the translation-caused errors before finding the qualified correspondences; 3) "G-BoW + AGV", which applies both the *"AGGC"* constraint and the *"injection"* constraint when finding the qualified correspondences; and 4) "G-BoW + TC + AGV" which combines Configurations (2) and (3). In this experiment, the number of coarse-level grids n in (3) is set to 4×4=16 (we will discuss another experimental study for the effect of varying n in Section 4.3).

We perform all the experiments with a single CPU on a 3.0GHz Core Duo desktop with 12G memory. Fig. 8 compares the above six approaches with respect to accuracy and runtime. Three key observations can be made from these results. First, applying the *"injection"* and *"AGGC"* constraints significantly improves the accuracy, which can be observed by comparing the results for "G-BoW+AGV" to "baseline-BoW". For the 100k+ dataset, the accuracy increases from 6% to 96.7% for the document images and from 22% to 94.3% for the natural images. Second, the accuracy and runtime improvements achieved by AGV are much more significant when combined with "G-BoW" than with "baseline-BoW". The accuracy is 11% higher for both document images and natural images on the 100k+ dataset, if we compare the curves of "G-BoW+AGV" and "G-BoW". However, the accuracy of "baseline+AGV" is only 2% higher for the document images and 11% higher for the natural images in comparison with the accuracy of "baseline". For runtime comparison, the time increase changing from "G-BoW" to "G-BoW+AGV" is less than the time increase changing from "baseline" to "baseline+AGV" - 159ms less for a document image and 54ms less for a natural image. These results further validate the claim made in Section3.3

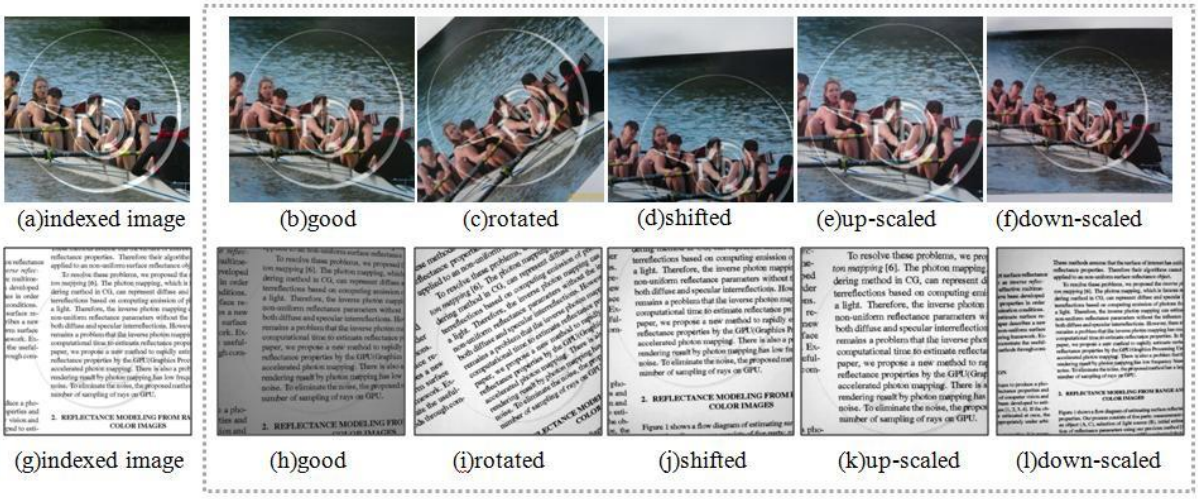


Figure 7: (a) and (f) are exemplar indexed natural image and document image, respectively. (b)-(f) and (h)-(l) are corresponding exemplar hand-captured query images from testing set-595 and testing set-540.

that too much noise in the correspondence set may greatly degrade the effectiveness of AGV. Finally, translation compensation boosts the accuracy for both approaches. With translation compensation ("G-BoW+TC+AGV"), we can achieve the highest accuracy 96.7% for the document images and 94.5% for the natural images on the 100k+ dataset.

4.3 Coarse Level Grids Number Selection

The number of grids *n* in Formula (3) is a key parameter for the G-BoW matching and translation compensation, and affects the results of approximate geometric verification. We tested the performance using the most comprehensive setting ("G-BoW+TC+AGV") with different *n* values. Using a hierarchical strategy with 32×32 grids as the finest resolution for AGV and any level above it for the G-BOW matching and translation compensation, there are 4 options for *n*: 2×2, 4×4, 8×8 and 16×16. Fig. 9 compares the effect of these four options on accuracy and runtime. For both document images and natural images, the results show that *n*=4×4 results in the highest accuracy. Furthermore, the runtime for *n*=4×4 is just slightly higher than that for *n*=2×2, and consistently lower than those of the other two options.

4.4 Variants of Vocabularies

The vocabulary should have sufficient descriptive ability to accurately distinguish a wide range of images. The descriptive ability of a vocabulary is determined by the set of training images. Therefore in this part we test the performance of our method using three vocabularies constructed from different training images: 1) VoB-Doc-Oxb, which uses EMM-ICME2K and EMM-Oxford5K as training images; 2) VoB-Doc-Dis, which uses EMM-ICME2K and 5k randomly selected images from EMM-Flickr200K as training images; 3) VoB-Dis, which uses 7k randomly selected images from EMM-Flickr200K as training images. The size of all the three vocabularies is set to 100k and the accuracy and runtime are evaluated based on the "G-BoW+TC+AGV" approach. The results are summarized in Table 2.

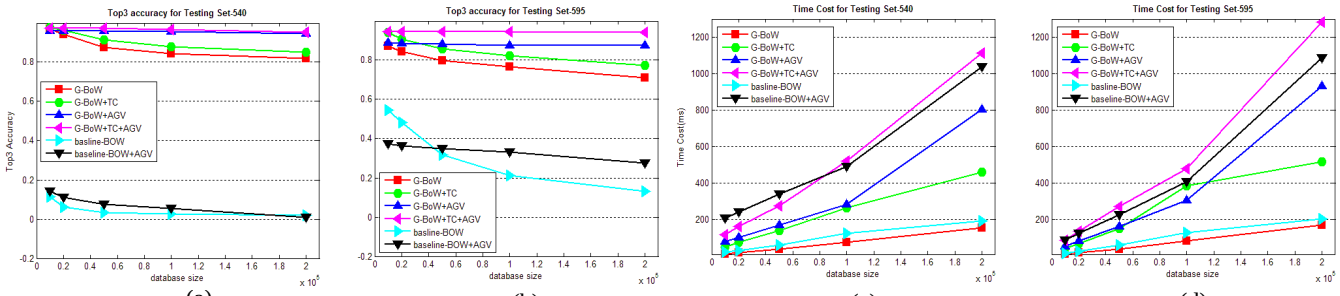
We made the following observations: 1) For testing set-540, there is an approximate 43% drop in accuracy when using the vocabulary VoB-Dis, which was trained solely based on natural images. This result indicates that in order to achieve high accuracy for certain types of images, images with a similar nature must be included in the training set. 2) For testing set-595, the accuracy remains consistent for all three vocabularies. This result implies that excluding the target images (i.e. source images used for testing) from the training set for constructing the vocabulary does not hurt the accuracy for certain types of images as long as there are sufficient training images of similar nature. This property is very important for real-world applications.

Table 2. Top3 Accuracy under different vocabularies. The vocabulary size is 100k and the dataset size is 100k+.

VoB. / Testing Set	Testing Set-540	Testing Set-595
VoB-Doc-Oxb	96.7%	94.5%
VoB-Doc-Dis	97.4%	92.3%
VoB-Dis	53.9%	94.5%

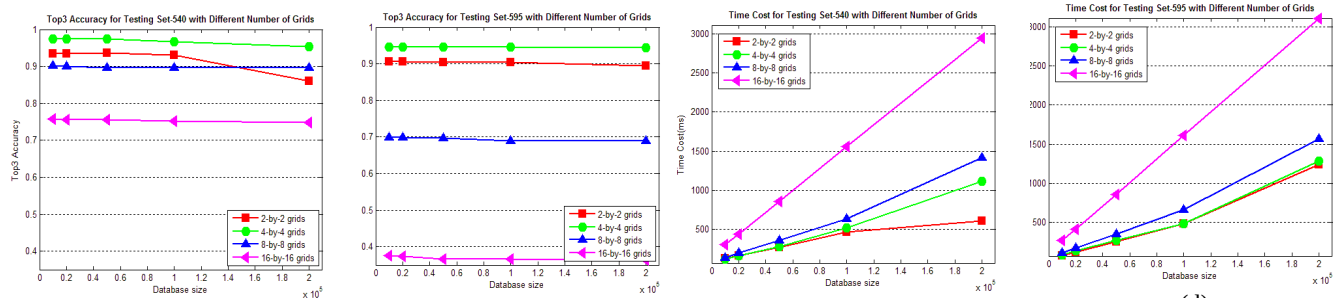
5. CONCLUSION AND FUTURE WORK

We present a scalable EMM identification system which utilizes visual features within EMMs to link dynamic media with a static paper document via camera-phones. *"Injection"* and *"Approximate Global Geometric Consistency"* are two unique constraints for EMM identification. Taking into account these two constraints, we propose a novel image matching scheme, which significantly outperforms the baseline method. The success of our approach relies on three key steps. First, translation compensation, which roughly aligns indexed images with a query image, greatly reduces the translation-caused errors and lays the groundwork for the subsequent grid-based process. Second, by utilizing the *"AGGC"* constraint, the grid-bag-of-words matching, which places coarse grids over the image and finds matches that occur in a common grid, removes lots of unqualified matches at an early stage. Third, approximate geometric verification, which conducts correspondences refinement at a fine-grid level, effectively and efficiently enforces the *"injection"* constraint and further increases the accuracy. Moreover, the proposed hierarchical encoding/decoding strategy minimizes the computational and memory cost of online correspondences searching.



(a) (b) (c) (d)

Figure 8: Performance on our dataset using 4×4 grids for G-BoW matching and 32×32 grids for AGV. (a) and (b) are Top3 Accuracy for Test Set-540 and Test-Set-595 respectively. (c) and (d) are Time Cost for Test Set-540 and Test-Set-595 respectively



(a) (b) (c) (d)

Figure 9: Performance on our dataset with different number of grids for G-BoW matching. (a) and (b) are Top3 Accuracy for Test Set-540 and Test-Set-595 respectively. (c) and (d) are Time Cost for Test Set-540 and Test-Set-595 respectively

In this EMM application, there is a practical limit on the amount of rotation and scaling for the user-captured image with respect to the index target image. Utilization of this additional constraint can be investigated to further improve the performance and accuracy of EMM identification. In addition, global features which are robust to the restricted geometric changes and photometric changes can be investigated to address the scalability for handling datasets with millions or even billions of images.

6. REFERENCES

[1] Wu, Z., Ke, Q.F., Isard, M., and Sun, J. Bundling Features for Large Scale Partial-Duplicate Web Image Search, *Proceedings of CVPR'09*

[2] Lazebnik, S., Schmid, C., Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proceedings of CVPR'06*

[3] Chum, O., Perdoch, M., and Matas, J. Geometric min-Hashing: Finding a (Thick) Needle in a Haystack, *Proceedings of CVPR'09*

[4] Jegou, H., Douze, M., and Schmid, C. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search, *Proceedings of ECCV'08*

[5] Lowe, D. Distinctive image features from scale-invariant keypoints. *IJCV*, 20:91–110, 2003.

[6] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. Object Retrieval with Large Vocabularies and Fast Spatial Matching. *Proceedings of CVPR'07*

[7] Qiong, L., Chunyuan, L., Lynn, W., Tony, D., and Bee, L. Embedded Media Marker: Linking Multimedia to Paper. *Proceedings of ACM Multimedia'10*

[8] Qiong, L., Chunyuan, L., Lynn, W., Tony, D., and Bee, L. Embedded Media Markers: Marks on Paper that Signify Associated Media, *Proceedings of IUI'10*.

[9] Sivic,J. and Zisserman, A. Video Google: A Text Retrieval Approach to Object Matching in Videos. *Proceedings of ICCV'03*.

[10] Barcode.http://en. Wikipedia.org/wiki/Barcode.

[11] Erol, B., Emilio, A., and Jull, J.J. HOTPAPER: Multimedia Interaction with Paper using Mobile Phones. *Proceedings of Multimedia'08*, pp. 399-408

[12] SnapTell. http://www.snaptell.com/

[13] Kooaba. http://www.kooaba.com/

[14] K. Grauman. Matching Sets of Features for Efficient Retrieval and Recognition. Ph.D. Thesis, MIT, 2006.

[15] Nister D., Stewenius H. Scalable Recognition with Vocabulary Tree. *Proceeding of the CVPR'06*.

[16] Donoser, M. and Bischof, H. Efficient Maximally Stable Extremal Region (MSER) Tracking. *Proceeding of CVPR'06*.

[17] Hare, J., P. Lewis, L. Gordon, and G. Hart. MapSnapper: Engineering an Efficient Algorithm for Matching Images of Maps from Mobile Phones. *Proceedings of Multimedia Content Access: Algorithms and Systems* II pp.