Unsupervised Ensemble Ranking: Application to Large-Scale Image Retrieval

Jung-Eun Lee, Rong Jin and Anil K. Jain[1]
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan, U.S.A.
{*leejun11, rongjin, jain*}*@cse.msu.edu*

***Abstract*—The continued explosion in the growth of image and video databases makes automatic image search and retrieval an extremely important problem. Among the various approaches to Content-based Image Retrieval (CBIR), image similarity based on local point descriptors has shown promising performance. However, this approach suffers from the scalability problem. Although bag-of-words model resolves the scalability problem, it suffers from loss in retrieval accuracy. We circumvent this performance loss by an ensemble ranking approach in which rankings from multiple bag-of-words models are combined to obtain more accurate retrieval results. An unsupervised algorithm is developed to learn the weights for fusing the rankings from multiple bag-of-words models. Experimental results on a database of $100,000$ images show that this approach is both efficient and effective in finding visually similar images.**

***Keywords*-Near-duplicate image retrieval, Bag-of-words models, Tattoo images, Ensemble ranking**

I. Introduction

Recent years have witnessed a dramatic increase in the amount of image/video data available on the Web, calling for efficient tools for browsing and searching of large image databases. For instance, Flicker, a well-known photo sharing site, hosts more than 3 billion images, with over 2.5 million new images uploaded to its database everyday [1]. To meet this demand, various methods are being developed for Content-based Image Retrieval (CBIR) to efficiently index and match images based on their visual content.

Although CBIR is inherently a difficult problem due to the gap between low-level image features and high-level semantics [2], CBIR techniques have been effective for *near-duplicate* image detection problems [3], [4], [5]. In particular, image similarity based on local image features, e.g. Scale Invariant Feature Transform (SIFT) descriptors [6], has shown the most promise for near-duplicate image retrieval problem [4], [7], [9], [10], [11]. But, this approach suffers from the scalability problem due to its requirement of linear scan of the entire image database. Bag-of-words model [12] addresses the scalability issue by clustering SIFT features into a small number of clusters. By treating each cluster center as a visual word in a codebook, bag-of-words model represents each image by a histogram of visual words. Despite its encouraging performance [4], [9], [10], [11], [12], there is loss in retrieval accuracy of the bag-of-words model when compared to the approaches that directly use the SIFT features.

In this paper, we have developed an image retrieval system that is not only accurate but is also scalable to large image databases. To this end, we propose an efficient ensemble ranking approach: each ranker within the ensemble is based on a different bag-of-words representation of SIFT features; an unsupervised learning algorithm is developed to learn the weights used to combine multiple rankers. Experimental results based on $1,000$ image queries to search a database of $100,000$ images show that our system is efficient as well as effective for finding near duplicate images.

II. Image retrieval for large databases

We first present image matching based on SIFT, and bag-of-words model. We then introduce the ensemble ranking method.

A. SIFT based image matching

Scale Invariant Feature Transform (SIFT) [6] is a well-known and robust local feature based approach used for object recognition. Previous studies have shown that SIFT based image representation is more effective for near-duplicate image retrieval than global visual features (e.g., color, texture and shape) [7]. SIFT extracts repeatable characteristic feature points at multiple image scales and resolutions, called keypoints, and generates descriptors representing the texture around the points. Given the extracted keypoints, the similarity between two images is determined by the number of matched keypoints, i.e., pairs of keypoints from two images that are separated by a small Euclidean distance. To improve the accuracy of keypoint matching, geometric constraints are used to reduce false matchings [3]. More details of SIFT based image matching can be found in [6].

B. Bag-of-words model

One limitation of the SIFT based matching is that it does not scale well to large databases because it has to

[1]Anil K. Jain's research was partially supported by World Class University (WCU) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.950

compute the similarity between the query image and every image in the database. Motivated by the success of text retrieval techniques [13], [14], the main idea of bag-of-words model [12] is to vector quantize (using *K*-means clustering) the collection of keypoints extracted from all the images into clusters which form the visual "words". Each image is then represented as a fixed length histogram of visual word frequency. Using the bag-of-words model, the image retrieval problem gets converted to a text retrieval system. Using the indexing technique, a text retrieval system is able to efficiently identify a subset of images that share similar visual content with the query without a linear scan of the database, leading to efficient image retrieval.

C. Unsupervised ensemble ranking

One limitation of the bag-of-words model is its loss in retrieval accuracy compared to image matching based on SIFT features. We improve the accuracy of the bag-of-words model by exploring the technique of ensemble ranking. The key idea is to combine a number of different image representations, i.e., multiple bag-of-words models in our case, for more accurate image retrieval. We construct multiple bag-of-words models by using different initializations of the K-means clustering in the construction of visual words. The remaining question is how to learn the optimal weights to combine the rankings computed from different bag-of-words models.

Various techniques have been proposed to learn the optimal weights for combining multiple rankings, including SVM [15], [16], [17], boosting [18], neural network [19] and semi-supervised [20] approaches. In these learning procedures, the training data usually consists of a number of queries, and each query is associated with a list of objects or labels. The relevance of these objects is manually judged by human subjects. One difficulty with the supervised learning approaches is that manual relevance judgments are not only expensive to acquire but also biased by the opinions of human subjects. The proposed system overcomes this difficulty by directly employing SIFT keypoint matching results for relevance judgment without any human intervention. If a retrieved image is within the top-N rank in SIFT keypoint matching, we regard the image as a *relevant* image, otherwise as an *irrelevant* image. This *unsupervised ensemble* ranking problem is formulated below.

Let $Q = \{q_i, i = 1, ..., N_q\}$ denote a collection of N_q query images in the training set, where q_i is associated with a list of N_i database images $D^i = (I^i_1, ..., I^i_{N_i})$ defined as the top-N_i similar images by SIFT keypoint matching. We assume that the first r images in the ranking list D^i, denoted by $D^i_r = \{I^i_1, ..., I^i_r\}$, are relevant and the remaining $D^i_{ir} = \{I^i_{r+1}, ..., I^i_{N_i}\}$ images are irrelevant.

Let $G = \{g_1(\cdot), ..., g_m(\cdot)\}$ denote the ensemble of m ranking functions, where each function is the mapping $g_i(I, q) : X \times X \rightarrow \mathbf{R}$. The goal of ensemble ranking is to combine the ranking functions in G to produce a ranking list that is better than any individual ranking function. In its simple form, the combined ranking function, denoted by $f_{\mathbf{w}}(\cdot)$, is expressed as $f_{\mathbf{w}}(I, q) = \sum_{k=1}^{m} w_k g_k(I, q)$, where w_k is used to weight the importance of the ranking function $g_k(\cdot)$. For the convenience of presentation, we define

$$\mathbf{s}^i_j = (s^i_{j,1}, ..., s^i_{j,m}) = (g_1(I^i_j, q_i), ..., g_m(I^i_j, q_i)),$$

and rewrite $f_{\mathbf{w}}(I^i_j, q_i)$ as $f_{\mathbf{w}}(I^i_j, q_i) = \mathbf{w}^\top \mathbf{s}^i_j$.

We adopt the Ranking SVM method [17] to learn the combination weights $\mathbf{w}$. The basic idea is to decompose a ranking list into a set of ordered example pairs and find the weights that are consistent with most of the pairs. Given a query q_i and two images $I^i_j \in D^i_r$ and $I^i_k \in D^i_{ir}$ for an ideal combination, one would expect I^i_j to be ranked before I^i_k, which implies the following constraint

$$\mathbf{w}^\top (\mathbf{s}^i_j - \mathbf{s}^i_k) \geq 1$$

By collecting constraints from all the queries, we have the following optimization problem for finding the optimal combination weights $\mathbf{w}$

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N_q} \sum_{j \in D^i_r} \sum_{k \in D^i_{ir}} \ell(\mathbf{w}^\top (\mathbf{s}^i_j - \mathbf{s}^i_k))$$

where $\ell(z) = \max(0, 1 - z)$ is the hinge loss, and C is a regularization parameter that is determined by a cross validation procedure.

III. Experiments

We verify both the efficiency and the efficacy of the proposed unsupervised ensemble ranking for large-scale tattoo image retrieval.

A. Tattoo image retrieval

A tattoo is a pattern imprinted onto the skin that has been found to be useful by law enforcement agencies for identifying a victim without any identity document or a suspect using a false identity. Another important application of tattoos is criminal identification since many gangs have a very distinctive tattoo which is used as a sign of gang membership and for intimidating others. Law enforcement agencies routinely photograph and catalog tattoo images with manually annotated class labels in the ANSI/NIST-ITL 1-2007 standard [21]. A tattoo search involves matching the label of a query tattoo with labels associated with tattoos in a database. This approach has many limitations: subjectivity in annotation, handling new tattoo types, and limited performance due to large intra-class variability in complex tattoo images (see Figure 1). We have developed a CBIR

Figure 1. Large intra-class variability in tattoo images. All the four images shown here belong to SYMBOL category.

Figure 2. Duplicate images: (a)-(d) show two different images of the same tattoo. Note the large variability in different images of the same tattoo.

system whose goal is to find tattoo images in the database that are near-duplicates of the query tattoo image (see Figure 2). Although our goal is near-duplicate detection, tattoo image retrieval is substantially more challenging than other application domains because of the large variation in the visual appearance of the same tattoo [3], [7].

We have access to ~$64,000$ tattoo images (640×480 color images) from the Michigan Forensics Laboratory. All the tattoo images were cropped to extract the foreground and suppress the background. A small fraction of the images in the database (~5%) are duplicates of the same tattoo (see Figure 2). These duplicates are introduced in the database due to multiple arrests of the same person or the multiple photographs of the same tattoo taken at the booking time. To evaluate the retrieval performance of our CBIR system, one of the duplicates is used as a query to retrieve the other duplicate(s) in the database.

B. Experimental setup

In order to verify the capability of our CBIR system for large scale databases, we increased the number of images in the tattoo database to $100,000$ by adding randomly selecting about $36,000$ images in the ESP game data set [22]. The retrieval experiments were done in a leave-one-out fashion in which $1,000$ tattoo image queries were searched against a gallery of $100,000$ images. We manually verified that these queries have at least one duplicate in the database.

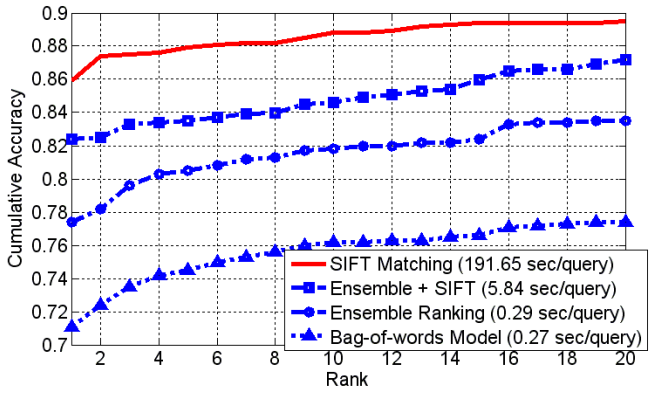

Figure 3. Retrieval Performance

The SIFT feature extraction is done offline and the total number of SIFT keypoints extracted from all the $100,000$ images in the database is ~7.8 million. Each point is represented in a 128-dimensional feature space. We applied the hierarchical K-means clustering algorithm [23] to quantize these points to build a bag-of-words model with $50,000$ visual words (number of clusters K is set to $50,000$). The clustering is also done offline which takes about 2.2 hours on Xeon 3.2 GHz, 32 GB RAM machine. A text retrieval system based on *tf-idf* weighting scheme is used to perform the image retrieval. Ten different bag-of-words models were constructed, each with $50,000$ visual words based on different initializations of K-means clustering. The number of bag-of-words models and number of visual words in the ensemble model were selected empirically. The training set consisted of randomly selected 100 images from the $1,000$ query image set. For each of these 100 images, two relevant and 10 irrelevant images were found based on SIFT keypoint matching. Learning the weights in ensemble ranking was done offline; the average learning time was 87 seconds on Intel Core 2, 2.4 GHz, 8 GB RAM processor.

We adopt the Cumulative Matching Curve (CMC) [24] as the evaluation metric. This measure shows that the probability of a matching, i.e. y-axis, if one looks at the first *N* images of the database, i.e., x-axis.

C. Experimental results

Figure 3 compares the retrieval accuracies and average matching time per query of SIFT keypoint matching, bag-of-words model, ensemble ranking and combination of ensemble ranking and SIFT keypoint matching. Among these methods, as expected, the SIFT keypoint matching shows the best retrieval performance, with 85.8% rank-1 and 89.4% rank-20 retrieval accuracies. But, the average retrieval time per query is ~191 seconds (on Intel Core 2, 2.66 GHz, 3 GB RAM processor) for a database containing $100,000$ images. Although there is a performance loss by applying the bag-of-words model (71.1% rank-1, and 77.4% rank-20),

Figure 4. Retrieval examples. Each row shows a query with the number of keypoints, top-8 retrieved image and the associated matching score (no. of matching keypoints). Note that four duplicates were found in the database for queries 1 and 2, and three duplicates for query 3.

Table I
COMPARISON OF ENSEMBLE RANKING FOR DIFFERENT NUMBER OF IMAGES RETURNED BY THE BAG-OF-WORDS MODEL.

No. selected images	1K	2K	3K	4K	5K
Rank-20 Acc. (%)	87.7	88.3	88.9	89.4	89.9
Ret. Time (s/query)	5.8	7.3	9.8	12.0	14.7

the average retrieval time per query is dramatically reduced from 191.65 sec to 0.27 sec. The proposed unsupervised ensemble ranking technique not only preserves the computational efficiency, 0.29 sec/query, but it also improves the accuracy of the best single bag-of-words model: 77.3% rank-1 and 83.5% rank-20 accuracies. A combination of ensemble ranking and SIFT keypoint matching, i.e., using the SIFT keypoint matcher to re-rank the first 1,000 tattoo images retrieved by the ensemble ranking algorithm, shows very similar performance to SIFT keypoint matching with 82.9% rank-1 and 87.7% rank-20 accuracies, with an average retrieval time of only 5.84 sec. Three retrieval examples are shown in Figure 3.

The main advantage of the ensemble ranking algorithm is that it is extremely effective in pruning the large image gallery. Additional experiments (see Table I) show that if the ensemble ranking algorithm returns 5,000 images from the database, the combination method outperforms (89.9%) the SIFT matching (89.4%) at rank 20 with only 14.7 sec average matching time per query.

IV. SUMMARY

We have presented an efficient ensemble ranking approach for large scale image retrieval. Multiple rankers are designed based on different bag-of-words representation of SIFT features, and then combined using weights learned from an unsupervised learning algorithm. Since the typical benchmark data sets such as the Oxford building data set is not large enough to validate the proposed method, we report results on a database of 100,000 images (tattoo images plus a subset of ESP images).

While the proposed system performs well in identifying duplicates for a given query image, its performance is highly dependent on the quality of the query images. When the quality of the query is poor, (i.e. faded tattoos or tattoos covered with hair), it is hard to extract distinctive features, leading to significantly lower retrieval accuracy. Figure 5 shows poor quality queries, and the images retrieved by the system. We plan to examine various techniques for image enhancement for more reliable tattoo image retrieval.

REFERENCES

[1] Reuters. "Flickr to map the worlds latest photo hotspots", 2007. http://www.reuters.com/article/technologyNews/idUSHO94233920071119.

[2] T. Pavlidis. "Limitations of content-based image retrieval". In ICPR, 2008.

Figure 5. Retrieval examples with poor quality tattoo image queries. In each row, the first image is the query (no. of keypoints is shown in parenthesis). The next seven images are the top 7 retrieved images with their scores. The last image shows the true match with match score and the rank at which it is retrieved.

[3] A. K. Jain, J.-E. Lee, R. Jin, and N. Gregg. "Content-based image retrieval: An application to tattoo images". In ICIP, 2009.

[4] Y. Ke, R. Sukthankar, and L. Huston. "Efficient near duplicate detection and sub-image retrieval". In ACM Multimedia, 2004.

[5] B.Wang, Z. Li, M. Li, and W. Ma. "Large-scale duplicate detection for web image search". In ICME, 2006.

[6] D. Lowe. "Distinctive image features from scale invariant keypoints". In IJCV, 1999.

[7] J.-E. Lee, A. K. Jain, and R. Jin. "Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification". In Biometrics Symposium, 2008.

[8] V. Lepetit, P. Lagger, and P. Fua. "Randomized trees for real-time keypoint recognition". In CVPR, 2005.

[9] D. Nister, and H. Stewenius. "Scalable recognition with a vocabulary tree". In CVPR, 2006.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. "Object retrieval with large vocabularies and fast spatial matching". In CVPR, 2007.

[11] V. Lepetit, P. Lagger, and P. Fua. "Randomized trees for real-time keypoint recognition". In CVPR, 2005.

[12] J. Sivic, and A. Zisserman. "Video Google: A text retrieval approach to object matching in videos". In ICCV, 2003.

[13] G. Salton, and M. J. McGill. "Introduction to modern information retrieval". McGraw-Hill, Inc., 1986.

[14] T. Joachims. "Text categorization with support vector machines: Learning with many relevant features". In ECML, 1998

[15] T. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. "Adapting ranking SVM to document retrieval". In SIGIR, 2006.

[16] R. Herbrich, T. Graepel, and K. Obermayer. "Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers". MIT Press, 2000.

[17] T. Joachims. "Optimizing search engines using clickthrough data". In ACM SIGKDD, 2002.

[18] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. "An efficient boosting algorithm for combining preferences". In Journal of Machine Learning Research, 2003.

[19] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. "Learning to rank using gradient descent". In ICML, 2005.

[20] S. C. Hoi and R. Jin. "Semi-supervised ensemble ranking". In AAAI, 2008.

[21] ANSI/NIST-ITL 1-2007: Standard Data format for the Interchange of Fingerprint, Facial, & Other Biometric Information, 2007.

[22] ESP Game. http://www.gwap.com/gwap/gamesPreview/espgame/.

[23] M. Steinbach, G. Karypis, and V. Kumar. "A comparison of document clustering techniques". In KDD Workshop on Text Mining, 2000.

[24] H. Moon and P. J. Phillips. "Computational and performance aspects of PCA-based face recognition algorithms". Perception, 2001.