Learning Metrics for Diversity in Instance Retrieval

Vidyadhar Rao, Ajitesh Gupta, Visesh Chari, C.V. Jawahar
Center for Visual Information Technology, IIIT Hyderabad, INDIA
Email: vidyadhar.rao@research.iiit.ac.in, ajitesh.gupta@students.iiit.ac.in, visesh@gmail.com, jawahar@iiit.ac.in

***Abstract*—Instance retrieval (IR) is the problem of retrieving specific instances of a particular object, like a monument, from a collection of images. Currently, the most popular methods for IR use Bag of words (BoW) features for retrieval. However, a prominent problem for IR remains the tendency of BoW based methods to retrieve near-identical images as most relevant results. In this paper, we define *diversity* in IR as variation of physical properties among most relevant retrieved results for a query image. To achieve this, we propose both an ITML algorithm that re-fashions the BoW feature space into one that appreciates diversity better, and a measure to evaluate diversity in retrieval results for IR applications. Additionally, we also generate 200 hand-labeled images from the Paris dataset, for use in further research in this area. Experiments on the popular Paris dataset show that our method outperforms the standard BoW model in many cases.**

I. Introduction

Instance-level image retrieval algorithms have gained recent prominence because of their applicability to two main areas, image recognition for product search like retail products [1] [2] and localization [3] [4]. The task of searching in an image database for specific "instances" of an object or subject is called instance retrieval (IR). For example, when searching for images of "Maruti car", a generic image search might retrieve various cars like Maruti, Toyata, BMW, etc. IR algorithms, on the other hand, retrieve images of various models like "Maruti Celerio", "Maruti Swift" etc. IR methods are expected to perform under several physical constraints like varitions in viewpoint of camera, time of day, camera zoom etc.

The success of IR algorithms usually depends on the low-level image features, such as color, texture, and shape, that represent the visual content present in the images. The most popular image representation is the Bag of Words (BoW) model [5]. A typical BoW pipeline for representing images is composed of the following steps: (i) **extracting** the local features from each image, (ii) **encoding** the local features to the corresponding visual words and (iii) performing **spatial binning**. Initially, a large set of local features are extracted from a training image corpus. These features are clustered to divide the local feature space into informative regions (called "visual words") and the collection of the obtained visual words is the visual vocabulary. Feature extraction is carried using the popular SIFT [6] which is designed to capture appearance and local image structures that are invariant to image transformations such as translation, rotation, and scaling. Next, in the encoding step, the local features of an image are assigned to the nearest visual word's centriod (in Euclidean distance) and a histogram of visual words is generated. Finally, spatial information is encoded by dividing the image into several (spatial) regions, compute the encoding of each region and concatenating all the resulting histograms. Thus, in IR, when a query image is given, one computes the SIFT features and encodes the visual information in the form of a histogram and retrieves relevant images that are close in the Euclidean distance.

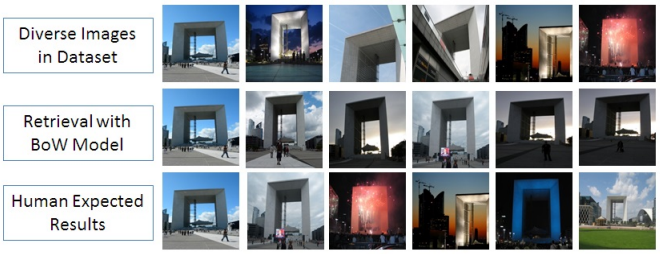

Fig. 1: Sample images of the Paris dataset for the monument "La Grande Arche de la Dfense". Top row shows the diverse images present in the dataset. Middle row shows a sample query image, and corresponding retrieved results using a BoW model. Bottom row shows the human expected diversity in results. (*Images best viewed in color*)

However, when a database has many similar images, IR methods result in near identical images at the top. This is beacause of two properties: Firstly, local features like SIFT are more adept at identifying near identical images and often confuse between different views of the same image and different but similar looking images. Such a differentiation requires higher order features to be computed. Secondly, these approaches do not penalize duplicate results aggressively. Therefore, the retrieved results are more homogenous with little diversity. We define diversity in IR as accurate retrieval of instances that show variations in physical properties like geometry and illumination.

Further more, in the BoW methods the distance metric, often pre-defined, used for image similarity is detremental to accuracy and diversity of the results. This limits the capacity of IR algorithms, because they usually assume that the distance between two similar objects is smaller than the distance between two dissimilar objects. This assumption may not hold, especially in the case of IR when the input space is heterogeneous i.e., diverse in visual content. For instance, the outdoor images (like monuments) are most effected by natural light, position from the which images are captured, and camera zoom that is intrinsic property of an image. Product search might have other properties like occlusion, but this is out of scope of this work.

We illustrate these characteristics with an example in Figure 1. Given a dataset containing several distinct views of a monument (La Grande Arche de la Defense in Paris, first row, Figure 1), BoW based algorithms [5] [7] typically retrieve near similar results for a query image (second row, Figure 1), even when the database itself contains diverse images. It can be easily seen that users searching for images of La Grande

978-1-4673-8564-0/15/$31.00

Arche de la Defense, might better appreciate the set of results shown in the third row of Figure 1, because its diversity in viewpoint, camera zoom, time of day etc. gives much better visual understanding of the monument itself. We thus make the case that diversity is an important characteristic for an IR algorithm to have.

In this work, we show that the key to encoding diversity is to find appropriate distance metric which allows for variations these physical properties. In particular, we consider a Mahalanobis distance function whose general form is given by,

$$d_A(x, y) = (x - y)^T A(x - y) \tag{1}$$

where, A is symmetric, positive semi-definite matrix. If $A = I$, the above equation is same the popular Euclidean distance metric. We look at the development and evaluation of methods for retrieving diverse images with respect to viewpoint of image capture, time of day, and camera zoom.

II. Proposed Method

As discussed earlier, one of the more important choices to make in IR is the distance metric used for retrieval. In contrast to the Euclidean distance metric, we propose an approach to learn the Mahalanobis distance metric from the data such that diversity in IR can be increased while still retrieving accurate results.

Learning distance metric from available domain has attracted much interest in recent studies [8] [9] [10]. The domain information is usually cast in the form of two pairwise constraints: must-link and cannot-link constraints. The must-link constraints enforce smaller distances for the pair of "similar" objects, and cannot-link constraints enforce large distances for the pair of "dissimilar" objects. The optimal distance metric is found such that majority of these pairwise constraints are satisfied. Our goal is to learn this distance metric (A), under certain physical constraints, to improve diversity in IR.

A. Metric Design

In this work, we use a popular metric learning approach called Information theoretic metric learning (ITML) [11]. ITML algorithm uses an information-theoretic cost model which iteratively enforces pairwise similarity/dissimilarity constraints, yielding a learned Mahalanobis distance metric, A. The Mahalanobis distance is a bijection to a Gaussian distribution with its covariance set as an inverse of A.

Exploiting this bijective property, ITML poses the metric learning problem as a convex optimization of a relative entropy between a pair of Guassian distributions with unknown A and the identity I or A_0 a prior knowledge about the inter-point distances, under simple distance similar(S)/dissimilar(D) constraints.

$$\begin{aligned} \min \ & D_{ld}(A, A_0) \\ \text{s.t.} \ & A \succeq 0 \\ & d_A(x_i, x_j) \leq u \qquad (i, j) \in S \\ & d_A(x_i, x_j) \geq v \qquad (i, j) \in D \end{aligned} \tag{2}$$

where, $D_{ld}(A, A_0) = tr(AA_0^{-1}) - \log det(AA_0^{-1}) - d$; v and u are large and small values, respectively. Solving Eq.(2) involves repeatedly projecting the current solution onto a single constraint, via an update:

$$A_{t+1} = A_t + \beta_t A_t (x_{i_t} - x_{j_t})(x_{i_t} - x_{j_t})^T A_t, \tag{3}$$

In the equation, x_{i_t} and x_{j_t} are the constrained data points for iteration t, and β_t is a projection parameter computed by the ITML algorithm. This formulation regularizes the optimization problem so as to seek a metric that satisfies the given constraints and is closest to the Euclidean distance.

Note that, for a pair of points, the distance computation in Eq.(1) can also be realized by first performing a linear transformation $\mathcal{X} \rightarrow \mathcal{T} = A^{\frac{1}{2}}\mathcal{X}$ and by computing the L^2 or Euclidean distance for the pair in $\mathcal{T}$. This linear transformation makes similar data points in $\mathcal{X}$ closer together and dissimilar data points farther apart in $\mathcal{T}$, and yields more computationally efficient pairwise computation.

Adapting this property, we treat the ITML's result A as a post feature transformation and evaluate it with different qualities of images like geometric and illuminance constraints. Once this feature transformation is performed, we simply solve the problem of nearest neighbor retrieval to report the relevant images. We demonstrate that the learned metric improves the diversity in the retrieval.

B. Constraint Generation

Metric learning can be seen as a data-driven transferring of semantic information from the class labels to input feature space. In this work, physical properties are assigned as class labels (refer Figure 2) while BoW forms the feature space. The semantic information is represented in the form of similar/dissimilar constraints over a pair of images. The standard distance metric learning involves pairs of images that are randomly sampled from a database. However, in the IR, we need to visually identify images that are similar/dissimilar to the query image.

In order to learn a diversity metric for BoW features, we have to define the constraints $d_A(x_i, x_j) \leq u$ or $d_A(x_i, x_j) \geq v$ for a pair of feature vectors x_i and x_j, corresonding to images that are similar and dissimilar, respectively. In the section III, we describe how the labels are obtained based on the visual aspects of the images. Using these labels, we can formulate the constraints in terms of similarity and dissimilarity between the feature vectors.

Algorithm 1: Diverse Retrieval Using Metric Learning

Input: $\mathcal{X} = \{x_1 \ldots, x_n\}$, where $x_i \in R^d$, a query $q \in \mathbb{R}^d$ and k an integer. $A_0 = I$, is prior about the inter-point distances. S, D are similar and dissimilar constraints.

1 $A \leftarrow$ ITML($\mathcal{X}$, A_0, S, D u, v) *// Learn metric*
2 $\mathcal{X} \rightarrow \mathcal{T} = A^{\frac{1}{2}}\mathcal{X}$ *// Apply linear transformation*
3 $\mathcal{R}_q \leftarrow \phi$ *// Retrieved images*
4 **for** $i \leftarrow 1$ ***to*** k **do**
5 $\quad t^* \leftarrow \text{argmin}_{(t \in \mathcal{T})}(\|q - t\|^2)$
6 $\quad \mathcal{T} \leftarrow \mathcal{T} \setminus t^*$
7 $\quad \mathcal{R}_q \leftarrow \mathcal{R}_q \cup t^*$

Output: $\mathcal{R}_q$, set of retrieved images

To summarize, our algorithm executes in three phases: i) perform metric learning using ITML, Algortihm 1 in line (1), to find appropriate metric and ii) transform the initial feature space to $\mathcal{T}$ using $A^{\frac{1}{2}}$, Algortihm 1 in line (2), and iii) find the k nearest neighbors, Algortihm 1 in lines (4-7), to report the set of retrieved images, $R_q \in \mathcal{X}$.

Throughout the algorithm, several variables are used that are specific to the quality of diversity. The essential control variables that direct the behaviour of the algorithm are: i) the choice of u and v in the ITML and ii) the number of constraints, $|S| + |D|$. See Algorithm 1 for a detailed description of our approach.

C. Diversity in Instance Retrieval

In order to perform an IR, we first extract SIFT [6] features from the input query image and compute visual words using the cluster centers of the database to be searched. In our experiments, we extract 100 visual words using the popular VLFeat library [12]. We set the variables v and u in Eqn.(2) to the 97^{th} and 3^{rd} percentiles of the distribution of pairwise Euclidean distances within the dataset, respectively.

We randomly sample 100 pairwise constraints from a pool of annotated images to learn the distance metric and, apply the transformation on the basic BoW features as discussed in Section II-A. As a result, the matching procedure using the learned distance metric takes the same time as the BoW method. We next include results of some of the queries peformed on these techniques based on traditional BoW model and the proposed metric design model for monument retrieval.

III. Experimental Results

A. Experimental Setup

1) Datasets: We use the Paris dataset [13] which consists of approximately 6K high quality (1024 X 768) images of monuments in Paris like La Defense and Pantheon. Note that this collection of paris images is considered to be a challenging dataset. Since the images are not tagged based on monument visibility, we manually annotated 200 images with 12 labels in the following categories: viewpoint (frontal, up, down, left, right), camera zoom (zoomed in, zoomedout, normal), time of day (morning, afternoon, evening and night). Figure 2 shows labels for a sample monument image in the Paris dataset.

2) Evaluation Criteria: There is no evaluation metric that seems to be universally accepted as the best for measuring the performance of methods that aim to obtain diverse retrieval [14]. Diversity necessarily depends on the collection over which the search is being run [15] [16]. Diversity also depends on a system's performance at basic ad-hoc retrieval i.e., how many images are relevant to the user query. Therefore, similar to precision and recall, there is a need to balance between accuracy and diversity in the retrieval. In this work, we keep a balance between accuracy and diversity by maximizing the harmonic mean of these two criteria. We believe that this performance measure is suitable for different kinds of diversity and helps us empirically compare different methods.

Accuracy: We measure the accuracy of the retrieval in terms of the proportion of relevant images (to the given query) in the retrieved results, aggregated over 50 trails.

Diversity: We measure diversity in terms of the entropy as $-\Sigma_{i=1}^{m} s_i \log s_i$, where s_i is the fraction of images of i^{th} tag, and m is the number of possible labels, aggregated over 50 trails.

B. Quantitative Results

To empirically evalaute the methods, we pick 50 random query images and retrieve results for these queries from the 200 labeled images. We use the labels of the top-5 results to compute accuracy and diversity scores. As discussed above, we show in Table I the overall performance measure as the harmonic mean of accuracy and diversity. We report results for viewpoint, time of day and camera zoom diversity.

TABLE I: Paris Dataset: Comparison of the retrieval performance for BoW and learned metrics using top-5 results. V-Viewpoint, T - Time of Day, Z- Zoom, Div-Diversity. H is the harmonic mean of accuracy with their respective diversity scores. Notice the best peformances are marked in **bold**.

Method	Accuracy	V-Div	H-V	T-Div	H-T	Z-Div	H-Z
BoW [5]	0.817	**0.412**	**0.511**	0.537	0.625	0.384	0.445
ITML [11]	**0.822**	0.391	0.495	**0.592**	**0.652**	**0.434**	**0.474**

TABLE II: User Study: Results averaged over 210 queries, answering which method produced more useful results.

Method	BoW	Our Approach	Tie
User Preference	84/210 = 40%	**97/210 = 46.19%**	29/210 = 13.81%

In our results, we observe an improvement in the accuracy of the retrieval using ITML. Notice that ITML outperforms BoW model in terms of h-score. This demonstrates the effectiveness of using metric learning to obtain both relevant and diverse set of manument images. In order to measure diversity, we use the distribution of the histogram labels (in Figure 3), with an equal distribution over all labels being the most desirable result.

Notice how ITML improves the "night" and "morning" labels by supressing the "afternoon" and "evening" labels. We also see a rise in "ZoomOut" label with a drop in "ZoomIn" label. It is important to notice that the diversity with respect to viewpoint is low for ITML approach (see Table I column 2), and this can also be observed in the rise of spikes at "Frontal" and "Up" labels for ITML approach in Figure 3.

C. Qualitative Results

The first rows of Table III give a visual representation of the top 5 retrieved images given the query images (shown in the first column). Note how the retrieved results are visually very similar to the query image in many aspects like appearance, viewpoint, zoom and even to some extent the time of day. This highlights the problem that we alluded to earlier, about the absence of diversity in results with traditional BoW model. As can be seen in Table III, our approach shows a greater visual diversity in the retrieved images. These visual results convincingly prove the ability of learning metrics (from pairwise constraints) can be helpful to improve the diversity by as much as 5% (in the case of time of day and camera zoom, refer Table I column 5) while still retaining similarity among the results.

D. Evaluating Human Expectations

We evaluate the utility of our approach based on testimonials from 14 different users randomly selected for trails. We asked them to rate 5 queries by pointing out which among IR method between BoW and our approach gave the most relevant results. We the averaged results for the 210 queries i.e., 14 users X 5 images X 3 criteria. Table II shows that users in general rated our approach superior to the BoW based IR approach.

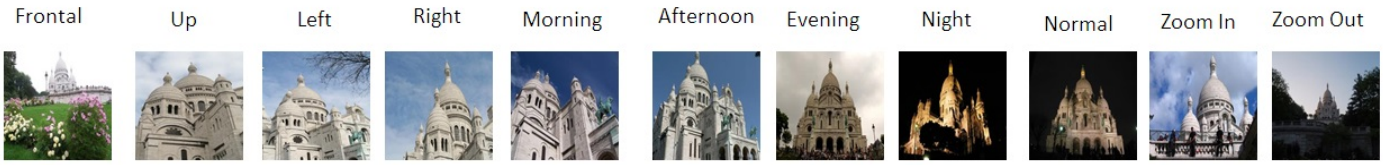

Fig. 2: Images with labels for the "Sacre Coeur" monument in the Paris dataset. (*Images best viewed in color*)

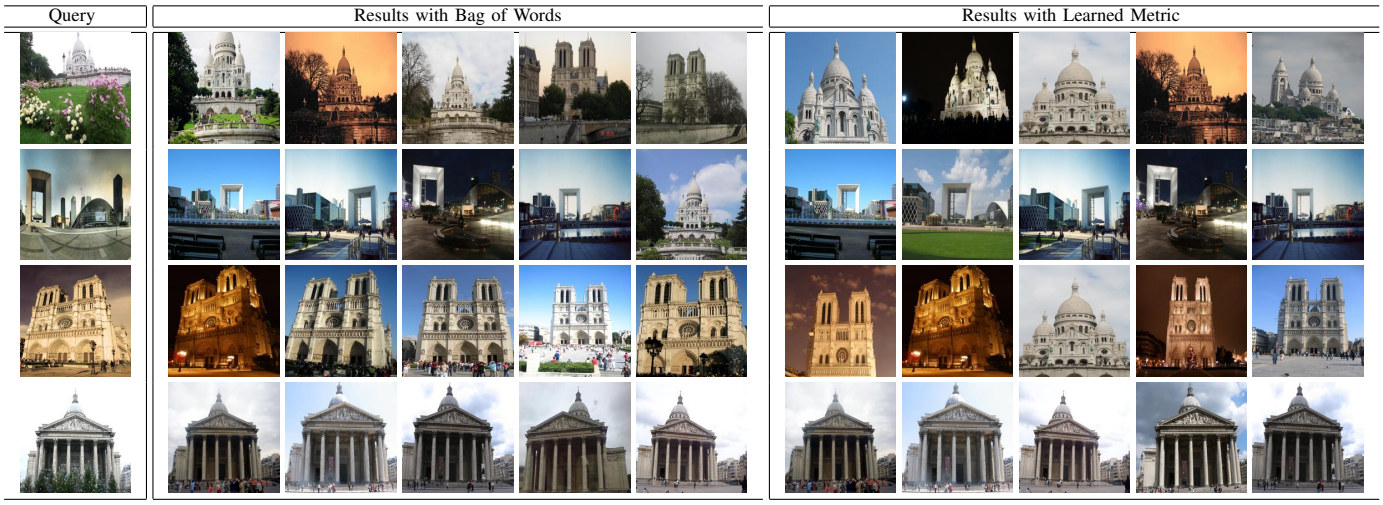

TABLE III: Five pairs of retrieval results from the Paris dataset. Top-5 candidates are shown for visual comparison between BoW and Learned metric based approaches. For each query (column one from top to bottom), accuracy for BoW and ITML Methods are $\{0.6, 0.8, 1, 1\}$ and $\{1, 1, 0.8, 1\}$, respectively. (*Images best viewed in color*)

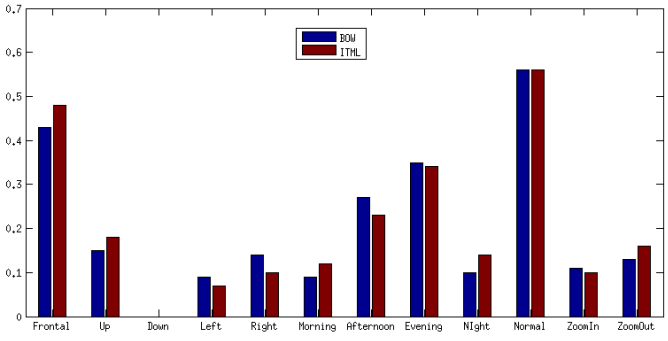

Fig. 3: Histogram of labels over 50 queries for BoW and ITML based retrieval algorithms. Notice the improvements in the "Morning", "Night" using ITML approach and also note the rise in "Zoom Out" label with a drop in "ZoomIn" label. (*Image best viewed in color*)

IV. Conclusions

This paper proposed a metric learning-based diverse IR method and presented a systematic experimental comparison with traditional bag of visual words model. Although retrieving visually similar images is arguably the most obvious application where metric distance learning plays an important role, we showed its application to diverse IR where a good distance metric is essential in obtaining competitive performances.

References

[1] M. George and C. Floerkemeier, "Recognizing products: A per-exemplar multi-label image classification approach," in *ECCV*, 2014.

[2] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Mobile product image search by automatic query object extraction," in *ECCV*, 2012.

[3] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *CVPR*. IEEE, 2013.

[4] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *CVPR*. IEEE, 2012.

[5] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

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

[8] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *ICML*. ACM, 2004.

[9] J.-E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *CVPR*. IEEE, 2008.

[10] A. López-Méndez, J. Gall, J. R. Casas, and L. J. Van Gool, "Metric learning from poses for temporal clustering of human motion." in *BMVC*, 2012.

[11] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*. ACM, 2007.

[12] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *ACM Multimedia*, 2010.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[14] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," in *SIGIR Forum*. ACM, 2009.

[15] P. B. Golbus, J. A. Aslam, and C. L. Clarke, "Increasing evaluation sensitivity to diversity," *Information Retrieval*, 2013.

[16] P. B. Golbus, V. Pavlu, and J. A. Aslam, "What we talk about when we talk about diversity," in *Proceedings of Diversity in Document Retrieval*, 2012.