CASCADED SPARSE COLOR-LOCALIZED MATCHING FOR LOGO RETRIEVAL

Rohit Pandey[⋆]*, Wei Di*[†]*, Vignesh Jagadeesh*[†]*, Robinson Piramuthu*[†]*, Anurag Bhardwaj*[†]

[⋆] Dept. of Computer Science & Engg., University at Buffalo, Amherst, NY 14260
[†] eBay Research Labs, eBay Inc., San Jose, CA 95125

ABSTRACT

In this paper we present a framework for logo retrieval in natural images. Color-localized spatial masks are used as an alternative to computationally expensive spatial verification techniques like RANSAC. First, keypoints are detected using traditional techniques such as the SIFT detector. Local masks are defined around each keypoint that take its scale and orientation information into account. To exploit inherent color information presented in brand logos, ordered color histograms are extracted from masked regions. A separate vocabulary is constructed for both SIFT descriptors (visual word) and color histograms (color word). For faster matching during runtime, a two-stage cascaded index is designed, which maps the visual word and color word tuple to a list of relevant images. This list is finally re-ranked with BoW cosine similarity to generate relevant matches for the input query. To demonstrate the efficacy of our method, we conduct experiments on two popular logo datasets: Flickr27 and Flickr32. Our experimental results illustrate State-of-the-art performance on these datasets with added speedup and lower memory footprint in terms of response ratio.

***Index Terms*—** Logo retrieval, sparse features, color information, fast spatial verification, cascaded index.

1. INTRODUCTION

Logo retrieval from natural images is a challenging problem with potentially wide commercial applications. Most of such applications require real-time indexing and retrieval of logo images with high accuracy. One such typical application is "Shop By Brand", where a user takes the picture of any merchandise (containing brand logo) and is immediately presented with the products from the same brand. Existing logo retrieval solutions take an object retrieval based approach to solving this problem. One such popular method has been the Bag-of-Words (BoW) paradigm [1]. In a typical BoW representation, localized descriptors (e.g. SIFT,SURF) are clustered into a vocabulary of visual words which is used to encode each image as a histogram over visual words. Often an inverted index is also built for faster lookup during query time. Local descriptors in general are robust to occlusion and minor distortions. However, one of their major limitations is lack of spatial information captured when used with vanilla BoW based encodings. To address this problem, often a spatial verification technique such as RANSAC is applied. This technique acts as a post-processing routine on top of BoW model returned image list, i.e. the first retrieved images are pruned to satisfy spatial consistency. Though RANSAC like approaches perform well in practice, they are computationally expensive. Real-time logo retrieval applications that have a fixed computing budget cannot use such expensive operations and hence fail to incorporate such spatial verification techniques. In this paper, we propose a novel spatial verification technique especially suited for logo retrieval applications.

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

Fig. 1. Examples of color-localized keypoints with spatial masks drawn around them (best viewed in color).

Our proposed technique is fast, efficient and is also motivated by presence of strong color components in real-world brand logo images.

A number of techniques have been proposed to improve the performance of Bag-of-Words methods which include variants over popular descriptors such as SIFT[2], RANSAC methods[1] and very recently using RootSIFT and query expansion based techniques[3]. Few works have focused on spatial re-ranking for large scale retrieval such as Phibin *et al.*[4] and Cao *et al.*[5]. Color has also been used as an important cue in object retrieval in form of color correlograms[6]. However, most of the logo related works have only used these approaches in a typical logo recognition system where the goal is not to classify a given logo image into known classes. Romberg *et al.* proposed a technique that uses both edges and triangles formed with local features [7]. For faster retrieval, they also proposed a bundle min-hashing technique [8].

Our motivation comes from the observation that logos contain specific patterns that are designed to be prominent as well as discriminative in order to convey a significant and distinct brand value. Regions containing such patterns usually correspond to high contrast areas of the image. Hence local descriptors such as SIFT are especially suited to these tasks. However, one consistent information which is usually ignored in such descriptors is the color layout information presented in brand logos. This information present multiple advantages from a feature extraction perspective - (i) they are computationally inexpensive and easy to compute, (ii) they capture

a global property of image sub-regions that complements the local properties captured by SIFT-like descriptors, (iii) A large number of brands are associated with distinct colors (i.e. pepsi is associated with blue color whereas coke with red) which makes color an important feature for brand logo recognition. In this paper, we propose to use this information to design a localized spatial color histogram feature on top of BoW encoding of logo images. Our contributions in the paper are as follows:

- We propose a novel color-localized image representation that captures some spatial color information around the local region of each visual word.
- We propose a two-level cascaded indexing process that integrates both color and SIFT-like information into a single indexing framework. This has an additional advantage of a much faster lookup time over traditional BoW and spatial verification techniques (i.e. RANSAC).
- We present State-of-the-art results on publicly available logo retrieval datasets *Flickr27* and *Flickr32*, for which we report nearly 14% relative improvement over existing benchmarks.

2. RETRIEVAL FRAMEWORK

2.1. Bag of Words

We build our retrieval framework upon the popular bag of words (BoW) model. First keypoints are detected using a difference of gaussians (DoG) local extrema detector. SIFT descriptors are computed from keypoints in each of the training images and stored. Let us represent each such descriptor x_σ^θ to be d-dimensional vector where σ, θ represent the scale and dominant angle of the descriptors respectively. These stored descriptors are then clustered using k-means to form the descriptor vocabulary or visual word $\mu = [\mu_1, \mu_2, ..., \mu_{K_1}]$ with a chosen size K_1. This vocabulary is now used to encode each training image $I = [I_1, I_2, ..., I_n]$ (n is the total number of images in the training set). The encoding process uses nearest-neighbor schema $NN(x_\sigma^\theta)$ that maps the d-dimensional input descriptor x_σ^θ to its nearest d-dimensional visual word μ_i where $1 \leq i \leq K_1$. Thus, each descriptor x_σ^θ is replaced by a visual word id $1 \leq i \leq K_1$. An inverted index is built by storing each visual word id i as key and the set of training images I_m closest to μ_i as value which is represented as:

$$IND_{BoW}(i, \{I_m : I_m = NN(\mu_i)\}) \quad (1)$$

2.2. Sparse Color-localized Keypoints (SparCl)

SIFT descriptors extracted from all keypoints in the image are representative of the whole image in general. However, for the task of logo retrieval we are particularly interested in identifying those keypoints which correlate strongly to logo-like information. One such information that we propose to use is color cues present in logos. Hence, in addition to SIFT descriptors, we also extract color signatures from all keypoints and use it to select Sparse Color-localized (Sparcl) keypoints that activate strongly to both SIFT and color information presented in logo regions of interest detected by the DoG local extrema detector. This is achieved by constructing a localized square mask or a patch $P(x_\sigma^\theta)$ where the diagonal of the mask is along θ and is of size σ. To capture spatial information around this local region, we construct a square spatial mask with 4 regions as shown in fig. 1. The diagonal of the square lies along the direction of the key point and the regions are the 4 symmetric triangles formed by the 2 diagonals of the square. We set the order of the triangles to a clockwise direction starting from the diagonal pointed along the direction of the key point.

This localized mask is essentially an ordered set of regions placed in some local area of the image. Information is extracted from each region (i.e. triangle) and stored in the defined order to form a new feature set for local spatial verification. As described, size and orientation of the mask depends on the scale and direction of the detected key point. By fixing the size and orientation of the mask to scale and orientation of descriptor, we ensure scale and rotation invariance of features and also weigh the contribution of each descriptor accordingly. The shape and number of the regions inside mask may vary but it is important to include multiple regions inside the mask for feature computation since its their relative ordering that captures localized spatial information about the logo image.

Once we have our localized spatial mask with multiple regions, we extract features from each region. For the task of logo retrieval, we choose RGB color histograms. Using color not only captures the inherent color information present in brand logos but also serves as a cheap and efficient feature that can be quickly computed. This step generates 4 RGB histograms (one from each triangle) which are then concatenated in the order determined by the mask (clock wise from the direction of the key point in our case) and stored as a larger spatial color histogram.

These stored descriptors are then clustered using k-means to form the color vocabulary or color word $\gamma = [\gamma_1, \gamma_2, ..., \gamma_{K_2}]$ of a chosen size K_2. An encoding process similar to BoW is performed that replaces each descriptor y_σ^θ by a color word id $1 \leq j \leq K_2$. An inverted index is built by storing each color word id j as key and the set of training images I_n closest to γ_j as value which is represented as:

$$IND_{CoL}(j, \{I_n : I_n = NN(\gamma_j)\}) \quad (2)$$

2.3. Cascaded Index

To perform fast matching over both visual and color words, we build a cascaded index represented as:

$$IND_{Bow,Col}(< i, j >, I_l : I_l = NN(\mu_i), I_l = NN(\gamma_j)) \quad (3)$$

This cascaded index takes a tuple $< i, j >$ as key where i is a visual word id and j is the color word id. The value of this tuple is a set of images I_l such that the closest visual word to any image in set I_l is i and the closest color word to any image in set I_l is j. Thus in this way we add an additional localized spatial verification layer on top of the BoW based inverted index. The maximum possible number of keys in this cascaded index is $K = K_1 \times K_2$ which could be prohibitively large. However in practice, these keys are highly sparse in nature, and the resulting cascaded index has only few non-empty keys. Our experiments show that roughly only 7.5% of the keys contain non-empty image sets. Such sparse nature can be attributed to the lower probability ($\frac{1}{K_1} \times \frac{1}{K_2}$) of each descriptor in the image being assigned a particular $< i, j >$ tuple.

2.4. Retrieval and Ranking

The cascaded inverted index ensures quick retrieval of relevant training images as follows.

$$\begin{aligned} I_r = \{ & I_r^{i,j} : \forall x_\sigma^\theta \in I_q, \\ & NN(BoW(x_\sigma^\theta)) = i,\ NN(Col(x_\sigma^\theta)) = j, \\ & IND_{Bow,Col}(< i, j >, I_r^{i,j}) \neq \varnothing \} \end{aligned} \quad (4)$$

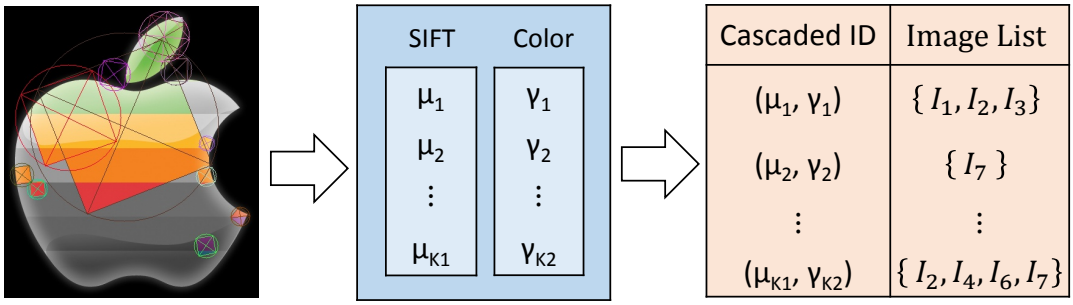

Fig. 2. Pipeline of cascade multi-inverted indexing for logo matching.

where I_r is the retrieved relevant image set by this two-level cascaded indexing.

Given a test or query image, we first extract SIFT descriptors as well as spatial color histograms from it. They are replaced by their corresponding visual word id i and color word id j. Given the $< i, j >$ tuple as key, the cascaded index ensures the retrieval of only those training images which have seen both a given visual word id as well as a color word id at the same key point. This process is also shown in fig. 2. Additionally, there is no significant computational overhead of this method on the bag of words model. In our experiments we found that our unoptimized code takes 0.34 seconds on an average for constructing spatial masks and extracting features on the high-resolution Flickr32 images. Once we have a list of relevant training images we re-rank them according to cosine similarity between the visual word histogram of the test image and each of the training images. The same vocabulary size used for building the index is also used to calculate the bag of words histograms. The average precision is then calculated on this final re-ranked list.

3. EXPERIMENTS

3.1. Dataset

For our experiments, we use two real world logo databases: Flickr32 and Flickr27. The Flickr32 consists of 32 different logo brands collected from Flickr [1], including brands like Adidas, Aldi, Apple and Becks. Flickr27 [2] has 810 training images, in total 27 logo classes/brands (30 images for each class). The query set consists of 270 images, among which 135 images contain logos, with 5 from each of the 27 annotated classes. The other 135 images do not depict any logo class. In our experiment, we only use images with logos presented as queries.

3.2. Performance Evaluation

Mean average precision(mAP) is chosen as a performance measure for retrieval. This takes into account the relevance of the retrieved results as well as how far apart they appear in the list. It is calculated for the re-ranked retrieved list as follows:

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{5}$$

where Q is the total number of queries, $AveP$ is the average precision of a given query.

[1] http://www.multimedia-computing.de/flickrlogos/
[2] http://image.ntua.gr/iva/datasets/flickr_logos/

We also employ the "Response Ratio"(RR) [8] as an indicator of the efficiency of the cascaded inverted indexing. It represents the ratio of the number of retrieved images to total size of database. The lower the response ratio, the fewer images are in the result list, which can help to reduce the computational complexity of the post-processing. Thus, a lower RR implies better efficiency of the search.

3.3. Baseline

For baseline, we chose standard bag of words model without any enhancements or post processing steps. Descriptors are first computed and stored for each of the training images. They are then clustered using k-means to form the vocabulary of visual words. Two kinds of descriptors namely, standard SIFT descriptors and the Opponent SIFT descriptors which encode RGB color information in the image, are used. In essence, the Opponent SIFT is the concatenation of the 128 dimension SIFT descriptor calculated from the RGB channels of the image separately to form a 384 dimension descriptor. Given the generated vocabulary, each descriptor is assigned with a word id, indicating its the closest cluster. An inverted index is built by storing word ids along with the training images for which they were seen in. At testing stage, all key points from the query image is detected and each associated descriptor is mapped to the vocabulary to get the word id. Relevant images associated with these word-id that are seen in the query images are retrieved and ranked. The ranking is based on cosine similarity between the bag of words histograms of the query image and each of the retrieved ones.

Vocabulary sizes were varied from 20k to 200k for Flickr32 and 10k to 100k for Flickr27. The mAP results for two datasets are shown in Table 1 and Table 2. As expected, the results are improved with larger vocabulary size. Standard SIFT is found to be consistently better at all vocabulary sizes as compared to Opponent SIFT. We also tried Opponent SURF as an alternative, but found it to be even poorer than Opponent SIFT.

Descriptor	Vocab	mAP / RR
SIFT	20k	0.287 / 0.969
	50k	0.340 / 0.916
	100k	0.370 / 0.820
	200k	0.383 / 0.681
Opp. SIFT	20k	0.237 / 0.912
	50k	0.281 / 0.848
	100k	0.304 / 0.761
	200k	0.311 / 0.652

Table 1. Performance of bag of words baseline on Flickr-32: Mean Average Precision (mAP) and Response Ratio (RR)

Descriptor	Vocab	mAP / RR
SIFT	10k	0.312 / 0.663
	20k	0.335 / 0.567
	50k	0.369 / 0.394
	100k	0.432 / 0.208
Opp. SIFT	10k	0.277 / 0.585
	20k	0.304 / 0.504
	50k	0.348 / 0.346
	100k	0.422 / 0.180

Table 2. Performance of bag of words baseline on Flickr-27: Mean Average Precision (mAP) and Response Ratio (RR)

3.4. Performance of proposed retrieval framework

For our proposed algorithm, we also test on both datasets with varying vocabulary sizes for both SIFT descriptors and localized color descriptor. For Flickr32 dataset, we vary the visual word vocabulary from 20k to 200k, while varying the localized color vocabulary from 500 to 5000. On the smaller Flickr27, we vary visual word vocabulary from 10k to 100k while varying localized color vocabulary from 100 to 1000. The retrieval results can be seen in Tables 3 and 4. We found interesting observation that it appears to be a trade off between the sizes of the two vocabularies. Performance is better when using higher visual word vocabularies and lower spatial color vocabularies, or using lower visual word ones with higher spatial color vocabularies.

SIFT vocab	Sp. color vocab	mAP / RR
20k	500	0.486 / 0.063
	1000	0.521 / 0.040
	5000	0.550 / 0.013
50k	500	0.581 / 0.031
	1000	0.602 / 0.020
	5000	0.578 / 0.007
100k	500	0.624 / 0.019
	1000	0.634 / 0.012
	5000	0.568 / 0.004
200k	500	0.647 / 0.012
	1000	0.646 / 0.007
	5000	0.564 / 0.003

Table 3. Performance of SparCl (BOW + Sp. color hists) on Flickr-32: Mean Average Precision (mAP) and Response Ratio (RR)

It can be seen that our framework offers a significant jump in terms of mAP from the baseline numbers. Also, it has an order of magnitude lower response ratio, indicating the proposed framework not only has higher precision, but is also much more efficient.

In Table 5, we also compare our results to the previously reported retrieval results on Flickr32 dataset [8]. Our method significantly outperforms the best result, which is achieved by using RootSIFT features, tf-idf weighting, RANSAC spatial verification and bundle min hashing. Unfortunately, for Flickr27 dataset, previous papers have only dealt with recognition. No retrieval results have been reported so far. We hope that our results provide a baseline for future retrieval studies on this dataset as well.

SIFT vocab	Sp. color vocab	mAP / RR
10k	100	0.515 / 0.090
	500	0.574 / 0.037
	1000	0.554 / 0.025
20k	100	0.538 / 0.062
	500	0.594 / 0.026
	1000	0.596 / 0.018
50k	100	0.587 / 0.032
	500	0.590 / 0.014
	1000	0.570 / 0.010
100k	100	0.590 / 0.012
	500	0.523 / 0.006
	1000	0.525 / 0.004

Table 4. Performance of SparCl (BOW + Sp. color hists) on Flickr-27: Mean Average Precision (mAP) and Response Ratio (RR)

Method	mAP
BOW, tf-idf with Root Sift (1M vocab) [8]	0.545
Bundle Min Hashing (200k vocab, 4 sketches) [8]	0.554
1P-WGC-RANSAC no LO (1M vocab) [8]	0.568
Proposed: SparCl (200k, 500 vocab)	**0.647**

Table 5. Comparison with other bench-mark retrieval methods on Flickr-32

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel logo retrieval technique based on a sparse color-localized representation of images. To ensure faster retrieval, a cascaded indexing mechanism is also proposed. Our retrieval algorithm outperforms all existing State-of-the-Art methods on standard logo datasets. We achieve significant speedup over existing RANSAC based spatial verification techniques. We also achieve higher precision with a much smaller response ratio demonstrating the low memory footprint and computational load of our approach. Our future work focuses on extending this idea to achromatic logo images (no color information).

5. REFERENCES

[1] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.

[2] David G Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[3] Relja Arandjelovic and Andrew Zisserman, "Three things everyone should know to improve object retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2911–2918.

[4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern*

Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007, pp. 1–8.

[5] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang, "Spatial-bag-of-features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3352–3359.

[6] Jing Huang, S Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih, "Image indexing using color correlograms," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 762–768.

[7] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof van Zwol, "Scalable logo recognition in real-world images," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011, p. 25.

[8] Stefan Romberg and Rainer Lienhart, "Bundle min-hashing for logo recognition," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 113–120.