Logo Retrieval with A Contrario Visual Query Expansion

Alexis Joly
INRIA Rocquencourt, IMEDIA team
Le Chesnay, 78153, France
alexis.joly@inria.fr

Olivier Buisson
INA
Bry sur Marne, France
obuisson@ina.fr

ABSTRACT

This paper presents a new content-based retrieval framework applied to logo retrieval in large natural image collections [1]. The first contribution is a new challenging dataset, called *BelgaLogos*, which was created in collaboration with professionals of a press agency, in order to evaluate logo retrieval technologies in real-world scenarios. The second and main contribution is a new visual query expansion method using an a contrario thresholding strategy in order to improve the accuracy of expanded query images. Whereas previous methods based on the same paradigm used a purely hand tuned fixed threshold, we provide a fully adaptive method enhancing both genericity and effectiveness. This new technique is evaluated on both OxfordBuilding dataset and our new *BelgaLogos* dataset.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information storage and retrieval

General Terms: Algorithms

1. INTRODUCTION

Content-based logos and trademarks retrieval in large natural image collections is of high interest for many applications, including dissemination impact evaluation, prohibited or suspicious logos detection, automatic annotation, etc. Trademarks recognition has been widely addressed by the pattern recognition community in the last decades but there is surprisingly very few works dealing with **natural images collections**. To the best of our knowledge there is no publicly available dataset to evaluate most recent object retrieval techniques on this logo application. The main singularity and difficulty of logo objects mainly come from the fact that they are usually printed on very heterogeneous supports including non rigid ones (T-shirts, shoes, etc.) and non-flat and reflecting ones (bottles, cars, etc.). The other issue is that logos are usually very small and may contain very few information. We discovered for example that even with 4000 SIFT features [6] extracted per image, a

---

[1] work funded by the European Project VITALAS

large part of annotated logos still have only 0, 1 or 2 points, which is insufficient for recognition.

The first contribution of this paper is thus the creation of a new challenging logo retrieval benchmark available to the community [2]. Its description is reported in section 2, whereas section 3 presents a *baseline* logo retrieval method and its performances on the new dataset.

The second and main contribution of the paper is a new *a contrario* adaptive thresholding method improving the effectiveness and genericity of visual query expansion methods. It is described in section 4 and experiments are reported in section 5.

2. BELGALOGOS DATASET

2.1 Images

The images of BelgaLogos dataset have been provided and are copyrighted by Belga press agency (http://www.belga.be/). The dataset is composed of 10,000 images covering all aspects of life and current affairs: politics and economics, finance and social affairs, sports, culture and personalities. All images are in JPEG format and have been re-sized with a maximum value of height and width equal to 800 pixels, preserving aspect ratio. Examples of images are illustrated Figure 1.

Figure 1: Some images of BelgaLogos dataset

2.2 Logos Annotation

The 10,000 images of BelgaLogos dataset have been manually annotated for 26 logos. Each image is labelled for each logo with 1 if the logo is actually present in the image and with 0 if it is not. A given image can contain one or several logos or no logo at all. The localization of the logo is not provided for all 10K images, but only for the queries (see section 2.3). The list of logos that were annotated is given in table 1 with an illustration of the targeted

---

[2] http://www-rocq.inria.fr/imedia/belga-logo.html

object. Logos having a bounding box with a minimum value of height and width lower than 10 pixels were not annotated.

| Logo name	Illustration	nb of images
Addidas		114
Addidas-text		47
Base		408
Bouigues		11
Citroen		46
Citroen-text		97
CocaCola		32
Cofidis		17
Dexia		494
Standard Liege		372
Eleclerc		8
Ferrari		76
Gucci		2
Kia		82
Mercedes		73
Nike		116
Peugeot		5
US President		14
Puma		54
Puma-text		22
Quick		56
Roche		2
SNCF		6
StellaArtois		15
TNT		66
VRT		9

Table 1: Annotated logos in BelgaLogos dataset

2.3 Queries

We created two pools of queries **Qset1** and **Qset2**:
Qset1 is composed of 55 internal queries, each defined by an image name and the coordinates of the logo bounding box in this image. Logos being the most frequent in the dataset (see table 1) are represented by more queries than less frequent ones.
Qset2 is composed of 26 thumbnails downloaded from Google first result page after querying [**logo** *logoname*]. The logo illustrations provided in table 1 are re-sized versions of the 26 thumbnails composing **Qset2**.

3. BASELINE METHOD AND RESULTS

3.1 Baseline method

We developed the following retrieval framework as a baseline method for BelgaLogos dataset. Let Ω be the dataset of N images I_j, $j \in \{1, \ldots, N\}$. Each image I_j is represented by a set

of n_j SIFT [6] features $\mathbf{F}_{ji}$, with corresponding positions $\mathbf{P}_{ji}$ for $i \in \{1, \ldots, n_j\}$.
Let Q be a query image (or a query image region) represented by a set of n_q features $\mathbf{F}_{qi}$, with corresponding positions $\mathbf{P}_{qi}$ for $i \in \{1, \ldots, n_q\}$. The baseline retrieval method works as follows:
STEP 1 - SIFT's matching: Each query feature $\mathbf{F}_{qi}$ is matched to the dataset thanks to an efficient approximate similarity search technique. We used the recent Multi-Probe Locality Sensitive Hashing method proposed by joly et al. [5] that allows sublinear search time with reduced memory space costs. We used $K = 300$ nearest neighbors queries with L_2 distance and a quality control parameter of $\alpha = 0.90$. When multiple matches occurred for a given query feature and a given retrieved image, we keep only the best match according to the feature distance.
STEP 2 - Stop list: All retrieved images are ranked in decreasing order of the number of found matches and we keep only the top 1000 results for the next step.
STEP 3 - geometric consistency: For each remaining image result, we compute a geometric consistency score by estimating an affine transformation model between the query and the retrieved images (as done e.g. in [4, 1]). An affine transformation model $(\mathbf{A}_j, \mathbf{B}_j)$ with 6 degree of freedom is first estimated for each image of the stop list by a RANSAC algorithm. Finally, the similarity score of an image I_j for query Q is given by the number of inliers according to the affine transformation model:

$$S_Q(I_j) = \sum_{m=1}^{M_j} \delta \left(\|\mathbf{P}_{qm} - \mathbf{A}_j \mathbf{P}_{jm} + \mathbf{B}_j\| \geq t \right) \qquad (1)$$

where $\delta(d \geq t)$ equals 1 if $d \geq t$ and 0 otherwise. t is a fixed threshold setting the position error tolerance ($t = 7$). M_j is the number of matches kept in image I_j after steps 1 and 2, $\mathbf{P}_{qm}$ and $\mathbf{P}_{jm}$ are the query and matched spatial coordinates of the p-th match in image I_j. The output is a list of images ranked in non-increasing order of the number of inliers $S_Q(I)$.

3.2 Baseline results

Retrieval results are evaluated by the Mean Average Precision metric. Average Precision is first computed for each query and the mean is finally estimated over all queries (or over all queries of a given logo to assess individual performances). Results obtained with the baseline method for **Qset1** and **Qset2** are summarized in the **baseline** column of table 3.

4. QUERY EXPANSION WITH A CONTRARIO ADAPTIVE THRESHOLDING

4.1 Visual query expansion based on geometrically verified results

In [1], Chum et al. introduced a new effective retrieval paradigm, referred to as visual query expansion by analogy to the text retrieval . The principle is that a number of highly ranked documents from the original query are reissued as a new query to improve performance. However, as mentioned by Chum et al. [1], improvements can be achieved only if no false positives (or very few) are included in the expanded query. To achieve this, Chum et al. suggested the use of a spatial verification criterion derived by thresholding a geometric consistency score very close to the one we described before (section 3). However, they do not provide a way to estimate this crucial threshold and simply suggest a fixed hand tuned threshold equal to 20 inliers. This threshold however depends on many factors including *global* factors, e.g. average number of features per

image, redundancy of the features in the dataset, parameters of the retrieval algorithm (e.g spatial threshold t of accepted inliers), etc. It also depends on factors varying for each query, such as the size of the query, the redundancy of the features in the query, the spatial distribution of the query features, etc. In this paper, we propose to solve this issue by a new a contrario adaptive thresholding method described below.

4.2 A contrario normalization of geometric consistency scores

The a contrario framework was initially proposed by Desolneux et al. [2] in order to group low-level visual features. The basic principle is to detect events in images a contrario to a random situation modelled by a *background model*. Usually, the unlikeliness of a given event is ensured by controlling the expected number of false detections. This generic approach has been applied with success to, among other things, the detection of alignments [2], contrasted edges, vanishing points, and grouping [3], or shape matching [7]. Recently, in [9], Rabin et al. proposed an a contrario matching of SIFT like features. However they aim at thresholding directly the SIFT feature distances and not a global geometric consistency as in our case. In this paper, we propose a new a contrario methodology adapted to geometric consistency scores. We first perform steps 1, 2 and 3 of our baseline retrieval framework without any changes. We then compute a new 4-th step that transforms original scores $S_Q(I)$ in new a contrario normalized scores $\hat{S}_Q(I)$.

STEP 4 - A contrario normalization: The normalization is based on an estimation of the false alarms distribution $\hat{N}_{fa}(S)$ with respect to the discrete random variable $S = S_Q(I)$. According to Equation 1, S_Q depends only on the set of M_j spatial coordinates pairs corresponding to the n_p matches found in step 1. High S_Q scores are thus directly related to the statistical dependence between the spatial positions of the query and the matched features. We thus define our a contrario background model by the probability mass function $\hat{p}_{fa}(S)$ of the variable S under the hypothesis $\mathcal{H}_0^Q$ that $\mathbf{P}_{qm}$ and $\mathbf{P}_{jm}$ are mutually independent random variables for all j:

$$\hat{p}_{fa}(S) = Pr[S_Q(I) = S \mid \mathcal{H}_0^Q]$$

The cumulative distribution function $\hat{N}_{fa}(S)$ can be obtained by:

$$\hat{N}_{fa}(S) = \sum_{s=0}^{S} \hat{p}_{fa}(s)$$

We finally keep as normalized score $\hat{S}_Q(I)$ an estimation of the results precision according to $\hat{N}_{fa}(S)$:

$$\hat{S}_Q(I) = \frac{\#\{I_j \in \Omega, S_Q(I_j) > S_Q(I)\} - N.\hat{N}_{fa}(S_Q(I))}{\#\{I_j \in \Omega, S_Q(I_j) > S_Q(I)\}}$$

In practice, we estimate the probability mass function $\hat{p}_{fa}(S)$ for each query Q by a Monte Carlo simulation. To generate independent spatial matches according to hypothesis $\mathcal{H}_0^Q$, we simply randomize the spatial positions of the query features $\mathbf{P}_{qm}$ and we keep the matched positions $\mathbf{P}_{jm}$ unchanged. More precisely, we affect to a given query feature F_{qi} a new spatial position $\mathbf{P}_{qj}$ randomly selected among the other points positions of the query. Compared to a purely uniform random generation of points position this method has the advantage to preserve some prior knowledge about the points distribution, such as bounds and principal orientations. We then simply re-compute step 3 and we estimate $\hat{p}_{fa}(S)$ by counting the number of results having a score S_Q equal to S. To limit the estimation bias due to the presence of correct images in the random results list, we keep only in the count the results having
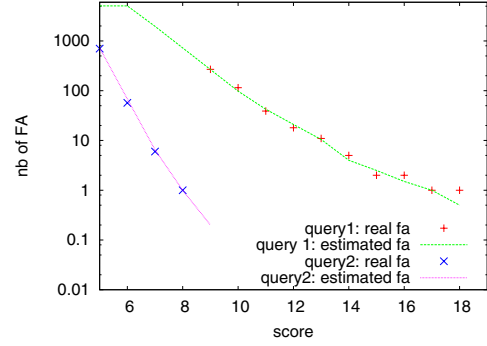

Figure 2: Real and estimated false alarms distributions for two distinct queries

a score higher than the one obtained with the normal query. To reduce the noise of the estimated distribution, we perform the Monte Carlo simulation 10 times and we average the obtained distributions.

As an illustration, Figure 2 shows the real false alarm distribution of two queries of the OxfordBuilding dataset (see experiments section 5) compared to the estimated distributions $\hat{N}_{fa}(S)$. The real false alarm distribution is computed by applying the baseline retrieval method and counting the number of real non correct results having a score S_Q higher than S. The figure shows that the accuracy of the estimated distribution is very good and that there is consistent variations between different queries.

4.3 Implemented query expansion method

The query expansion method we implemented is very similar to the transitive closure method described in [1] except that geometrically verified results are selected by a threshold $\hat{S}_t$ on our new adaptive measure $\hat{S}_Q(I)$ instead of a threshold on $S_Q(I)$. After performing steps 1, 2, 3 and 4, we keep all results having a score $\hat{S}_Q(I)$ higher than $\hat{S}_t$ and we insert them in a priority queue keyed by $\hat{S}_Q(I)$. Then, an image is taken from the top of the queue and the region corresponding to the original query region is used to issue a new query. Verified results of the expanded query that have not been inserted to the queue before are inserted (again in the order of $\hat{S}_Q(I)$). The images in the final result are in the same order in which they entered the queue. Note that setting $\hat{S}_t$ is much easier than thresholding $S_Q(I)$ since $\hat{S}_Q(I)$ corresponds to an effective estimation of the obtained precision independent of the dataset and the parameters.

5. EXPERIMENTS

5.1 Evaluation on OxfordBuilding dataset

Before applying our query expansion method to *BelgaLogos* dataset, we evaluate our new a contrario framework on *OxfordBuilding* dataset [8] (the one also used by Chum et al. [1]). We first compute the recall/precision curves of the baseline method (without expansion), with and without our a contrario normalization. Results are given Figure 3 and show that the a contrario score normalization significantly reduces the recall decrease for very high precision, which is exactly what we aim for a query expansion framework (including false alarms in the expanded queries is indeed prohibitive). The recall is even doubled for near perfect false alarms rejection.
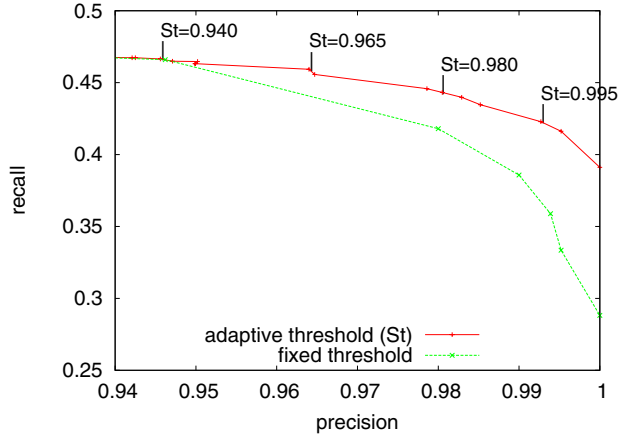
Figure 3: ROC curves for OxfordBuilding dataset with fixed and adaptive thresholds

Table 2 then gives the Mean Average Precision scores (MAP) for the baseline technique and for the two query expansion methods (with and without a contrario normalization). The expansion threshold was set to 20 for the fixed threshold method (as suggested by the authors of [1]) which corresponds to an average precision of 0.98. For our a contrario method, we thus choose a similar global precision by setting our threshold on estimated precision to $\hat{S}_t = 0.98$.

The results show that the MAP is consistently higher for most queries with up to 70% MAP improvements. Note that the improvement is limited by the fact that the query set is very homogeneous and that stronger improvements would be obtained on more heterogeneous queries and/or dataset. Moreover, we remind here that the main advantage of our technique is that it generalizes automatically to other data contrary to the fixed threshold approach. For example, applying the same fixed threshold of 20 on the **BelgaLogos** dataset don't provide any improvements whereas applying our approach with the same threshold $\hat{S}_t = 0.98$ provides significant improvements (see next section).

Query name	Baseline	Qexp fixed [1]	Qexp acontrario
all souls	51.7	83.7	**95.9**
ashmolean	43.9	75.1	**76.3**
balliol	46.7	81.2	**84.2**
bodleian	53.3	**82.2**	78.8
christchurch	63.4	71.4	**80.4**
cornmarket	**53.4**	53.3	49.1
hertford	71.3	77.4	**89.6**
keble	92.3	**92.9**	91.4
magdalen	19.7	26.52	**45.2**
pittrivers	**100**	**100**	**100**
radcliffe	72.4	87.2	**96.6**
All	60.8	75.5	**80.7**

Table 2: OxfordBuilding results

5.2 BelgaLogos results

Results obtained on BelgaLogos dataset with our new a contrario query expansion method are provided in table 3 for the query sets Qset1 and Qset2. It clearly outperforms the baseline retrieval method with better or equivalent performances for all queries, and

strong improvements higher than 50% for 37% of the queries. Weak improvements for about half of the queries seems to be mainly related to insufficient numbers of SIFT features in the corresponding queries and/or expected correct results. In future works, we plan to overcome this issue by using complementary interest points detectors or dense patches. Another perspective to improve both the effectiveness and the efficiency of the proposed query expansion method is to add a generative step that will filter the selected local features before submitting expanded queries.

Logo name	Baseline		Qexp acontrario	
	Qset1	Qset2	Qset1	Qset2
Addidas	7.8	0.7	**13.3**	0.7
Addidas-text	5.6	1.1	**7.8**	1.1
Base	14.4	38.9	**21.5**	**58.2**
Bouigues	18.2	11.3	**18.6**	**15.3**
Citroen	6.1	4.5	**38.4**	4.5
Citroen-text	5.3	0.1	**18.8**	0.1
CocaCola	23.0	0.1	**48.6**	0.1
Cofidis	26.0	55.2	**26.6**	**65.3**
Dexia	16.6	29.3	**24.0**	**51.3**
Ecusson	1.1	0.1	**5.9**	0.1
Eleclerc	78.1	74.1	**80.6**	**80.1**
Ferrari	24.7	7.5	**41.4**	**17.5**
Gucci	50.0	0.0	50.0	0.0
Kia	32.8	61.3	**67.5**	**75.6**
Mercedes	9.7	18.5	**15.0**	**19.2**
Nike	1.4	1.2	**3.5**	**2.6**
Peugeot	20.0	20.7	**20.2**	**23.2**
US President	64.3	60.3	**96.6**	**100.0**
Puma	8.6	2.2	**20.0**	2.2
Puma-text	51.6	0.7	**56.6**	0.7
Quick	24.4	39.0	**41.4**	**56.6**
Roche	50.0	0.2	50.0	0.2
SNCF	33.3	27.9	**35.4**	**33.7**
StellaArtois	32.7	31.8	**39.3**	**43.4**
TNT	22.5	2.5	**33.54**	4.4
VRT	11.1	5.8	**12.53**	**11.2**
All	20.8	19.0	**34.11**	**25.7**

Table 3: BelgaLogos results

6. REFERENCES

[1] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. of ICCV*, 2007.
[2] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *Int. J. Comput. Vision*, 40(1):7–23, 2000.
[3] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(4):508–513, 2003.
[4] A. Joly. New local descriptors based on dissociated dipoles. In *Proc. of ACM CIVR*, pages 573–580, New York, NY, USA, 2007. ACM.
[5] A. Joly and O. Buisson. A posteriori multi-probe locality sensitive hashing. In *Proceedings of ACM international conference on Multimedia*, 2008.
[6] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of Int. Conf. on Computer Vision*, pages 1150–1157, 1999.
[7] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An a contrario decision method for shape element recognition. *Int. J. Comput. Vision*, 69(3):295–315, 2006.
[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of CVPR*, 2007.
[9] J. Rabin, J. Delon, and Y. Gousseau. A contrario matching of sift-like descriptors. In *ICPR*, pages 1–4. IEEE, 2008.