Recognizing Thousands of Legal Entities through Instance-based Visual Classification

Valentin Leveau, Alexis Joly, Olivier Buisson, Pierre Letessier, Patrick Valduriez

Recognizing Thousands of Legal Entities through Instance-based Visual Classification

Valentin Leveau
INA / INRIA Zenith, France
valentin.leveau@ina.fr

Alexis Joly
INRIA Zenith, France
alexis.joly@inria.fr

Olivier Buisson
INA, France
olivier.buisson@ina.fr

Pierre Letessier
INA, France

Patrick Valduriez
INRIA Zenith, France

ABSTRACT

This paper considers the problem of recognizing legal entities in visual contents in a similar way to named-entity recognizers for text documents. Whereas previous works were restricted to the recognition of a few tens of logotypes, we generalize the problem to the recognition of thousands of legal persons, each being modeled by a rich corporate identity automatically built from web images. We introduce a new geometrically-consistent instance-based classification method that is shown to outperform state-of-the-art techniques on several challenging datasets while being much more scalable. Further experiments performed on an automatic web crawl of 5,824 legal entities demonstrates the scalability of the approach.

1. INTRODUCTION

Legal entities (such as firms, government bodies, political parties, societies, associations, etc.) are entities other than natural persons (human being) created by law and recognized as having duties and rights. It does not exist any estimation of the number of such legal entities but they are omnipresent in our all day life as well as in all media contents. Beyond their legal identity, most of them also have a corporate visual identity, that is a set of graphical rules and elements providing an organisation with visibility and recognizability (graphic charter, logotype, insignia, colors, polices, fonts, etc.). As for natural persons, it is therefore possible to recognize them automatically in visual contents in order to provide automatic annotations. This is of high interest for many applications involving huge amounts of weakly annotated image or video contents (YouTube, social media, TV archives, etc.).

Whereas legal entity recognition is considered as an important challenge in the text community (e.g. the freebase repository contains 741K organisations, 160K educational institutions or 31K sport teams), the problem has received little attention in the multimedia community. Existing works and techniques (see section 2) are limited to the recognition of few tens of classes (37 logos in BelgaLogos dataset[1], 32 logos in FlickrLogos dataset [18]), each targeted entity being modeled by a single logotype. In this paper, we generalize the problem to the recognition of thousands of legal persons, each being modeled by a rich visual identity automatically built from web images. The proposed method relies on a geometrically-consistent instance classification technique that is described in section 3 and evaluated in section 4.

MM '14, November 03 - 07 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2647868.2655038 ...$15.00.

2. RELATED WORKS

As the number of legal entities to be recognized is very large, the problem is primarily related to large-scale image classification. Existing methods and techniques for this problem are typically based on local descriptors pooling techniques (BoW [19], Fisher vectors [13], SPM [10]) and the use of efficient classifiers in the high-dimensional embedded space such as linear support vector machines [14]. An alternative is *deep convolutional neural networks* that have been recently proved to achieve similar results on large-scale image datasets such as ImageNet [9]. The problem we are targeting is however slightly different. Because the visual identity of a legal entity is actually aimed at guarantying its recognizability, it relies on visual objects with small intra-class variations (such as logos, landmarks, insignias, etc.) but in highly cluttered contexts (very small objects & weak image-level annotations). This problem has been referred as *instance classification* in a recent paper of Krapac et al. [8] and is at the crossroad between *object recognition* and *instance-level image retrieval*. The method they propose is based on a feature-wise prototype selection approach: local descriptors are all kept in their original form (without quantization) and a distance-adaptive prototype is trained for each of them in a supervised way. They report some consistent performance improvements over several state-of-the-art classification methods (including Fisher Vectors [14]).

The other family of techniques related to our work is *instance-based image retrieval* techniques [2] and in particular the ones focused on logo retrieval [18, 6]. These techniques are primarily aimed at retrieving instances of a given query object in an unsupervised way but any of them can be used for classification purposes when the search is performed on

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

a labelled set of pictures (typically by voting on the top-K retrieved images or through any other instance-based classifier). They usually rely on local descriptors quantization methods (such as hamming embedding [1], data-dependent hashing [7] or product quantizer [4]) and scalable indexing structures (such as inverted lists or hash tables). This allows matching all local descriptors one by one in the full index and favors a more precise matching of small objects in highly cluttered pictures. Using geometry in addition to this raw scheme has been proved to achieve further performance gains, either by the inclusion of weak geometry constraints in the matching phase [3] or by post-checking the geometry consistency of the matches, typically by using a RANSAC algorithm [6]. The method proposed in this paper improves such techniques in order to construct precise class-specific saliency maps and build a strong image classifier highly robust to noise and clutter.

3. PROPOSED METHOD

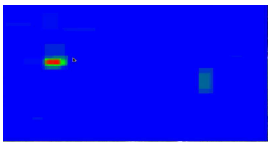
We consider a training set S of $|S|$ pictures weakly annotated with one or several labels among $|C|$ classes (corresponding to one or several legal entities presumably recognizable in the picture). Each picture is described by a set of local descriptors (SIFT [11] in our experiments) forming a reference dataset of N local features $\mathbf{x}_i$. Our instance-based geometrically-consistent classification scheme can be summarized as follows: local descriptors are extracted from the query image I_q and searched independently in the reference set using an efficient K-NN search scheme. A local geometry consistency checking is then performed at every potential region of interest using a newly introduced sliding RANSAC procedure. The resulting lists of checked patches are then back-propagated in the query image and merged in order to produce pixel-wise saliency maps for each of the retrieved label. A strong classifier is finally derived from the class-specific saliency maps through a max-pooling strategy.

Hash-based K-NN search. The approximate K-Nearest Neighbors (K-NN) of each local feature $\mathbf{x}_j^Q$ belonging to a query image I_Q are computed efficiently thanks to the hash-based multi-probe search method introduced in [5]. Its principle is to train an adaptive search model at indexing time through kernel density estimates computed on exact K-NN samples. This model is used at search time to select the most probable buckets to be visited so as to retrieve on average a fraction α of the real K-NN's (α was set to 0.80 in all our experiments). The advantage of this method compared to other state-of-the-art methods (such as PQ-code [4] or Soft-Assignment [16]) is that it allows controlling accurately the quality of the retrieved K-NN while being applicable to any quantization function and metrics.
In our case, the original scheme is transposed to the use of a more effective hash function and to the application of the Hamming Embedding principle [1]. More precisely, we used RMMH [7], a recent data-dependent hash function family, in order to embed the original feature vectors in compact binary hash codes of 128 bits (the parameter M of RMMH was fixed to $M = 32$). The $k = log_2(N)$ first bits of the hash codes are used as keys of a hash map containing the N 128-bits hash codes divided into the 2^k buckets. Still at indexing time, the search model is estimated based on the exact K-NN's in the 128-bits Hamming space rather than in the original space. It has actually be shown in [7] that RMMH somehow works as an unsupervised metric learning technique and can provide better matches in the embedded space than in the original one. At search time, the query feature $\mathbf{x}_j^Q$ is also embedded via RMMH and the resulting 128-bits hash code $\mathbf{h}_j^Q$ is passed to the probabilistic multi-probe algorithm (see [5] for more details). Once the most probable buckets have been selected, the top-K matches are computed as the K nearest hash codes according to the Hamming distance between the query hash code $\mathbf{h}_j^Q$ and the hash codes belonging to the selected buckets.

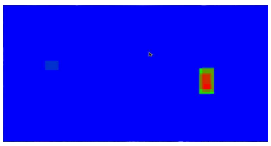
Sliding RANSAC. Post-checking the geometric consistency of the raw visual matches is an efficient strategy to filter false positives and consolidate good matches. The RANSAC algorithm has notably been successful in rigid objects retrieval [15] in particular logos [6]. Some recent variants such as MAC-RANSAC [17] provide even better results in the context of multiple localized objects by either constraining the distance of the randomized candidate pairs or by iteratively removing validated matches from the pool of candidates. But even with such improvements, a global RANSAC algorithm applied at the image-level is not adapted to the detection of very small objects in highly cluttered images for which the percentage of inlier pairs of matches can be typically lower than 0.1% of the whole set of possible pairs. Furthermore, as it is computed on the retrieved images one by one, it does not allow consolidating locally the matches from different training images.
To address these issues, we introduce a *sliding* RANSAC strategy aimed at checking the geometric consistency locally for each of the N_Q query features of the query image I_Q. More precisely, for a given local feature $\mathbf{x}_j^Q \in I_Q$, its m spatial nearest neighbors are computed so as to define a candidate region of interest to be geometrically checked in all the retrieved pictures (i.e. in the ones having some visual matches within the $m+1$ lists of K-NN's). For a given candidate region of interest and a given retrieved image, the RANSAC algorithm then works in a standard fashion (in our experiments, we used a scale-rotation-translation model and 100 iterations per local run). Note that the support of both the *random sampling* and the *consensus* phases is bounded by the set of local features belonging to the current region of interest. This allows improving the recall and the precision of the inliers compared to classical global RANSAC algorithm. The parameter m controls the locality constraint of the geometry consistency analysis. Ideally, it should fit the size of the targeted objects of interest. Too large values of m would lead to the same problem than a global RANSAC. Too small values of m would degrade the dynamic of the number of inliers and possibly miss some consistent matches. In our experiments, m was trained by cross-validation.

Class-specific geometry consistency maps. The output of the sliding RANSAC algorithm is a set of N_Q lists of consolidated results (i.e. one list per query feature $\mathbf{x}_j^Q$). Each consolidated result $R_{j,t}^Q$ is itself defined as a set of individual matches of the form $(\mathbf{x}_{j'}^Q, \mathbf{x}_{t'})$ where the $\mathbf{x}_{j'}^Q$ belong to the m spatial neighbors of $\mathbf{x}_j^Q$ and the $\mathbf{x}_{t'}$ belong to an image I_t of the training set. In order to construct saliency maps, we first associate each consolidated result $R_{j,t}^Q$ with an individual geometry consistency score $f_{j,t}^Q$ and a bounding box

(a) Raw query image



(b) Class *Base*



(c) Class *Quick*

Figure 1: Class-specific geometry consistency maps.

$B^Q_{j,t}$ in the query image. Rather than simply counting the number of inlier matches, the score $f^Q_{j,t}$ is computed as the sum of the inverse rank of the matched features $\mathbf{x}_{t'}$ (rank in the K-NN's of $\mathbf{x}^Q_{j'}$). This allows giving more importance to the most confident visual matches. The bounding box $B^Q_{j,t}$ is defined as the minimum bounding rectangle containing all the individually matched features $\mathbf{x}^Q_{j'} \in R^Q_{j,t}$.
The pixel-wise consistency score $g^Q_c(w,h)$ of a pixel (w,h) according to class label c is then computed by (i) selecting the consolidated results $R^Q_{j,t}$ whose bounding box $B^Q_{j,t}$ intercepts (w,h) (ii) grouping them according to the provenance image I_t and averaging the scores $f^Q_{j,t}$ for each group (iii) summing the averaged scores of the groups whose provenance images I_t are labeled with c. This allows voting on the number of pictures retrieved for label c and weighting each vote by an average geometry consistency in each image. Figure 1 displays two saliency maps $g^Q_c(w,h)$ computed for two distinct class labels in a single query image.

Multi-label Scoring. As illustrated by Figure 1, the saliency maps produced by the previous step could be easily used for a precise localization of the visual patterns recognized for each of the retrieved legal entity. The scope of the present paper is however only on classification so that we only use the maps to build a strong classifier at the image level. This is done by simply taking the value of the most salient pixel in each map (i.e for each returned label):

$$s^Q(c) = \max_{(\mathbf{w},\mathbf{h})} g^Q_c(w,h) \quad (1)$$

where $s^Q(c)$ is the detection score of the label c.
As shown in Figure 1, this last step also acts as a disambiguation procedure where different classes can co-occur in different images and bring geometry consistency in other class-specific saliency maps.

To decide whether a given legal entity is detected or not, a threshold τ_s is applied on the $s^Q(c)$ scores (several annotations can thus be produced for each image). To better model the density distribution over the classes, a normalization is then applied according to:

$$p^Q(c) = \frac{s^Q(c)}{\sum_{c'} s^Q(c')} \quad (2)$$

where $p^Q(c)$ is the probability estimation of the presence of the label c.

Discussion. The main line of this paper is to use online geometry consistency checking to disambiguate instance-based matches rather than training discriminative models offline. This is justified in several ways. First, our training phase is reduced to a simple indexing process with a linear time and space complexity $O(N)$. The prototype selection technique of [8] requires computing the $20,000$-NN's of each of the N features of the training set, leading to a much more important training time (over-linear in N). Concerning the memory storage, their method requires at least 8 times more RAM to store the original SIFT features. Besides, the complexity of other state-of-the-art methods making use of pooling and SVM's is typically $O(N + |C|.|S|^2)$ so that they are less scalable in both the number of classes and the number of images. Beyond scalability, our method has several other advantages including the easy management of multi-labeled images, the fine grained localisation of the recognized patterns making them highly interpretable and the possibility of dynamically inserting additional training images in an incremental way.

4. EXPERIMENTS

Evaluation metric and datasets. Classification effectiveness is measured by the Mean Average Precision (mAP), which is commonly defined as the mean of the per-class average precisions. The evaluation is performed on 3 datasets of the literature and one newly introduced dataset dedicated to the large-scale recognition of legal entities:
FlickrLogos32 [18] - 8,240 images composed of 2,240 images labelled with 32 logo classes and 6000 distractors considered as a no-logo class. Evaluation protocol is the same than [8].
BelgaLogoII [6] - 10,000 images containing 2695 instances of 37 different logos. The main difference with FlickrLogos32 is that images are multi-labeled and that that the average size of the logos is much lower. Evaluation is completed by leave-one-out queries.
Vehicles29 [8] - 10,622 images labelled with 29 car models of 7 brands (divided into 5,266 training and 5,356 test images). Evaluation protocol is the same than [8].
LegalEntities5K - 371,924 images noisy labelled with 5,824 legal entities. This dataset was automatically created by querying Google Image search engine with the entities names. The list of the entities is the union of several thesaurus found on the web and contains world-wide companies, associations, organizations and sport teams. To try limiting noise, we kept only the top-20 to top-1000 results as a function of the popularity of the tag (number of pages returned by Google). A test set was built by taking the $2,500$ of the most popular tags of the dataset. A second test set was created by intersecting the whole Flickr32 dataset with the labels of LegalEntities5K (540 test images belonging to 18 classes).

Method	FlickrL32	Vehicles29
Fisher Vectors (128x4,096)	0.866	0.497
Prototype voting [8]	0.914	0.557
Our method (S-Ransac)	**0.928**	**0.597**

Table 1: Classification performances.

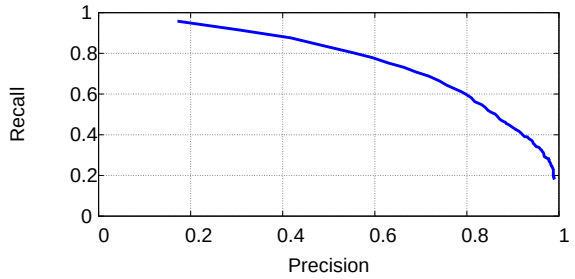


Figure 2: Recall-precision curve of our method on the multi-labeled dataset BelgaLogos.

Features and parameters setting. Local features are SIFT features [11] extracted around interest points detected by the rotation-invariant Harris-Hessian-Laplace detector [12]. The two main parameters of our method, i.e. the spatial neighborhood size m and the detection threshold τ_s were trained by cross-validation on the training sets of *FlickrLogos32* and *Vehicles29*. m was set to 5% of the number of descriptors for each query image ans τ_s was set to 5.

Comparative study. Table 1 reproduces the results of [8] and reports our own results using the same evaluation protocol. It can be seen that our method outperforms the previous baseline of [8] (and de facto the other state-of-the-art classification methods) on the two experimented datasets, whereas the training stage of our method is much more scalable. It took respectively 13 minutes and 22 minutes to index and to compute the a posteriori multi-probe search model of the 51,054,054 descriptors of the Vehicles29 training set and the 91,800,540 descriptors of the FlickerLogos32 training set (including distractors images). The good results achieved by our method on the *Vehicles29* dataset shows that it is well suited for such fine-grained image classification tasks even if it is not the primary goal of this paper.

Multiple labels. To evaluate our method in the context of multi-labeled images, we used the challenging BelgaLogos dataset. For this experiment, we randomly chose 1,000 images of the dataset that contain at least one labeled logo. Our method achieves a mAP of 82.30 which is quite impressive knowing that this dataset involves very small objects in highly cluttered contexts. To better illustrate the relevance of the produced labels for each played queries, Figure 2 displays the precision-recall curve of our method when varying the detection threshold τ_s. It shows that very high precision or recall values might be obtained depending on the applicative constraints.

Large-scale experiment on thousands of legal entities. It took 1 hour and 55 minutes to index the 500,957,407 SIFT features of the *LegalEntities5K* dataset and to compute the a posteriori multi-probe search model. Table 2 reports the results achieved on the two test tests in terms of mAP and search time. It shows that that the effectiveness of our method is still very satisfactory considering that (i) the number of classes in the training set is two orders of magnitude higher (ii) the training set was built automatically without any human validation and therefore contains a high level of noise.

Benchmark	mAP	Avg Search time
2.5K images / LegalEntities5K	**0.686**	7.4 sec
FlickrLogos / LegalEntities5K	**0.648**	6.3 sec

Table 2: Classification results and computation time on the *LegalEntities5K* dataset (Intel(R) Xeon(R) E5-2650 CPU 2.00GHz).

5. CONCLUSION

The work reported in this paper was a first attempt towards deploying real-world visual-based entity recognizers at the web scale. It did show the importance of using fine-grained geometry in this regard and demonstrated the advantage of instance-based classification methods in terms of scalability. Future works will focus on collecting and denoising training data in a more elaborated way than the proof of concept presented in this paper.

6. REFERENCES

[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*, 2006.
[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.
[3] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.
[4] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 2011.
[5] A. Joly and O. Buisson. A posteriori multi-probe locality sensitive hashing. In *ACM MM*, 2008.
[6] A. Joly and O. Buisson. Logo retrieval with a contrario visual query expansion. In *ACM MM*, 2009.
[7] A. Joly and O. Buisson. Random maximum margin hashing. In *CVPR*, 2011.
[8] J. Krapac, F. Perronnin, T. Furon, and H. Jégou. Instance classification with prototype selection. In *ICMR*, 2014.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[11] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.
[12] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *ECCV*, 2002.
[13] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
[14] F. Perronnin, J. Sánchez, and T. Mensik. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.
[15] J. Philbin. *Scalable Object Retrieval in Very Large Image Collections*. PhD thesis, University of Oxford, 2010.
[16] 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.
[17] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. Mac-ransac: a robust algorithm for the recognition of multiple objects. In *3DPTV*, 2010.
[18] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol. Scalable logo recognition in real-world images. In *ICMR*, 2011.
[19] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *CVPR*, 2003.