Interactive Objects Retrieval with Efficient Boosting

Saloua Litayem Ouertani, Alexis Joly, Nozha Boujemaa

Interactive Objects Retrieval with Efficient Boosting

Saloua Litayem
INRIA Rocquencourt, France
saloua.litayem@inria.fr

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

Nozha Boujemaa
INRIA Rocquencourt, France
nozha.boujemaa@inria.fr

ABSTRACT

This paper presents an efficient local features boosting strategy for interactive objects retrieval tasks such as on-line supervised learning or relevance feedback. The prediction time complexity of most existing methods is indeed usually linear in dataset size since the retrieval works by applying a trained classifier on the images of the dataset one by one. In our method, the trained classifier can be computed directly on the whole dataset in sublinear time thanks to distance-based weak classifiers. The idea is to speed-up drastically the prediction of each weak classifier on the whole dataset by performing approximate range queries with an efficient similarity search structure. Experiments on Caltech 256 dataset show that the technique is up to 250 times faster than the naive exhaustive method. Thanks to this efficiency improvement, we developed a relevance feedback mechanism on image regions freely selected by the user and we show how it improves the effectiveness of the retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback, retrieval models*

General Terms

Algorithms, Performance, Experimentation

1. INTRODUCTION

Contrary to usual supervised object recognition schemes, interactive object retrieval tasks consist in learning models and retrieving relevant objects in a large dataset as an **on-line** process. This includes common relevance feedback mechanisms but also other scenarios, such as submitting on-line some illustrations of a given object (e.g. crawled from the web). In such contexts, the efficiency and the scalability of the retrieval are critical and this is the main point addressed in this paper. The training time cost can also be an issue but is not addressed here since we suppose that the training set provided by the user is of relatively small size.

MM'09, October 19–24, 2009, Beijing, China.

Recent works [1, 13] address the problem of speeding-up the retrieval of top-k instances identified by SVM algorithms thanks to the use of efficient index structures. They show that interesting speed-ups from 5 to 100 times can be obtained with acceptable lost in quality. However, as pointed by [13], these methods fail to reduce the retrieval complexity when using very high-dimensional and sparse features. Their applicability is thus restricted to global visual features of moderate dimension of about 150. On the other side, most recent and effective recognition techniques [8, 5, 12] are precisely based on high-dimensional and sparse representations induced by the large number of local visual features extracted in the images. Classifiers learned on such representations are usually applied to test images one by one and the complexity in a retrieval context is intrinsically linear in dataset size.

In this paper, we explore an efficient boosting strategy that aims to reduce the retrieval complexity when using feature rich representations of images. In our method, the trained classifier can be computed directly on the whole dataset in sublinear time thanks to efficient range queries used as weak classifiers. The proposed framework is presented in section 2 and experiments are reported in section 3.

2. EFFICIENT INTERACTIVE OBJECT RETRIEVAL FRAMEWORK

Since our boosting training method is similar to the one introduced by Opelt et al. [12], we first briefly summarize it in subsection 2.1. We then introduce our new retrieval framework in section 2.2 and 2.3, section 2.2 being related to the general retrieval algorithm and section 2.3 describing how we speed up the prediction of the weak classifiers with a posteriori multi-probe LSH [7]. Finally, section 2.4 presents a relevance feedback application of the proposed method.

First of all, we introduce here some general notations. Let

$$\Omega = \{\mathbf{I}_i\}_{i\in[1,N]}$$

be the dataset of N images in which we would like to retrieve objects. And let ω be the training set of M labeled images provided to the learning algorithm for a given searched object:

$$\omega = \{(\mathbf{I}_m, l_m)\}_{m\in[1,M]} \qquad l_m \in \{-1, +1\}$$

A positive label indicates that a relevant object appears in the image. ω can be either a subset of Ω if the user provides on-line labels on some images of the dataset or an external training dataset (e.g. images crawled from the web representing a given object).

For simplicity we assume here that each image I is represented by a set of d-dimensional local features $\{\mathbf{v}_{I,j}\}$, but the technique can be easily extended to several sets of heterogeneous features as suggested in [12]. The set of all local features extracted from the

images of Ω is noted V_Ω and the set of all features extracted from the training images is noted V_ω.

2.1 Learning by boosting distance-based weak classifiers

The learning model described by Opelt et al.[12] is based on the AdaBoost algorithm [4] with a specific weak learner based on distances between training local features. As output, the learning algorithm delivers a final classifier which predicts if a relevant object is present in a new image $\mathbf{I}$:

$$H(\mathbf{I}) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})\right) \quad (1)$$

Each weak classifier $h_t(\mathbf{I})$ can be expressed as:

$$h_t(\mathbf{I}) = sign\left(\theta_t - \min_j d(\mathbf{v}_t, \mathbf{v}_{I,j})\right) \quad (2)$$

where $d(.,.)$ is a distance metric between features (typically L_2), $\mathbf{v}_t$ is a local feature of the training set V_ω selected by the t-th weak learner, θ_t is a threshold on the distance estimated by the t-th weak learner. The final classifier $H(\mathbf{I})$ is thus entirely parametrized by a set of T triplets $(\alpha_t, \mathbf{v}_t, \theta_t)$ where each pair $(\mathbf{v}_t, \theta_t)$ defines the t-th trained weak classifier.
We now briefly describe the weak learner algorithm defined by Opelt et al.[12] but we let the reader refer to the paper for more information. The input of the t-th weak learner is the set of labeled training images ω and a set of weights W_{t-1} on these images produced by the previous step of the AdaBoost algorithm:

$$W_{t-1} = \{w_m\}_{m\in[1,M]}$$

where w_m is the weight of the m-th training image. The initial weights W_0 are set to $1/M$ for all images. Relatively to the weights W_{t-1}, the weak learner selects the best weak classifier h_t among a set of possible classifiers $h_{\mathbf{v},\theta}$ constructed from all features of the training set. Formally:

$$(\mathbf{v}_t, \theta_t) = \underset{\mathbf{v}\in V_\omega, \theta}{\operatorname{argmax}} \sum_{m=1}^{M} w_m \; h_{\mathbf{v},\theta}(\mathbf{I}_m) \; l_m \quad (3)$$

where $\mathbf{v} \in V_\omega$ is a feature of the training set and

$$h_{\mathbf{v},\theta}(\mathbf{I}) = sign\left(\theta_t - \min_j d(\mathbf{v}, \mathbf{v}_{I,j})\right)$$

Solving the maximization problem of Equation 3 is then done in two steps: first by determining the best distance threshold θ for each $\mathbf{v} \in V_\omega$. Then by selecting the best feature $\mathbf{v}_t \in V_\omega$. In practice, a matrix with all minimum distances is computed before AdaBoost starts so that the quantity $\min_j d(\mathbf{v}, \mathbf{v}_{I,j})$ is known for all features $\mathbf{v} \in V_\omega$ and all training images I_m.

2.2 Retrieval algorithm with range queries

First of all, since we are interested in objects retrieval more than objects recognition we define the following ranking score rather than a binary decision as expressed in Equation 1:

$$S(I) = \sum_{t=1}^{T} \alpha_t h_t(I) \quad (4)$$

The higher $S(I)$, the higher the confidence that I contains the trained object. The baseline exhaustive approach to perform the on-line retrieval would be to compute the score $S(I)$ on all images of the dataset Ω, one by one, and then to rank the dataset by decreasing scores. The time cost of the prediction being mainly related to the number of distance computations, this would lead to a time complexity $O(TN)$, linear in dataset size.
In our case, instead of predicting the scores of the images one by one, we directly predict scores from the whole dataset Ω in a two steps process:

1. **Range queries:** we first perform T range queries in the dataset according to the T weak classifiers parameters $(\mathbf{v}_t, \theta_t)$. Each range query returns a set of features R_t such as:
$$R_t = range_\Omega(\mathbf{v}_t, \theta_t) = \{\mathbf{v} \in V_\Omega \| d(\mathbf{v}, \mathbf{v}_t) < \theta_t\}$$
2. **Prediction:** for each query results set R_t, we can construct a set of images having at least one feature in R_t, denoted as Ω_t. It is then easy to show that:
$$\forall\, I \in \Omega_t \;\; h_t(I) = 1$$
$$\forall\, I \notin \Omega_t \;\; h_t(I) = -1$$
Thus, for any image I of the dataset, the prediction score can be estimated by:
$$S(\mathbf{I}) = \sum_{t=1}^{T} \alpha_t \delta_{I\in\Omega_t} \quad (5)$$

The time cost of the whole prediction being mainly related to the number of distance computations, the cost of the second step is negligible compared to the exhaustive prediction cost. However, when using an efficient technique for the range queries step, the cost of the second step can be similar as the range query step. In practice, to reduce the prediction cost, we thus limit the scope of the retrieval to the images contained in

$$\Omega_T = \Omega_1 \cup \cdots \cup \Omega_t \cup \cdots \cup \Omega_T$$

Note, that the *missing* images are only the one having no features in all result sets R_t, i.e the ones that are labelled negatively by all weak classifiers, i.e the ones labelled negatively by the strong classifier $H(I)$.

2.3 Efficient retrieval with a posteriori multi-probe LSH

To process the T range queries of step 1 efficiently, we use an approximate similarity search structure based on Multi-Probe Locality Sensitive Hashing (MP-LSH) [7, 10]. MP-LSH methods are built on the well-known LSH technique [2], but they intelligently probe multiple buckets that are likely to contain query results in a hash table. Such techniques have been proved to overcome the over-linear space cost drawback of common LSH while preserving a similar sublinear time cost (with complexity $O(N^\gamma)$). The a posteriori version proposed by Joly et al. [7] improves the former one by defining a more reliable a posteriori model taking account some prior about the queries and the searched objects. This prior knowledge allows a better quality control of the search and a more accurate selection of the most probable buckets. In our case, the main advantage of the second method is that it allows to change the query parameters on the flight without creating a new index and to control the quality of the returned results. For each weak classifier $(\mathbf{v}_t, \theta_t)$, we thus issue a range query centered on $\mathbf{v}_t$ and with radius θ_t. By default, we set the quality control parameter α to $\alpha = 0.9$, which means that we can expect to retrieve 90% of the exact results.

2.4 Relevance feedback on freely selected image regions

Based on the previously described framework, we built a new relevance feedback application on freely selected image regions. Classical relevance feedback methods usually consist in labeling on-line positive and negative samples from a previous search result (text based query, query by example, random query, etc.). A classifier is trained on the selected samples and then applied on the whole dataset. Based on the choice of the active learning strategy, most positive or most ambiguous image results are presented to the user who can iterate again to enrich the underlying classifier. In our scheme, the user can freely select and label image regions instead of entire images. Relevance feedback strategies on image regions were already proposed in the literature (e.g [3, 11]) but most of them were based on previously segmented regions. Our scheme differs from these approaches in the sense that the user can select any region of interest he wants in an image and label it positively or negatively. Note that the efficiency provided by our method is crucial for such paradigm.

Within this new interactive mode, only the local features belonging to the selected regions will be kept in the learning stage of our method. The learning algorithm is still provided with a set of labeled training images $(\mathbf{I}_1, l_1), \ldots, (\mathbf{I}_m, l_m)$, but only the features belonging to the selected regions are included in the training features set V_ω. The effectiveness improvement obtained thanks to this interactive mode is studied in section 3.4.

3. EXPERIMENTS

3.1 Experimental setup

All experiments were computed on a 2.83GHz CPU with 8 GB RAM. As visual local features we used common SIFT features [9] with L2 distances. The evaluation benchmark is built from Caltech256 dataset [6]. It is indeed well appropriated to assess on-line and interactive retrieval techniques, with few training images per class and a large number of images. A single MP-LSH index is built offline on all SIFT features extracted from the $30,607$ images of the full dataset. As query objects, we used only 10 classes of varying complexities (according to [6]), namely: airplanes, american-flag, chess-board, golf-ball, mars, motorbikes, sunflower, swiss-army-knife, tennis-racket and tower-pisa. For each query object, we created a learning set of 295 images, composed of 40 positive examples randomly selected in the object class and 255 negative examples randomly selected from the other classes (1 per class). The $30,312$ remaining images for each query object are considered as test data. In practice, the prediction of each trained model is done on the whole dataset using the MP-LSH index. We then remove from the results list the images corresponding to the learning set and we compute the Average Precision measure for each query object. The Mean Average Precision (MAP) over all classes is then computed by averaging the 10 Average Precision scores. Time efficiency of the prediction is measured for each query object and finally averaged over all queries.

We compared our method to the original exhaustive approach of [12] classifying each image of the dataset one by one.

3.2 Objects retrieval performances

Our main goal here is to measure the efficiency gain of our method relatively to the effectiveness bias induced by the approximate similarity search technique. Table 1 reports the retrieval Times and the MAPs for both the exhaustive original approach of [12] and our new approximate efficient version. It shows that our method provides very strong efficiency improvements with a retrieval time about 250 times faster. Concerning the quality, our approximate method does not degrade the performances and even surprisingly provides significantly better results than the exact method. Our interpretation to this phenomenon is related to the bias introduced by the approximate range query search. Approximate similarity search techniques tend to miss the features that are the farthest from a given query point. Furthermore, the quality control is not guaranteed precisely for each query but only on average and queries with larger radii tend to be more degraded. The resulting effect is that the feature points that should be labeled positively by the weak classifier and which are the closest from the frontier are labeled negatively. This induces a lost in recall but also a gain in precision that seems to be predominant in the final Average Precision. This phenomenon is augmented by the fact that the proportion of negative images is much wider in the test dataset than in the training data we use. So that the weak classifiers of the exhaustive method tends to label positively a wide number of negative features.

	Exhaustive [12]		**Approximate**	
airplanes	9008.92	0.2037	35.43	0.3881
american-flag	8935.19	0.2922	35.72	0.3903
chess-board	9537.36	0.7156	33.21	0.7446
golf-ball	8908.83	0.1156	39.31	0.2361
mars	9017.57	0.1603	31.2	0.0909
motorbikes	9001.33	0.2863	34.43	0.4516
sunflower	8942.48	0.5797	32.63	0.6214
swiss-army-knife	1604.16	0.0201	31.52	0.1196
tennis-racket	8923.87	0.2266	33.08	0.2715
tower-pisa	8911.52	0.2683	40.01	0.5512
Means	**8279.123**	**0.2868**	**34.65**	**0.3865**

Table 1: Mean Average Precision and Prediction time for the 10 studied classes of Caltech256

3.3 Scalability evaluation

Our goal here is to evaluate the influence of the dataset size on the retrieval time. Figure 1 shows the average prediction time of our method and the exhaustive method for varying size of the dataset (in images number). Each evaluated size corresponds to a random subset of the Caltech256 dataset for which a new MP-LSH index was built. The log curves clearly show the sublinear complexity of our method compared to the linear complexity of the exhaustive approach. For datasets larger than 200 images, the complexity is of the form $O(N^\gamma)$ with $\gamma = 0.35$, i.e. the efficiency gain compared to the exhaustive approach is multiplied by about 20 each time the dataset is multiplied by 100.

3.4 Relevance feedback experiments

To evaluate the relevance feedback scenario, we did implement a batch feedback mechanism. The process is initialized with 29 images randomly selected from the training set, with 4 positive images and 25 negative images (in order to keep the same ratio than the previous experiments, with 40 positives and 255 negatives). A classifier is then trained and applied on the dataset, returning a ranked list of results sorted by decreasing order of $S(I)$. We then simulate an ideal user who would label negatively the top-25 negative results and positively the top-4 positive results containing an instance of the targeted object. The selected images are added to the training

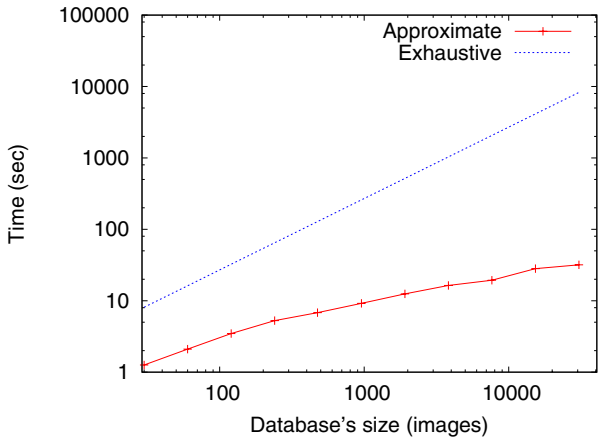

Figure 1: Retrieval time when varying dataset size

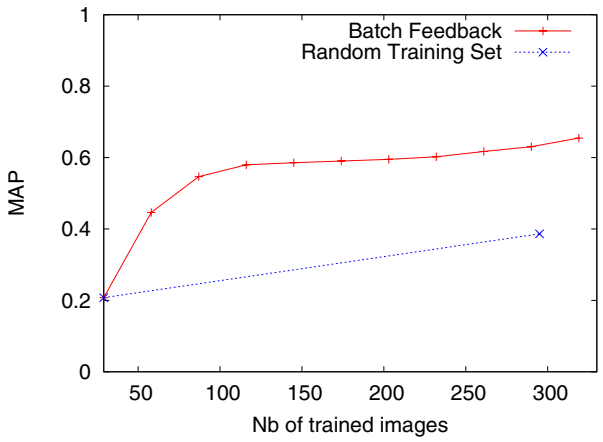

Figure 2: The Mean Average Precision for the probabilistic prediction method vs the one in the case of exhaustive prediction.

set and we iterate again. Note that a new test ground truth is built at each iteration to remove the new training images.
Figure 2 displays the MAP over all queries at each iteration. The x-axis is given in the number of trained images. The dot line corresponds to the result obtained with randomly selected training samples as done in previous experiments. The figure shows that the active training provides strong effectiveness improvements. The MAP obtained previously with 295 training images (0.38) is now reached with only 50 training images. With the same final number of training images (295), the MAP is 68% higher.

4. CONCLUSIONS AND PERSPECTIVES

In this paper we proposed an efficient local features boosting strategy for interactive objects retrieval tasks such as on-line supervised learning or relevance feedback. The main contribution is to speed-up drastically the prediction of distance-based weak classifiers by performing approximate range queries in an efficient similarity search structure. Experiments on Caltech 256 dataset show that the technique is about 250 times faster than the naive exhaustive method with surprisingly better performances. Another important contribution was to apply the proposed method to a real time relevance feedback mechanism based on freely selected image regions. Experiments show that the active learning provides significant effectiveness improvements. We think that such strategy is very promising to build accurate objects model in a semi-supervised way.
Future works will focus on several aspects: we first would like to study more in depth the effect of the weak classifier choice since our results show that approximate range queries seem to give better results than exact range queries. Since the MP-LSH technique we use allows to issue more complex probabilistic queries than range queries, we would like to train Bayesian weak classifiers instead of distance-based ones. Another perspective is to include geometry informations in the trained classifier by analysing the local geometry of the local features (scale and orientation). Finally, we would like to improve the efficiency of the learning step for large training datasets.

5. REFERENCES

[1] M. Crucianu, D. Estevez, V. Oria, and J.-P. Tarel. Speeding up active relevance feedback with approximate knn retrieval for hyperplane queries. *Int. J. Imaging Syst. Technol.*, 18(2-3):150–159, 2008.

[2] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proc. of Symposium on Computational geometry*, pages 253–262, 2004.

[3] H. Z. B. Z. F Jing, M Li. Relevance feedback in region-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):672–681, 2004.

[4] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *2nd European Conference on Computational Learning Theory (EuroCOLT'95)*, pages 23–37, 1995.

[5] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, 2007.

[6] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[7] A. Joly and O. Buisson. A posteriori multi-probe locality sensitive hashing. In *Proceedings of ACM international conference on Multimedia*, 2008.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of Int. Conf. on Computer Vision*, pages 1150–1157, 1999.

[10] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li. Multi-probe lsh: efficient indexing for high-dimensional similarity search. In *Proc. of Conf. on Very Large Data Bases*, pages 253–262, 2007.

[11] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Region-based image retrieval using an object ontology and relevance feedback. *EURASIP J. Appl. Signal Process.*, 2004:886–901, 2004.

[12] A. Opelt, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):416–431, 2006.

[13] N. Panda and E. Y. Chang. Efficient top-k hyperplane query processing for multimedia information retrieval. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 317–326, New York, NY, USA, 2006. ACM.