Large-Scale Image Mining with Flickr Groups

Alexandru Lucian Ginsca, Adrian Popescu, Hervé Le Borgne, Nicolas Ballas, Phong Vo, Ioannis Kanellos

Large-scale Image Mining with Flickr Groups

Alexandru Lucian Ginsca[1,2], Adrian Popescu[1], Hervé Le Borgne[1], Nicolas Ballas[3], Phong Vo[1], Ioannis Kanellos[2]

[1]CEA, LIST, Vision & Content Engineering Laboratory, Gif sur Ivette, France
[2]TELECOM Bretagne, France, [3]Université de Montréal
{alexandru.ginsca, adrian.popescu, herve.le-borgne, phong.vo}@cea.fr
nicolas.ballas@umontreal.ca, ioannis.kanellos@telecom-bretagne.eu

Abstract.
The availability of large annotated visual resources, such as ImageNet, recently led to important advances in image mining tasks. However, the manual annotation of such resources is cumbersome. Exploiting Web datasets as a substitute or complement is an interesting but challenging alternative. The main problems to solve are the choice of the initial dataset and the noisy character of Web text-image associations. This article presents an approach which first leverages Flickr groups to automatically build a comprehensive visual resource and then exploits it for image retrieval. Flickr groups are an interesting candidate dataset because they cover a wide range of user interests. To reduce initial noise, we introduce innovative and scalable image reranking methods. Then, we learn individual visual models for 38,500 groups using a low-level image representation. We exploit off-the-shelf linear models to ensure scalability of the learning and prediction steps. Finally, *Semfeat* image descriptions are obtained by concatenating prediction scores of individual models and by retaining only the most salient responses. To provide a comparison with a manually created resource, a similar pipeline is applied to ImageNet. Experimental validation is conducted on the ImageCLEF Wikipedia Retrieval 2010 benchmark, showing competitive results that demonstrate the validity of our approach.

1 Introduction

As predicted a few years ago [20], research in visual and multimedia recognition has strongly benefited from the availability of manually labeled large-scale image and video collections. The ImageNet representation [7] of nearly 22,000 concepts with approximately 14 million images according to a hierarchy of concepts was thoroughly exploited. For instance, a subset of ImageNet was used to train large convolutional neural networks (CNNs), leading to the design of powerful visual features such as Overfeat [21]. In the domain of large-scale recognition, another important advancement came from the use of classifier outputs as descriptors, with classemes [22] being one of the early methods of this type. The central novelty of this approach was not the use of classifier outputs as feature in itself, but that the signatures were built from many category-specific classifiers. It therefore needs a sufficiently large-scale set of annotated data.

These approaches are very promising but raise new problems, concerning, in particular, the availability of the underlying resources. Manually labeled datasets are the result of sustained effort provided by motivated communities of researchers [21], eventually supplemented with crowdsourcing [20, 7]. An important limitation of this approach is that manual annotation is a repetitive task and annotators tend to become demotivated. In addition, when conducted on a large scale, crowdsourcing has a non-negligible financial cost and dedicated funding is difficult to obtain. A promising way to circumvent the lack of annotated data is to use images shared on multimedia social networks (OSNs), such as Flickr. An advantage of this type of resource compared to formal "annotation tasks" is that data are annotated by a community of users motivated to make their content accessible.

Here, we firstly propose an approach to design semantic image features which are built on top of an automatically processed large-scale collection. Then we test these features in image retrieval; a particular attention is paid to the reduction of the noise inherent to Web collections. We address the following open and recurring research questions:
Q1 - Do features, established using automatically and manually built resources, have similar performances?
Q2 - Is it possible to build image representations whose parts efficiently convey semantic meaning?
Q3 - Can we design features which ensure a good coverage of the semantic space?
Q4 - How to build semantic representations which are at the same time compact and efficient?
Q5 - Can we learn semantic features with a resource and exploit them to mine other datasets?

Q1 is the central question addressed here. Our approach is validated only if performances obtained with the two types of resources are comparable. To our knowledge, the direct comparison of automatically and manually built large-scale resources was not properly addressed in literature. *Q2* relates to the use of semantic features as an alternative to low level features, such as bags-of-visual-words, Fisher Kernels [11] or CNN features [21]. Unlike low-level feature vectors, semantic features directly convey humanly understandable information. Consequently, they would be a promising candidate for bridging the semantic gap if they had been both precise and comprehensive, two conditions which are not yet met. For instance, the performances of meta-classes [1], which exploit a large part of ImageNet, lag behind those of Fisher Kernels, whose performances are in turn lower than those of CNN features [13]. We assume that Flickr groups are a good candidate to answer *Q3*, provided that they are properly selected and "cleaned" via image reranking. Regarding *Q4*, we note that textual documents can be searched efficiently because they are sparse and a similar property is desirable for semantic image features. Surprisingly, sparsification is not directly addressed in existing work, with the closest proxy being quantization done for more efficient signature storage [1]. *Q5* is a hot topic in computer vision and relates to transferring knowledge gained from a dataset to other datasets. Among others, it was recently tackled by [17] but not in the context of semantic features, as it is in this paper.

2 Related work

Flickr groups. Flickr groups are often used as a playground for investigating photo sharing communities. In a pioneering work, Negoescu *et al.* [16] looked at the involvement of users in groups and found, among others, that user group loyalty is generally low and most users share the same photos in different groups. Grabowicz *et al.* [10] make a distinction between topical and social groups and proposed several methods to classify a group into one of these two categories.

While informative for our purposes, such works do not tackle the exploitation of group content for image mining. Chen *et al.* [4] are among the first to exploit the visual content of groups. They use Flick group search for a set of 62 concepts and rank the returned groups based on 4 factors related to group popularity. They train dedicated SVM models for concepts and use them independently to recommend tags and groups. More closely related to our work is the idea presented in [24], where the similarity between two images is computed by leveraging the prediction scores of a set of 103 hand picked Flickr groups. Each probability is estimated using a SVM classifier trained over low-level visual features. The resulting vectors are also briefly tested in clustering and classification tasks, for which comparable results with visual features are reported.

Key differences with our work arise from the way groups are modeled. We propose several image ranking methods that improve individual classifier performance when using an initial training set of only 300 images, whereas in [24] the learning is performed on a large training set (15,000 to 30,000 images). In addition, we sparsify the features and thus enable fast retrieval over large datasets.

Semantic Image Representation. The availability of large image collections and of scalable machine learning techniques has led to a resurgence of semantic representation for image classification [14, 22]. Li et al. [14] introduced Object Bank, where an image is represented as

a scale-invariant response map of 200 pre-trained object detectors. Torresani et al. [22] also introduced a semantic representation using a fix number of hand selected binary classifiers. Each classifier is applied on the whole image input. Due to the relatively small number of visual concepts considered, early semantic representations ensured only a limited coverage of the semantic space. To tackle this issue, Bergamo and Torresani [1] learned the visual concepts of the semantic representation directly from the data. They however use 13 different features and "lift-up" each one to approximate a non-linear kernel, that is a much more costly approach than ours.

Convolutional neural networks (CNN) for Large-Scale Image Classification. CNNs have recently shown impressive image classification performances in the large-scale visual recognition challenge ILSVRC [13]. Compared to traditional low-level features such as Fisher Vector [18], the use of CNN brought down the ILSVRC error rate from 0.26 to 0.15 in 2012 and 0.11 in 2013 [13, 21]. Moreover, CNN-based feature extractors, such as *Overfeat* [21], were publicly released. These extractors provide pre-trained weights files and facilitate the extraction of features for new image collections. The outputs of their final layer are semantic image representations but they are limited to the 1,000 ILSVRC concepts, due to computational complexity of the algorithm and their need of labeled training data. Very recently, the authors of [17] and [8] exploit CNN to build mid-level features and report impressive results on various image classification datasets. The focus here is not on building new CNN representations but rather on exploiting them as basic features in order to build powerful semantic representations.

3 Datasets

We collected Flickr groups starting with an initial list of 100 million metadata pieces from which we extracted the most frequently occurring groups. Then we downloaded group metadata for the most frequent 50,000 of them and retained the 38,500 groups which include at least 300 images. Given that some images were withdrawn by users before crawling, the initial dataset contains approximately 11 million images. This selection is done in order to ensure that a reasonable amount of data is fed into the visual classifiers which are build from Flickr groups.

ImageNet [7] is a visual resource built on top of WordNet. It contains manually labeled examples for nearly 22,000 concepts. From this dataset, we selected the 17,462 concepts which have at least 100 associated images so that the resulting subset includes around 13 million images.

3.1 Dataset preprocessing

Visual preprocessing focuses on the extraction of Overfeat features [21] of Flickr groups and of ImageNet concepts. Groups and concept Overfeat features are written as $FG_o = \{I_i\}_{i=1...N}$ and respectively $IN_o = \{I_i\}_{i=1...N}$ where I_i is Overfeat feature associated to an image. The default configuration, i.e. layer 19 of the small network provided by Overfeat, is used for representing the datasets and for experiments. All images are represented by a vector of 4096 dimensions which is further normalized using L2. A similar extraction process is applied to a set of 4,000 diversified images from which we select negative examples during group/concept modeling.

Text preprocessing consists in extracting the most salient tags of each group. Groups are structured thematically but a single tag might not be sufficient to describe them. Tags are ranked by the number of unique users which annotate images of a group with them. This measure is chosen instead of tag frequency, which is sensitive to bulk uploads, in order to maximize the social relevance of tags. After an initial examination, we empirically retain the top three tags as a textual representation of groups and write this representation as $FG_t = \{T_1, T_2, T_3\}$.

4 Flickr group modeling

4.1 Group analysis

Flickr groups may be formed around specific concepts (brands of cars, animals etc.), abstract concepts (beauty, frightening imagery) or they may gather images taken with a specific brand

of camera or camera setting (black and white, light setting). We investigate how much visual variability there is among the images of a group. We measure the visual coherence of a group by the 5-fold cross-validation score reported by the model trained on that group, as described in Section 5. Figure 1 confirms that the ranking based on the cross-validation score manages to separate groups depicting concepts with a clear visual representation from those focusing on photography or generic concepts. Only the first 30,000 groups ranked by cross-validation are retained for the experiments described in Section 6. Most other groups are not conceptually oriented and have little added value as dimensions of a semantic image representation.

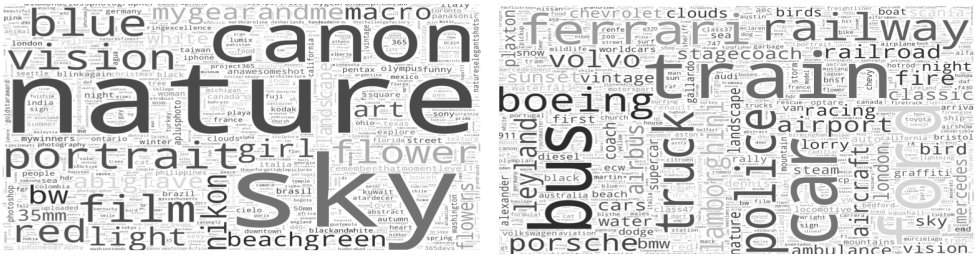

Fig. 1. Word clouds of the most frequent tags found in the last 10% groups (left word cloud) and the first 10% groups (right word cloud) in a ranking induced by the cross-validation accuracy score.

Performances of semantic descriptors obtained from ImageNet and Flickr groups are compared throughout the paper. For a better understanding of the results, we are interested in the semantic overlap between ImageNet concepts and groups and the particularities of each data source. We consider that an ImageNet concept and a group match if at least one term describing the concept has an exact match in the FG_t representation of the group. From the total of 17,462 ImageNet concept names, only 2,567 are found in groups. The concepts from ImageNet that do not appear in groups include species of animals (*African elephant, eastern grey squirrel*) or technical equipment (*computer keyboard, microphone*). When first looking at Flickr groups, we find 28,243 groups that have at least one tag matching an ImageNet concept. Among the first groups ranked by the number of contributing users that do not have an ImageNet correspondent, we notice a high frequency of geographical locations ({*paris, france, eiffel*}, {*croatia, sea, dubrovnik*}) and car brands ({*bugatti, veyron, supercar*}, {*lamborghini, gallardo, murcielago*}). The main advantage over ImageNet comes from the nature of the concepts found in groups that are formed through social consensus, as opposite to ImageNet, where the specific concepts come from the leafs of the WordNet hierarchy and may be less encountered in images shared through online platforms.

4.2 Group image reranking

A part of the images associated to Flickr groups are irrelevant and direct learning of visual models with all group images is sub-optimal. We introduce image reranking techniques in order to automatically reduce the amount of noise present in groups. Existing approaches either exploit tags or rely both on tagging and/or visual content [3]. We test two classical methods and introduce moreover one which highlights a social cue [12], i.e. the identity of the uploader. Focus is put on scalability in order to be able to process Flickr groups efficiently. All methods use the Overfeat representation of group images described in 3.1. Furthermore, we implemented:

- avg_{sim} - this baseline method computes I_{avg}, the average Overfeat representation of each group, and ranks the images of the group by considering $sim(I_i, I_{avg}) = \frac{1}{||I_i - I_{avg}||^2}$, the inverse of the L2 distance from the average representation. The intuition supporting this method is that the similarity with group average is a good indicator of image relevance.
- *kNN* - classical method which compares group images with a set of diversified negatives in order to favor images which are best linked to other images of the same group. We keep

computation cost low by choosing as many negative as there are images in the target group. The reranking score of each target image is given by the position of the 10^{th} group image in the list of similar images which includes both positives and negatives. The higher this position is, the better the image rank will be. This reranking approach is motivated by the assumption that relevant images are more similar to other images of the group than to images from other groups.

– *skNN* - is a "social" version of *kNN* in which all images which come from the same user as the target images are excluded from the list of similar images. Here we assume that an image is more likely to be relevant if it is visually similar to images uploaded by other users. *skNN* is more robust to bulk upload behaviors than the simple *kNN* algorithm.

Training images are sorted according to one of the methods described above. Then only the top of the reranked list, referred as *cut*, is retained for group modeling.

5 Semantic features building

We combine a visual resource, initial low-level features, such as FG_o or IN_o, and an array of individual concept classifiers in order to compute a semantic representation of images. Simply put, our approach can be summed-up as: "use an appropriate large-scale visual resource, represent concepts with linear models, exploit a good low-level feature, and sparsify to retain the most salient dimensions". The focus is on the exploitation of automatically built resources, using for instance Flickr groups, but the pipeline is generic and is also applied to ImageNet. The two semantic features are named $Semfeat^{FG}$ and $Semfeat^{IN}$ respectively.

5.1 Model learning and aggregation

The individual dimensions of semantic features can be modeled with different types of classifiers. Considering the size of the problem we tackle, binary classifiers are used. In comparison with a multiclass classifier, they present the advantages of (i) remaining computable for any number of classes and having lower constraints on the training dataset size (ii) being easily "extensible", i.e. a given concept can be added/removed independently from other dimensions. For scalability, we model each concept with linear models, which are very fast to compute and exhibit good performance in practice [18, 11]. Hence, each individual model is learned from a set $\{(I_i, y_i)\}_{i=1...N}$ of training images and their corresponding binary label ($y_i \in \{-1, +1\}$). Models are learned with L2-regularized logistic regression, which solves the following unconstrained optimization problem:

$$W^c = \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} log(1 + e^{-y_i \mathbf{w}^T x_i}) \quad (1)$$

Where $x_i \in \mathbb{R}^{S_f}$ is an image low-level feature reflecting the visual content of image I_i. This feature is later augmented with a last dimension fixed to 1 to take into account the model bias: $\mathbf{f}_i^T \leftarrow [x_i^T \ 1] \in \mathbb{R}^{(S_f+1)}$. In practice, (1) is solved in the primal using a trust region Newton method, relying on the liblinear implementation [9].

Individual visual models map image features into a semantic space of size *sup* defined by Flickr groups or ImageNet concepts. We denote by $\mathbf{W} = \{\mathbf{W}^1, ..., \mathbf{W}^{sup}\} \in \mathbb{R}^{(S_f+1)\times sup}$ as the matrix concatenating all individual visual models learned by (1). Using the $\mathbf{W}$, an initial image feature $\mathbf{f} \in \mathbb{R}^{(S_f+1)}$ is mapped to its semantic representation $\mathbf{x} \in \mathbb{R}^{S_s}$ through

$$\mathbf{x} = \mathbf{W}^T\mathbf{f}. \quad (2)$$

$\mathbf{x}$ (a short notation for *Semfeat*) contains semantic information as it aggregates the classification scores of $\mathbf{f}$ given all available Flickr groups or ImageNet concepts. (2) is therefore comparable to a soft assignment encoding since all binary classifiers contribute to the semantic representation.

This feature is dense since all classifier outputs are taken into account. One drawback of such representation is that the effects of a relevant classifier can be smoothed by the accumulation of weights of an array of poorer classifiers. For instance, a concept whose output is 0.95 has lower importance than a combination of 5 concepts with 0.2 outputs.

5.2 Semantic feature sparsification

Liu et al. [15] showed that soft assignment encoding is not optimal as it discards the manifold geometric structure of the mapped space. Beyer et al. [2] demonstrated that distances between points often become less meaningful in high-dimensional space. Consequently, the expressive power of high-dimensional features is limited. Moreover, empirical analysis of high-dimensional features [26, 25] shows that they often lie on a manifold which has a much smaller intrinsic dimensionality. Such a manifold structure implies that the neighborhood of a feature point is homeomorphic to the Euclidean space into a local region only. Thus, computing distance (or proximity) between features is meaningful within a local region only. Outside of this region, two local points considered similar using a distance measure might actually be far from each other.

Given this manifold assumption, a classification score indicating the proximity of a feature $\mathbf{f}$ to a concept c is reliable only when $\mathbf{W}^c\mathbf{f}$ is large. The use of all classifiers degrades the semantic representation as concepts distant from $\mathbf{f}$ do not bring information. To obtain a more reliable representation, we leverage the manifold geometry by adding a locality constraint [25] to $\mathbf{x}$,

$$\mathbf{x} = \arg\min_{\bar{\mathbf{x}}} \|\mathbf{W}^T\mathbf{f} - \bar{\mathbf{x}}\|_2^2 + \lambda\|\mathbf{D}\bar{\mathbf{x}}\|_2^2, \text{ s.t. } \mathbf{1}^T\bar{\mathbf{x}} = 1, \tag{3}$$

where $\mathbf{D} \in \mathbb{R}^{S_s \times S_s}$ is a diagonal matrix such that each $(\mathbf{D})_{cc}$ penalizes the c-th basis of $\mathbf{x}$ when the c-th concept is unlikely to be observed in $\mathbf{f}$. More specifically,

$$(\mathbf{D})_{cc} = e^{-\gamma \mathbf{W}^{cT}\mathbf{f}} \; \forall c \in [1..S_s]. \tag{4}$$

In (4), $\mathbf{W}^c$ is the c-th linear model and γ a parameter which controls locality. In (3), the first term verifies that the semantic representation $\mathbf{x}$ is close to $\mathbf{W}^T\mathbf{f}$, i.e. the projection of $\mathbf{f}$ in the semantic space. The second term penalizes the semantic representation $\mathbf{x}$ that are non-local. It ensures that only the semantic dimensions of $\mathbf{x}$ which yield high-likelihoods will be selected. λ is a trade-off parameter between both terms.

The locality constraint in (3) chooses dimensions $\mathbf{x}$ based on the likelihood of the corresponding concepts with $\mathbf{f}$. Given this observation, an efficient sparse approximation of (3) can be derived. We define the matrix $\mathbf{W}' = \{\mathbf{W}'^1, ..., \mathbf{W}'^{S_s}\} \in \mathbb{R}^{(S_f+1)\times S_s}$ which considers only the top K individual visual models ($K << S_s$) that yield the highest scores on $\mathbf{f}$: $\mathbf{W}'^c = \mathbf{W}^c$ if the c-th concept achieves one of the K highest scores, 0 otherwise. We obtain an semantic representation which approximates locality through

$$\mathbf{x} = \mathbf{W}'^T\mathbf{f}. \tag{5}$$

Compared to (3), the only parameter is K which directly controls the locality of the solution.

6 Evaluation

Our objective is to evaluate the effectiveness of our semantic features in image retrieval. The central question is whether the performances obtained using an automatically constituted resource, based for instance on Flickr groups, can be similar to those of a manually built resource, such as ImageNet. We also evaluate the effects of semantic feature coverage and sparsification. Popular wisdom considers that an image is worth a thousand words. We test how many of these words should actually be used for image retrieval. In a preliminary experiment, we tune the learning algorithms and select the best Flickr group image reranking method. Then we perform the main experiment in a CBIR scenario using the ImageCLEF Wikipedia Retrieval 2010 collection [23]. Additionally, we present a simple fusion experiment to assess the effect of combining text and content based IR.

6.1 Preliminary experiment

The purposes of this preliminary experiment are to tune individual model learning algorithm and to evaluate results obtained with the different image reranking methods. The validation dataset used here is created by matching Flickr groups, used for training, with ImageNet concepts whose images are used for testing. We first pre-select a list of groups which have the first tag of their textual representation FG_t, present in ImageNet. For instance, we match the Flickr group 1000405@N24 (top tags *memorial, war, warmemorial*) with the ImageNet concept *memorial*, defined as *a structure erected to commemorate persons or events*. To ensure diversity, each tag is used only once. We manually validate the alignment between groups and ImageNet concepts to obtain a final list of 367 pairs. Training is done using Flickr group images as positive examples and a diversified negative set extracted from Flickr. Test is performed with the images of the corresponding ImageNet concept as positives and a fixed list of over 4000 images of other concepts as negatives. Logistic regression tuning is done via a grid search on the parameter C. The best average classification scores are obtained with $C = 10$, which is used to learn all binary models included in *Semfeat*.

The reranking methods are compared by using them in an image retrieval scenario. For instance, assuming that *palm tree* is part of the 367 pairs, the purpose is to use the models trained with Flickr groups in order to classify ImageNet test images and ImageNet negatives. Using classification scores, we produce a ranking and assume that the best reranking method is the one which places the most *palm tree* among the top images. The precision at 100 (P@100), i.e. the number of positives among the first 100 results, is a used for assessment. This measure accounts for the capacity of the reranking methods to favor positive test examples over negatives and, indirectly, for the quality of the reranked training set. Also, compared to classical cross validation, all reranking cut-offs are evaluated using the same test set; thus, results are easier to compare.

Table 1. P@100 results for different reranking methods and different cut-off percentages (*cut*) for the selection of reranked images. The baseline corresponds to a no cut-off, i.e. the rightmost column.

	cut[%]			
	70	80	90	100
avg_{sim}	0.915	0.917	0.92	0.917
kNN	0.918	0.92	0.92	0.917
skNN	**0.922**	**0.921**	**0.922**	0.917

The results obtained with the three reranking methods and different cut-off points are presented in Table 1. While the P@100 differences with the baseline are small, some improvement is obtained with all reranking methods. More interestingly, the use of *skNN* provides slightly better results compared to *kNN*, indicating that the use of social cues for reranking is beneficial. Given the results presented in Table 1, we will use *skNN* at different cut-off points in retrieval experiments.

6.2 CBIR and fusion evaluation

The main objective here is to assess the usefulness of *Semfeat* in a CBIR task performed over a diversified dataset. Wikipedia Retrieval 2010 was created as part of the ImageCLEF evaluation campaign[1] and is publicly available. It includes 237,434 Wikimedia images which were extracted from a large set of Wikipedia articles and includes a wide range of content. It it thus fitted for ad-hoc image retrieval experiments, in which any query can be submitted to the process. To ensure comparability with other methods tried on this dataset, we report mean average precision (MAP) performances. The 2010 query set contains 70 diversified queries, with 118 associated image examples. Retrieval over the Wikipedia collection is challenging because the image content is highly diversified. The Wikipedia Retrieval ground truth has been built using a pooling approach

[1] http://www.imageclef.org/

and is therefore incomplete [23]. To improve comparability, we extend the original ground truth (noted $origGT$) by pooling the new runs proposed here. This extension (noted $extGT$) is realized using similar topic narratives and a majority voting with three relevance judgments per image. We had limited resources and assessed only the new images appearing in the top 20 results of a selection of runs are annotated, compared to a pooling depth of 100 used for establishing the initial ground truth[2]. If a collection image is similar to two examples, the highest similarity score is retained. Similarities between images are computed using the cosine measure. The following runs are used in our main experiment:

- *Fisher* - existing baseline which exploits a version of Fisher Vectors adapted for CBIR [5].
- *Overfeat* - baseline which exploits the default outputs of the CNN-based feature extractor presented in [21], using the small network to speed-up feature computation. These features are also exploited to build individual models included in *Semfeat*.
- $SemfFeat^{IN}$ - semantic feature based on the 17462 ImageNet concepts which were modeled here. Results are reported for sparsification $K = 10$, corresponding to the best $Semfeat^{IN}_{full}$ MAP in Figure 2.
- $Semfeat^{FG}_{cut}$ - semantic feature based on the 30,000 groups with the highest cross-validation scores from the initial 38,500 group dataset.
- *Txt* [19] - to our knowledge, text run with best performances on the collection.
- *Fusion* - fusion of image ranks from *Txt* and $Semfeat^{FG}_{80}$ with linear weighting.

Fisher Vectors and Overfeat are strong baselines, representing both pre-CNN and CNN features. Direct comparison to meta-classes and classemes was not performed since they show performances inferior to Fisher Vectors [1]. $Semfeat^{FG}_{80}$ and $Semfeat^{IN}$ are introduced here and, while based on *Overfeat*, they represent images semantically and cover significantly more concepts.

Sparsification evaluation One of our central objectives is to create features which are in the same time efficient and compact. To this end, sparsity is a desirable property of document representations since it allows one to use inverted indexes in order to search massive datasets in real time. Sparsification is applied to semantic features built on top of 30,000 Flickr Groups and of ImageNet. These datasets have respectively 30,000 and 17,462 concepts available in each case. In Figure 2, we vary the sparsification factor $K = \{5, 10, ..., 100\}$ in order to thoroughly evaluate the effect of sparsification. Very interestingly, all group based features compare favorably with *Semfeat* for $K > 10$. This finding confirms that Flick groups can be successfully used in CBIR as a substitute for manually built resources. The results presented in Figure 2 also show that the most interesting MAPs are obtained when 10 to 50 most salient concepts detected in them are used in the representation, with small peaks around $K = 30$. This finding has an important practical implication since *Semfeat* features can be efficiently represented using inverted indexes. Inverted search is much faster than a brute force search, which is needed for dense low-level features, including Overfeat and Fisher Kernels.

Confirming the results presented in Table 1, the performances obtained with cut-offs $cut = \{70, 80, 90\}$ are close to each other. The constant gap between reranked versions of $Semfeat^{FG}_{cut}$ and $Semfeat^{FG}_{full}$ indicates that removing poorly reranked images has a clear beneficial effect regardless of the sparsification factor. Results obtained with the three *cut* values are similar, indicating that the positive effect of noise reduction and the negative effect due to training set shrinking compensate each other. Beyond $K = 100$, performances drop continuously, confirming the uselessness of concepts with low scores.

Conceptual coverage evaluation Here we evaluate the influence of the semantic features support (*sup*), which is tightly linked with their capacity to cope heterogeneous datasets. We fix the *skNN* reranking percentage $cut = 80$ and sparsification at $K = 30$ and vary the *sup* between 1000 and 30,000 for $Semfeat^{FG}_{80}$ and from 1000 to 17,462 for $Semfeat^{IN}$. The concepts/groups included in each support are selected randomly and we give the percentage of the performance obtained with the full features. Performances of $Semfeat^{IN}$ for $sup = \{1000, 5000, 10000, 15000\}$ are 70.9%, 87.7%, 89.7%, 96.4%. Performances of $Semfeat^{FG}_{80}$ for $sup = \{1000, 5000, 10000, 15000, 20000, 25000\}$ are

[2] The extended ground truth will be made available to other researchers.

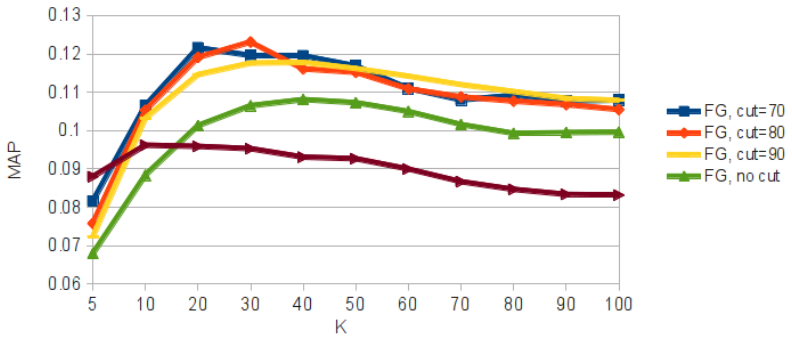

Fig. 2. Sparsification analysis in function of K, the number of most salient concepts retained in the semantic features built on top of Flickr groups and of ImageNet.

53.5%, 70.9%, 80.1%, 84.9%, 89.9%, 93.8%. *sup* effect is more important on $Semfeat^{FG}$, a behavior which is probably explained by the fact that automatically selected groups are more redundant than ImageNet concepts. Saturation is not reached in either case and this indicates that adding supplementary concepts or groups would probably further increase performances.

CBIR results analysis In Table 2, we compare the MAP scores obtained with *Fisher* and *Overfeat* baselines to those with *Semfeat* versions proposed in this paper. The Fisher vector is dense and contains over 100,000 dimensions [5]. The Overfeat vector is also dense and is obtained with the small network configuration [21]. For $Semfeat^{IN}$, we use a sparsification factor $K = 30$, which corresponds to the best MAP reported in 2. For $Semfeat^{FG}_{cut}$, results are reported for sparsification $K = 30$ and $cut = 80$, corresponding to the best MAP in Figure 2.

Table 2. Results for CBIR runs with the ImageCLEF Wikipedia Retrieval 2010. Both the original and the extended ground truths (*origGT* and *extGT*) are used. $Semfeat^{FG}$ results are reported for sparsification $K = 30$. *Fisher* performances do not change since this run was already pooled for *origGT*. For *Fusion*, contributions from image and text are weighted with 0.8 and respectively 0.2 but values are relatively stable for neighboring weights.

	Fisher	*Overfeat*	$Semfeat^{IN}$	$Semfeat^{FG}_{full}$	$Semfeat^{FG}_{80}$	*Txt*	*Fusion*
MAP *origGT*	0.0553	0.0986	0.0962	0.1065	0.1231	0.2786	0.2812
MAP *extGT*	0.0553	0.1149	0.1167	0.1267	0.1487	0.3358	0.3378

The results presented in Table 2 show that $Semfeat^{IN}$ has roughly the same performances as *Overfeat* while both $Semfeat^{FG}$ are better than this strong baseline. The best overall results are obtained with $Semfeat^{FG}_{80}$. This feature built on top of Flickr groups which exploits *skNN* reranking (24.8% and 29.4% relative improvement compared to *Overfeat* with *origGT* and *extGT*). A t-test shows that $Semfeat^{FG}_{80}$ is significantly different from all other CBIR runs with p at least 0.001. Compared to *Fisher*, the previous state-of-the-art method tested on this dataset, improvements are very consequent, 122.6% and 169% relative improvements for $Semfeat^{FG}_{80}$ for the two ground truths. As a complement to CBIR, we ran a simple fusion experiment. Even with a simple approach, results are however improved by over 20% compared to the text run and this difference is statistically significant with $p \leq 0.001$. This result confirms that text-image fusion is beneficial. The improvement could probably be further improved using more advanced fusion methods [6].

The query set includes 70 topics and it would be therefore impractical to plot individual bars in order to visualize results. Instead, we present best and worst 10 topics ranked by MAP scores in Table 3. Confirming intuition, topics with high MAPs correspond to common Flickr topics, well represented in $Semfeat^{FG}_{80}$. Topics with low MAPs often depict non-natural scenes and the bad behavior of $Semfeat^{FG}_{80}$ is explained by two factors: *Overfeat* was trained mostly with natural

Table 3. Best and worst 10 topics ranked by MAP score using $Semfeat_{80}^{FG}$ with *origGT*.

	MAP range	Textual topics
Best 10	0.52 - 028	stars and galaxies, tennis player on court, close up of bottles, polar bear, cyclist, race car, launching space shuttle, lightning in the sky, civil airplane, sailboat
Worst 10	0.003 - 0	paintings related to cubism, fractals, musician on stage, DNA helix, shiva painting or sculpture, solar panels, Oktoberfest beer tent, Rorschach black and white, videogame screenshot, Chernobyl disaster ruins.

images and many of these topics are poorly represented in Flickr groups. CNN retraining would be needed to deal with these cases but it falls outside the scope of the paper. Other examples of bad behavior include topics which are visually hard. For instance, *Oktoberfest beer tent* images often depict crowds which are difficult to distinguish, *solar panels* are visually similar to the surface of *skyscrapers* while *Chernobyl disaster ruins* can easily be mistaken for other ruins.

To give insight about $Semfeat_{80}^{FG}$ robustness, we compare its individual topic MAPs with those of two baselines. First $Semfeat_{80}^{FG}$ is better in 45 cases (average MAP gain of 0.068), *Fisher* in 22 cases (average MAP loss of -0.031) and there are 3 ties. The largest 3 gains are obtained for *tennis player on court* (0.469), *cyclist*(0.41) and *polar bear* (0.391). Inversely, the largest performance losses occur for *postage stamps* (-0.177), *brain scan* (-0.099) and *earth from space* (-0.095). Completing Table 3, these examples indicate that $Semfeat_{80}^{FG}$ is better for natural images whereas *Fisher* behaves better for other types of images.

We further illustrate the results obtained with $Semfeat_{80}^{FG}$ in Figure 3. The first two rows have high MAP scores in the original ground truth, the following two are in the middle of the topic ranking and the last two correspond to queries with poor results. Although not in focus in this paper, automatic image annotation with large vocabularies is a part of *Semfeat* pipeline and we present a list of 5 Flickr group tags which are automatically associated to query images. Interestingly, even though annotations are only partially relevant, their combination in *Semfeat* often favors the retrieval of relevant images, as this is the case for *tennis player*. The only image whose annotations are all conceptually unrelated to the image is *DNA helix*. Confirming the results in Table 3, the last two rows indicate that the conceptual support of *Semfeat* should be further extended.

7 Discussion

Advantages CBIR results show that the *Semfeat* version based on reranked Flickr groups significantly outperforms other existing methods which were tried on the Wikipedia Retrieval dataset. Moreover, text-image fusion further improves the quality of results. Beside competitive performances and contrary to widely used image features, such as bags-of-visual-words, Fisher Kernels [11] or CNN features [21], *Semfeat* directly conveys semantic meaning. Image similarities are based on the comparison of humanly understandable dimensions, a characteristic which enables result explainability and reduces the *semantic gap*.

Another advantage of *Semfeat* is its sparsity. The best performances are obtained when only a few dozens of concepts are kept for each image. In this configuration it is straightforward to efficiently represent images as inverted indexes in order to speed up retrieval. We tested inverted search with a simple in-memory C++ implementation and simulated datasets up to 100 million images with sparsity $K = 100$. Retrieval time grows linearly and is under 1 ms for 10 million images and under 10 ms for 100 millions. For comparison, we also tested forward search with Overfeat (4096 dimensions) and obtained a retrieval time in the range of 15 s for 10 million images. Even if one would use compressed versions of dense features, inverted search would still be faster.

Last but not the least, $Semfeat^{FG}$ is built with an automatically mined dataset, using simple reranking and learning. This pipeline facilitates resource extension to new concepts/groups.

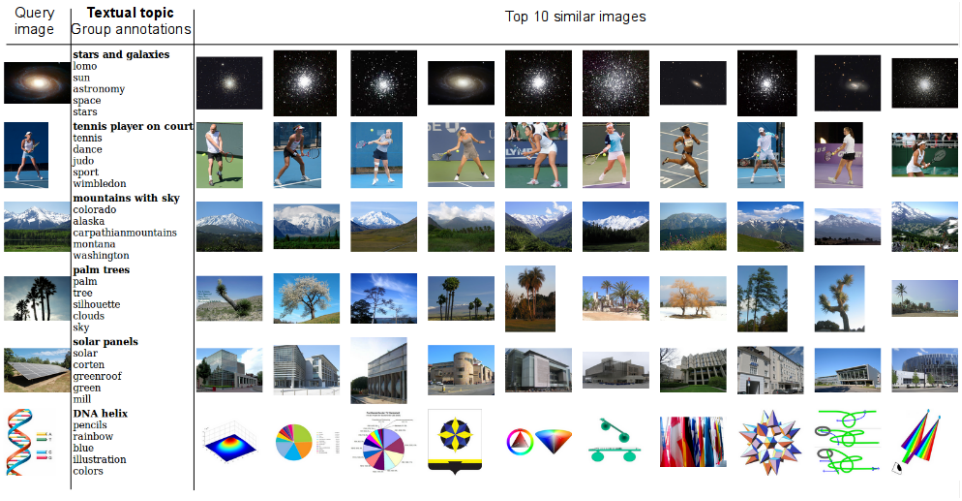

Fig. 3. Illustration of the CBIR process based on $Semfeat_{80}^{FG}$. We present the query image, the associated textual topic (bold face), 5 automatic annotations from Flickr groups and the most similar images from the Wikipedia collection. We present two highly ranked topics, two from the middle of the ranking and two from the bottom according to *origGT*. The *mountain* example illustrates well the incompleteness of the *origGT* because, while relevant, many of its neighbors were not found by official campaign runs.

Limitations We have mentioned some *Semfeat* limitations in the experimental section and extend the analysis. The learning methods used here are scalable but can be improved. With the use of more sophisticated models, the predictions associated to *Semfeat* dimensions would probably be more robust and have a positive impact on the overall results. However, when choosing the learning models to use, one should keep in mind that the prediction process needs to be fast, a constraint which is particularly relevant when the semantic features include a large number of dimensions. Consequently, as the authors of [1], we advocate for the use of linear models and will test different such models in the future. Another important limitation is the choice of positive examples which model individual groups/ concepts. We implemented a first version of *Semfeat* with a maximum of 300 images per group. While this volume is sufficient for simple visual concepts, it is probably insufficient to model complex concepts and future experiments should focus on enlarging the positive examples set. In such a setting, a larger amount of potentially noisy images could be removed while still having a sufficiently rich and diversified representation of the concept. Flickr groups mirror users' interests but they are often redundant. For instance, there are tens of different groups which focus on *classic cars* and several of them can be jointly activated in the *Semfeat* representation. These groups could probably be merged into larger meta-groups to reduce redundancy and propose more informative features.

8 Conclusions

We proposed a technique for the automatic mining of large-scale visual resources from Web data. With the support of this, we proposed a new semantic image representation for image retrieval. Returning to our initial research questions we can conclude that:

Q1 The direct use of Web corpora, as proposed in [22] or [1], yields lower performances compared to manually curated datasets. However, with an appropriate choice of the initial collection and with the introduction of efficient image reranking techniques, the results obtained with the automatically built resource can rival with those of the manual resource.

Q2 Efficient semantic representations can be built through the combined use of powerful initial features, such as ImageNet, and of an appropriate visual representation of feature components. Further investigation is needed concerning the choice of machine learning models and the number

of images, both positive and negative, used for learning individual models. It is probable that more sophisticated models combined with a larger number of training images will improve results.

Q3 A good coverage of the conceptual space is obtained with a good choice of the Web dataset. We explored the use of Flickr groups by other concept sets; but the pipeline presented here is easily applicable to larger datasets. The only potential constraints are the availability of data and the processing power needed to build individual models.

Q4 In image retrieval, compactness is achieved by sparsifying semantic features and by using inverted indexes. With this scheme, very large volumes of data can be search without precision loss, as it is the case for existing dense features [11].

Q5 The obtained results indicate that semantic features are useful for retrieval and, compared to [22, 1], we propose an efficient way to clean and exploit large-scale noisy Web corpora.

9 Acknowledgments

This work is supported by the MUCKE FP7 CHIST-ERA project, partly funded by ANR, France, and by the USEMP FP7 project, partly funded by the EC under contract number 611596.

References

1. A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *CVPR*, 2012.
2. K. Beyer and al. When is nearest neighbor meaningful? In *ICDT*. Springer, 1999.
3. E. Chatzilari. Using tagged images of low visual ambiguity to boost the learning efficiency of object detectors. In *ACM Multimedia*, 2013.
4. Hong-Ming Chen and al. Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. In *ACM Multimedia 2008*.
5. Stéphane Clinchant and al. Xrce's participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In *CLEF'10*.
6. Stéphane Clinchant and al. Semantic combination of textual and visual information in multimedia retrieval. In *ICMR*, 2011.
7. J. Deng and al. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
8. Jeff Donahue and al. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, 2013.
9. Rong-En Fan and al. Liblinear: A library for large linear classification. *JMLR*, 9, 2008.
10. Przemyslaw Grabowicz and al. Distinguishing topical and social groups based on common identity and bond theory. In *ACM WSDM 2013*.
11. Hervé Jégou and al. Aggregating local image descriptors into compact codes. *PAMI*, 2012.
12. Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *WWW 2008*, 2008.
13. A Krizhevsky and al. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
14. Li-Jia Li and al. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.
15. L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, 2011.
16. Radu Andrei Negoescu and Daniel Gatica-Perez. Analyzing flickr groups. In *ACM CIVR 2008*.
17. M. Oquab and al. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *CVPR*, 2014.
18. Florent Perronnin and al. Large-scale image retrieval with compressed fisher vectors. In *CVPR 2010*.
19. A. Popescu and G. Grefenstette. Social media driven image retrieval. In *ACM ICMR 2011*.
20. B. Russell and al. Labelme: a database and web-based tool for image annotation. *IJCV*, 77, 2007.
21. P. Sermanet and al. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013.
22. Lorenzo Torresani and al. Efficient object category recognition using classemes. In *ECCV*. 2010.
23. Theodora Tsikrika and al. Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. *IEEE MultiMedia*, 2012.
24. Gang Wang and al. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *CVPR*, 2009.
25. Jinjun Wang and al. Locality-constrained linear coding for image classification. In *CVPR*, 2010.
26. Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.