

Learning Compact Visual Attributes for Large-scale Image Classification

Yu Su and Frédéric Jurie

GREYC — CNRS UMR 6072, University of Caen Basse-Normandie, Caen, France
{yu.su, frederic.jurie}@unicaen.fr

Abstract. Attributes based image classification has received a lot of attention recently, as an interesting tool to share knowledge across different categories or to produce compact signature of images. However, when high classification performance is expected, state-of-the-art results are typically obtained by combining Fisher Vectors (FV) and Spatial Pyramid Matching (SPM), leading to image signatures with dimensionality up to 262,144 [1]. This is a hindrance to large-scale image classification tasks, for which the attribute based approaches would be more efficient. This paper proposes a new compact way to represent images, based on attributes, which allows to obtain image signatures that are typically 10^3 times smaller than the FV+SPM combination without significant loss of performance. The main idea lies in the definition of intermediate level representation built by learning both image and region level visual attributes. Experiments on three challenging image databases (PASCAL VOC 2007, CalTech256 and SUN-397) validate our method.

1 Introduction

Attribute based image classification [2–6] – in which an image is represented by a set of meaningful visual attributes – has several interesting properties such as the ability to handle large number of categories or the compactness of image representation. For example, in [5], an image is represented by a 2659-d binary vector, each of which corresponds to a visual attribute. However, the attribute based methods typically need a large amount of human efforts, *i.e.* manually defining visual attributes and labeling training images for these attributes. The only exception is [6] which learns both discriminative and nameable visual attributes without labeled training images. But this learning process still includes human supervision and therefore is semi-supervised. Another drawback of visual attributes is their classification performance, which is below or comparable to the simple Bag-of-Words histogram when using the same low-level features.

Indeed, recent literature in image classification have shown that the state-of-the-art results are typically obtained by combining Fisher Vectors (FV) and Spatial Pyramid Matching (SPM) which leads to very high dimensional image signatures. For example, as in [1], the fisher vector for an image is 32,768-d and the final image signature with a three level pyramid ($1 \times 1, 2 \times 2, 3 \times 1$) is 262,144-d. This is a hindrance to large-scale image classification since storing

high dimensional features for thousands of (or even millions of) images and learning classifiers based on them is very difficult if not impossible. Considering that, many methods were proposed to produce compact image signatures.

SPM [7] divides an images into fixed regions which are not guaranteed to be optimal. Thus, some extensions of SPM were proposed to either learn the positions and sizes of regions [8, 9] or learn a weight for each region [10]. Although these methods produce more compact image signatures than SPM by using less regions, the compression rate is only about 1/4 or 1/2 which is far from enough for large-scale classification tasks. The quantization based techniques were also proposed to compress the high dimensional image signatures (*e.g.* [11–13]). Especially, in [13] the product quantizers (PQ) are adopted to compress the FVs to 1/64 of their original size, without significant loss of performance.

In this work, we propose a novel way to automatically (*i.e.* without any additional annotations) learn both image-level and region-level attributes. The former encode the common visual structures of whole images (corresponding to the 1×1 channel of SPM), while the latter encode the common visual structures of image regions (corresponding to the $2 \times 2, 3 \times 1$ channels of SPM). More specifically, to learn the visual attributes, we first compute descriptors (FVs in our case) for training images or regions randomly sampled from training images. Then we build a small set of prototypes (clusters) from these descriptors and train one classifier per prototype. An image is then encoded by measuring the similarities between its descriptors and the prototypes using the pre-learned prototype classifiers. Since the prototypes usually encode high-level visual structures (see Fig.2), they can be also considered as visual attributes. In the follows, we use the words *attribute* and *prototype* interchangeably. The resultant image signature is called as visual attribute feature (VAF). We show by experiments that, compared with some best known methods, the VAF leads to much better trade-off between compactness and classification performance.

2 Method

Our method has two components: offline learning of image/region attributes and online prediction of them. The former learns a set of attributes from both images and image regions and, based on them, the latter produces compact image signatures. Fig. 1 illustrates how to learn and predict region attributes. The process for image attributes is similar.

2.1 Describing images and regions

Recently, Fisher Vector (FV) has shown state-of-the-art performance on image classification. FV characterizes the first and second order differences between the local features and the centers of a Gaussian mixture model (GMM) which is learned on the local features extracted from training images. Given a set of local features $\{x_i : i = 1, \dots, N\}$ extracted from an image or an image region,

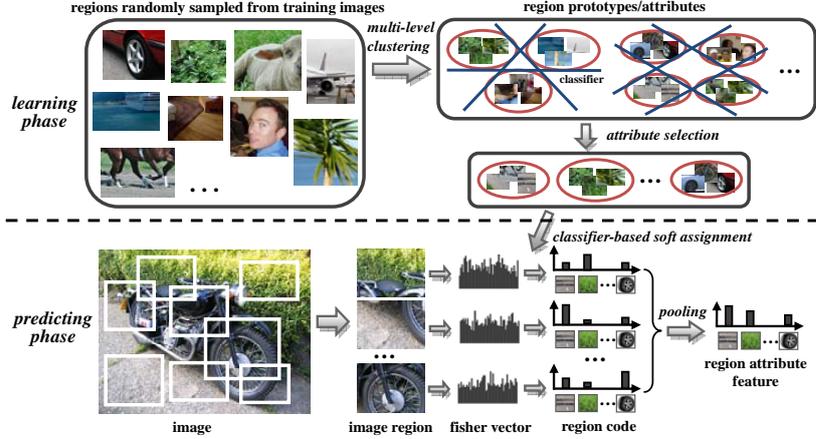


Fig. 1. Learning and prediction of region attributes. See text for details.

the FV for the k -th GMM component are computed as:

$$u_k = \frac{1}{N\sqrt{w_k}} \sum_{i=1}^N \gamma_{ik} \left(\frac{x_i - \mu_k}{\sigma_k} \right), \quad v_k = \frac{1}{T\sqrt{2w_k}} \sum_{i=1}^N \gamma_{ik} \left[\frac{(x_i - \mu_k)^2}{\sigma_k^2} - 1 \right] \quad (1)$$

where w_k, μ_k, σ_k are the parameters of GMM and γ_{ik} is the soft assignment value. Concatenating both u_k and v_k for all the K components leads to a FV of size $2KD$ where D is the dimension of local features. To speedup the computation of FV, we sparsify the γ_{ik} , *i.e.* set those for which $\gamma_{ik} \approx 0$ to 0. It is worth noting that in this case the FV is still a dense feature vector since the number of GMM components is very small.

In this work, an image is represented by one image-level descriptor and several region-level descriptors. The former is computed by aggregating all the local features from the image into a FV, while the latter are computed by randomly sampling rectangular regions from the image and aggregating the local features within each region into a FV.

2.2 Learning visual attributes and their predictors

Let $\{f_i : i = 1, \dots, M\}$ be the FVs extracted from either training images or image regions. Our objective is to obtain a set of visual attributes representing images and regions. We do this by performing spectral clustering [14] on image and region level FVs separately. Each cluster contains visually similar images or regions, and constitutes a visual attribute. In spectral clustering, the FVs are first projected into a low-dimensional manifold by using a similarity matrix of them, and then the traditional k-means is performed on the low-dimensional data to obtain the clusters. Compared with performing k-means directly on the

high dimensional FVs, spectral clustering can better capture the intrinsic visual structure of images or regions.

In our implementation, the similarity between two FVs f_i and f_j is computed using a Gaussian kernel $s(f_i, f_j) = \exp(-\|f_i - f_j\|^2/2\sigma^2)$ where the scaling parameter σ is determined by cross-validation. The visual structures of images or regions, which we aim to capture, can exist at different levels. Thus, in our method, we run the spectral clustering with different number of clusters and aggregate all the so-obtained clusters to form a vocabulary of attributes. We finally have two vocabularies of attributes: $[a_1^g, \dots, a_{C_g}^g, a_1^l, \dots, a_{C_l}^l]$, where a^g and a^l are the image and region attributes respectively.

After obtaining visual attributes, we train a classifier (linear SVM in this work) for each of them by the one-vs-rest strategy, producing a set of attribute classifiers $[\phi_1^g, \dots, \phi_{C_g}^g, \phi_1^l, \dots, \phi_{C_l}^l]$. The classifier training process is performed for the different clustering levels independently. These attribute classifiers are then used as predictors to produce the attribute features, as described in the next section.

2.3 Generating visual attribute feature

The generation of attribute feature can be considered as an encoding process. The simplest method is the hard assignment in which a vector (FV in our case) is represented by its nearest prototype. The underlying assumption of this strategy is that the vectors satisfy Gaussian mixture distribution and a vector can be represented by a single prototype. To relax this assumption, soft-assignment [15] has been proposed: a vector is assigned to multiple prototypes with assigned values proportional to its similarities to the prototypes. However, this soft-assignment model also assumes the Gaussian mixture distribution of vectors.

In practice, the assumption of Gaussian mixture distribution is not always well satisfied, especially when the dimensionality of feature space is high. In our case, the FVs are much higher dimensional than some common-used local features (*e.g.* 128-d SIFT). It explains why both the traditional hard-assignment and soft-assignment methods fail to perform well in our case. Thus, we propose in this work a classifier-based soft assignment to encode both image and region descriptors. Specifically, for a descriptor f , its assigned value to an attribute a is computed as

$$\Theta(f, a) = \frac{1}{1 + \exp(-\phi_a(f))} \quad (2)$$

where $\phi_a(f) = w_a^T f + b_a$ is the classifier (linear SVM) of attribute a and the output of ϕ is transformed to $(0, 1)$ by the sigmoid function.

As above mentioned, an image I is represented by an image-level FV and several region-level FVs. For the former, the image signature $\Psi^g(I, a^g)$ is computed by using Eq. (2) directly, *i.e.* $\Psi^g(I, a^g) = \Theta(f, a^g)$ where f is the image-level FV. For the latter, the image signature is computed by pooling all the encoded

image regions:

$$\Psi^l(I, a^l) = \frac{1}{R} \sum_{i=1}^R \Theta(f_i, a^l) \quad (3)$$

where f_i is i -th region-level FV extracted from image I and R is the number of regions. Finally, an image is represented by its visual attribute features (VAF): $A(I) = [\Psi^g(I, a_1^g), \dots, \Psi^g(I, a_{C_g}^g), \Psi^l(I, a_1^l), \dots, \Psi^g(I, a_{C_l}^l)]$.

2.4 Producing compact image signature

Since the learned visual attribute have large redundancy, a selection process is needed to get a compact subset of them. Given the original VAF $A = [r_1, \dots, r_C]$ obtained in the previous section, a sequential feature selection algorithm (similar to [16]) is used to select a compact subset of features (attributes) with low redundancy. At iteration p , the set A_{p-1}^s of the $p-1$ already selected features is extended by choosing a new feature in $A - A_{p-1}^s$ such as:

$$\hat{r}_p = \arg \min_{r \in A - A_{p-1}^s} \left(\frac{1}{p-1} \sum_{r_i \in A_{p-1}^s} MI(r, r_i) \right) \quad (4)$$

where $MI(r, r_i)$ is the mutual information between r and r_i which is estimated from the training set. From the information theory point of view, this criterion chooses for each step the feature with the lowest dependence (redundancy) to the set of already selected features. As to A_1^s , in our implementation, it includes a randomly chosen feature.

To get more compact image signature, the A^s is further compressed by using the Locality-Sensitive Hashing (LSH) [17]. Specifically, we draw B random vectors $\{h_b : b = 1, \dots, B\}$ and represent the image by the sign of $h_b^l A^s$ which is a B -bits binary vector.

3 Experiments

3.1 Databases

The proposed method is evaluated on three challenging image databases: PASCAL VOC 2007 [18], Caltech256 [19] and SUN-397 [20].

PASCAL VOC 2007 database contains 9,963 images of 20 object classes. Following the protocol in [18], the performance is measured by the mean Average Precision (mAP) of 20 binary classification tasks.

Caltech256 database contains 256 object categories with about 30K images. Following the protocol in [19], we run the experiments with different numbers of training images per category ($n_{train}=10$ and 30). One-vs-rest strategy is used for multiclass classification and the performance is reported as the average classification rate on 256 categories.

SUN-397 database contains 397 scene categories, each of which has at least 100 images collected from the Internet. The experimental setup [20] is similar to that of Caltech256 except that the training images per category is 50.

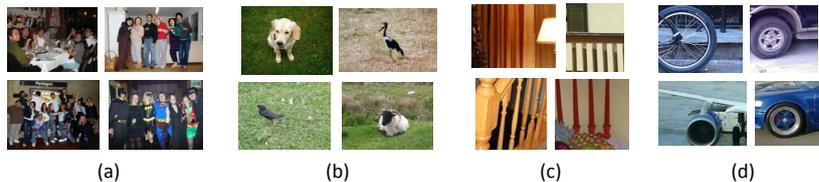


Fig. 2. Examples of image prototypes (a) (b) and region prototypes (c) (d).

3.2 Implementation details

For the local features, we adopt SIFT descriptors extracted on a dense grid of patches over 8 scales separated by a factor 1.2 and the step size is half of the patch-size. The smallest patch size is 16×16 pixels. As in [1], the SIFT descriptors are reduced from 128 to 64 by PCA and modeled by a GMM with 256 components, which results in a FV of 32,768-d ($64 \times 256 \times 2$).

When generating clusters by spectral clustering, 10 different clusterings are done for both image and region level FVs, with the number of clusters varying from 50 to 500 (with an increment of 50), which finally produces 5500 attributes. The train/validation set of PASCAL VOC 2007 is used learn the attribute classifiers and select a compact set of them.

For image classification, the classifier is also learned by linear SVM. The regularization parameter of SVM is also determined on the PASCAL train/validation set. It is worth pointing out that the randomness of the classification performance comes from the randomly sampled image regions, random initialization of attribute selection as well as the randomly selected training images (for Cal-Tech256 and SUN-397 databases). However, in the following experiments, only the averaged performances are reported since the variances in all the experimental settings are no more than 1%.

3.3 Evaluation of attribute learning and prediction

Attribute learning. As above mentioned, the learned prototypes tend to have semantic meanings and therefore can be considered as visual attributes. Fig.2 gives some examples. Specifically, the four prototypes from (a) to (d) can be interpreted as *group of persons*, *animal in the grass*, *vertical structure* and *circular object* respectively. We also compare the spectral clustering with k-means for attribute learning. It can be seen from Fig. 3 that spectral clustering gives better performance no matter how many attributes are selected, which is consistent with our analysis in Section 2.2.

Attribute prediction. In this experiment, we compare the different encoding strategies for attribute prediction as introduced in Section 2.3, *i.e.* traditional distance-based hard/soft assignment and classifier-based hard/soft assignment. Here the classifier-based hard assignment is to assign a image or region descriptor

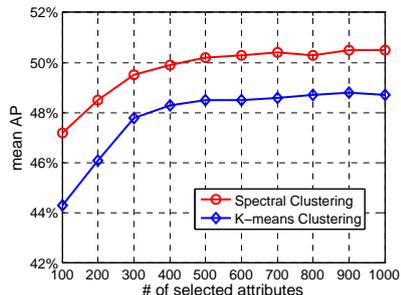


Fig. 3. Comparison of k-means and spectral clustering on the PASCAL validation set. The number of regions per image is 50.

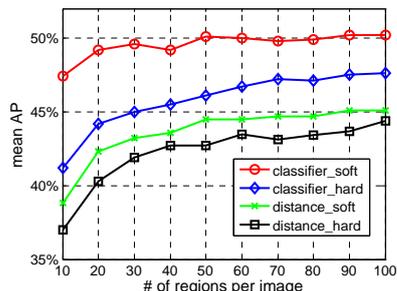


Fig. 4. Comparison of different encoding strategies on the PASCAL validation set. The number of selected attributes is 500.

to the attribute with the highest classifier output. It can be seen from Fig. 4 that the classifier-based soft assignment performs best, which validates our analysis in Section 2.3. In addition, 50 regions per image gives the best tradeoff between performance and computational cost. Thus, in the following experiments, this parameter is set to 50.

3.4 Evaluation of real-valued VAF

In this experiment, we compare the real-valued VAF with the original FV (or BoW histogram) with SPM (1×1 , 2×2 , 3×1 , making a total of 8 channels), as well as other two compact image signatures. The first one is obtained by using PCA to reduce the dimensionality of FV+SPM. The second one is the classes descriptor [5] which is the output of a large number of weakly trained category classifiers (called as "classes") on the image. The categories are selected from the LSCOM ontology and the training images are collected by the Bing image search engine. It is worth pointing out that in [5] multiple low-level features (*e.g.* GIST, HOG and SIFT) are used to learn the classes. The BoW histogram is built with 1,000 visual words (learned by clustering SIFT features), so its dimensionality with SPM is 8,000.

It can be seen from Fig.5 that the proposed VAF is very compact. Specifically, VAF with 500 dimensions performs slightly worse than the FV+SPM with 262,144 dimensions (less than 3% loss of performance for all the databases). In this case, the compression rate is about 1/500 since both VAF and FV are dense features. With this compact image signature, the time and memory costs for training image classifiers can be greatly reduced. Moreover, the VAF outperforms the PCA reduced feature which validates the effectiveness of representing images by high level visual structures. Compared with the classes descriptor which can be considered as the image level attribute feature, our method extract both image and region level attributes therefore produce more informative im-

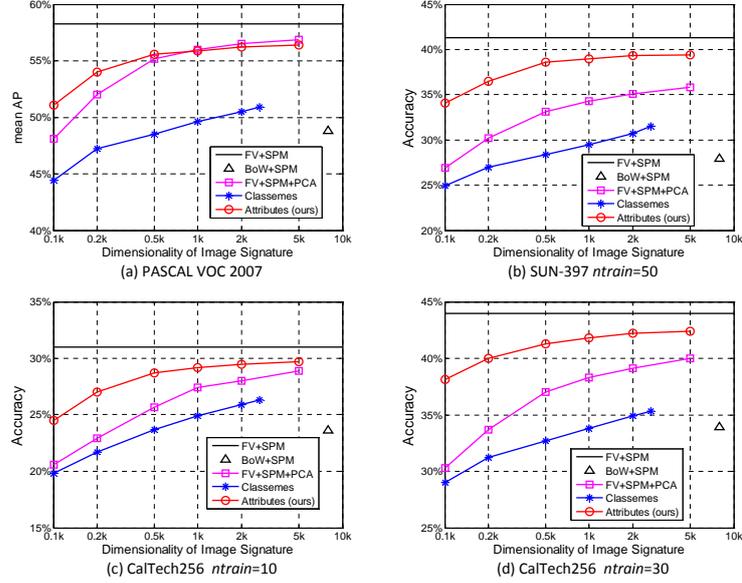


Fig. 5. Comparison of the real-valued VAF with FV+SPM, BoW+SPM, FV+SPM reduced by PCA as well as the classesmes descriptor.

age signature. Besides, the proposed VAF also outperforms the standard BoW histogram.

3.5 Comparison between binary VAF and state-of-the-art

In this experiment, we evaluate the binary VAF which is generated by applying LSH on the selected attribute feature (500-d), and compare them with two state-of-the-art binary image signatures. One is the binary classesmes descriptor [5]. The other is the PiCoDes [21] which is learned by explicitly optimizing the performance of classifying 2659 categories randomly sampled from the ImageNet dataset [22]. It can be seen from Fig.6 that the binary VAF outperforms both binary classesmes descriptor and PiCoDes except the case of small training samples ($n_{train}=10$ on Caltech256 database). Moreover, the VAF is built from single type of local features (*i.e.* SIFT) while both classesmes descriptor and PiCoDes are built from multiple types of local features. As to the runtime cost, on a machine with two 3.2GHz CPUs, it takes about 1 second to extract the binary VAF from an image of 500×500 pixels, while it takes about 2 seconds to extract binary classesmes descriptor or PiCoDes.

Especially, the performance of 4096-bits VAF is almost the same as the 500-d real-valued VAF. In this case, the compression rate relative to the original FV+SPM (with 4 bytes float point for each dimension) is $1/2048$ which is much higher than that of product quantizers ($1/64$) used in [13] to compress the FV.

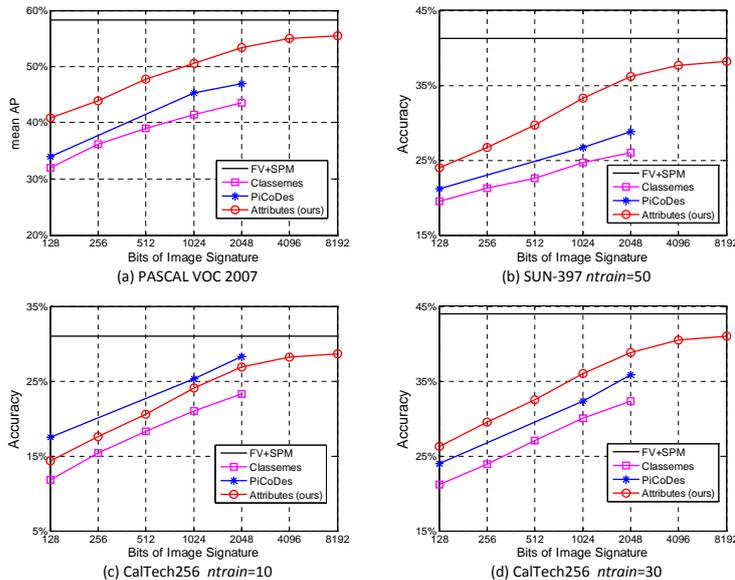


Fig. 6. Comparison of binary VAF with binary classemes descriptor and PiCoDes.

4 Conclusions

In this paper, we have introduced compact visual attribute features which encode both image and region level visual structures. Compared with the state-of-the-art fisher vector (with spatial pyramid), the proposed attribute feature is 2048 times smaller with about 3% loss of performance on all the evaluated databases. In the sense of compactness, the proposed attribute feature outperforms the best known methods, e.g. fisher vector with product quantizer [13], classemes descriptor [5] and PiCoDes [21].

It is worth noting that all the learning processes in our method (*e.g.* clustering and classifier training) are performed on PASCAL train/validation set and the learned attributes generalize well for both Caltech256 and SUN-397 databases. Thus, in practice, the visual attributes can be firstly learned in an offline manner and then applied to any classification task. Future works include applying the attribute feature to larger scale image classification (*e.g.* on ImageNet10K [22]) and image retrieval (*e.g.* on Holiday+Flickr1M [23]).

Acknowledgments.

This work was partly realized as part of the Quaero Program funded by OSEO, and partly as a part of the ITEA2 RECONSURVE project.

References

1. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010)
2. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In: NIPS. (2010)
3. Su, Y., Jurie, F.: Visual word disambiguation by semantic contexts. In: ICCV. (2011)
4. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. *International Journal on Computer Vision* **72** (2007) 133–157
5. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: ECCV. (2010)
6. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: CVPR. (2011)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
8. Cao, Y., Wang, C., Li, Z., Zhang, L., Zhang, L.: Spatial-bag-of-features. In: CVPR. (2010)
9. Sharma, G., Jurie, F.: Learning discriminative spatial representation for image classification. In: BMVC. (2011)
10. Harada, T., Ushiku, Y., Yamashita, Y., Kuniyoshi, Y.: Discriminative spatial pyramid. In: CVPR. (2011)
11. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010)
12. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: CVPR. (2010)
13. Sanchez, J., Perronnin, F.: High-dimensional signature compression for large-scale image classification. In: CVPR. (2011)
14. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS. (2001)
15. van Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J.M.: Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** (2010) 1271–1283
16. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2005) 1226–1238
17. Charikar, M.: Similarity estimation techniques from rounding algorithms. In: ACM symposium on Theory of computing. (2002) 380–388
18. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: (The PASCAL Visual Object Classes Challenge 2007 results)
19. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)
20. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010)
21. Bergamo, A., Torresani, L., Fitzgibbon, A.: Picodes: Learning a compact code for novel-category recognition. In: NIPS. (2011)
22. Deng, J., Berg, A., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: ECCV. (2010)
23. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: ECCV. (2008)