Learning to rank images using semantic and aesthetic labels

Naila Murray[1]
nmurray@cvc.uab.es
Luca Marchesotti[2]
luca.marchesotti@xrce.xerox.com
Florent Perronnin[2]
florent.perronnin@xrce.xerox.com

[1] Computer Vision Center
Universitat Autònoma de Barcelona
Spain

[2] Xerox Research Centre Europe
Meylan, France

Abstract

Most works on image retrieval from text queries have addressed the problem of retrieving semantically relevant images. However, the ability to assess the aesthetic quality of an image is an increasingly important differentiating factor for search engines. In this work, given a semantic query, we are interested in retrieving images which are semantically relevant and score highly in terms of aesthetics/visual quality. We use large-margin classifiers and rankers to learn statistical models capable of ordering images based on the aesthetic and semantic information. In particular, we compare two families of approaches: while the first one attempts to learn a single ranker which takes into account both semantic and aesthetic information, the second one learns separate semantic and aesthetic models. We carry out a quantitative and qualitative evaluation on a recently-published large-scale dataset and we show that the second family of techniques significantly outperforms the first one.

1 Introduction

Semantic retrieval is currently perceived by users as a commoditized feature of multimedia search engines. This is confirmed by a recent user evaluation [9] performed to determine the key differentiating factors of an image search engine. The top five factors were reported to be: "High-quality" (13%), "Colorful" (10%), "Semantic Relevance" (8%), "Topically clear" (7%) and "Appealing" (5%). Semantic relevance is only ranked as the third factor, whereas features related to the quality and aesthetics rank first and second.

In the past few years, the computer vision community has demonstrated a growing interest in the data-driven analysis of image aesthetics. Particular emphasis was given to the extraction of features which would suitably describe the aesthetic properties of an image. Several works in this vein proposed features which would mimic good photographic practices such as the rule of thirds [4, 7, 13, 15, 16, 21]. In a recent work, it was shown that generic image descriptors, *i.e.* descriptors which were not specifically designed for aesthetic image analysis, could yield state-of-the-art results [17]. Extracted features are used to train statistical models to discriminate between "high quality" and "low quality" images

http://dx.doi.org/10.5244/C.26.110

[5, 7, 12, 13, 15, 16, 17, 21], to predict the aesthetic score of an image [4, 22], or to rank images by their aesthetic quality [21]. These encouraging results have lead to the development of several prototypes for assessing and improving image aesthetics [12]. One such system, ACQUINE [3], predicts for a given image a corresponding aesthetic score . Another system, OSCAR [23], may be deployed to a mobile device such as a smart-phone and offers on-line feedback to help the user improve the composition or colorfulness of an image.

In this work, given a semantic (textual) query, we are interested in retrieving images which are *both relevant and score highly in terms of aesthetics*. To our knowledge only two papers have started exploring this problem. In [21], textual and visual features are used to predict the aesthetic scores of images retrieved using textual queries. The retrieved images are then re-ranked by the sum of their aesthetic score and their query relevance score. In our work, we do not assume the availability of textual features to score the semantic relevance of a new image. Geng *et. al* [9] propose to train a ranking-SVM using visual, textual and contextual features. Like [21], textual features are used for determining semantic relevance. For a given query, [9] enforces relevant high-quality images to rank higher than relevant low-quality images which should themselves rank higher than irrelevant images (whatever their quality). See their section 7.2 for more details. We believe that a significant limitation of this approach is that the model mixes both sources of variability (semantic and aesthetic), thus making the job of the ranker significantly more difficult. In this work, we advocate models which treat these two sources of variability separately.

In this paper, we make three main contributions. First, through a statistical analysis, we show that aesthetic rankings cannot be directly inferred from crowd-sourced aesthetic scores and we provide a strategy to derive meaningful relevance levels from these scores. Second, we show that the ranking approach of [9] can be significantly improved by an appropriate re-weighting of the training samples inspired by the re-weighting of positive and negative examples when learning binary classifiers. Finally, and more importantly, we propose two simple models which, as opposed to [9], separate the semantic and aesthetic components. In the case of the first model, the aesthetic part is independent of the semantic part while in the second case, the aesthetic part depends on the semantic part. Our experimental results demonstrate that it is preferable to train separate components for semantics and aesthetics rather than include them into a single model. The remainder of this paper is organised as follows: in section 2 we describe the database we used. In section 3 we describe and evaluate the three approaches for learning to rank images using aesthetic and semantic labels. Conclusions and future work are outlined in section 4.

2 Dataset and Experimental Protocol

2.1 The AVA dataset

A key aspect of our work is the study of methods that reuse existing corpora with aesthetic and semantic annotations. Recently, a large scale database (AVA, Aesthetic Visual Analysis [18]) containing such annotations was published. AVA was derived from the website WWW.DPCHALLENGE.COM, where photography hobbyists and professionals submit images in response to photographic challenges, defined by textual descriptions. The submitted images are then scored in terms of their aesthetics, taking into account the challenge description. AVA provides almost 1,500 such challenges. For each image, a score distribution is available to characterize its aesthetic quality. On average, each image is described with 200

votes between 1 and 10.

Semantic labels. Semantic information is available in the form of textual tags (at most 2 per image) and from the textual description of each challenge. Tags are assigned by photographers while challenges are created by the website moderators. To have an idea of the kind of semantic information that can be deduced from AVA, we manually inspected the textual description and title of each challenge. We discovered that most of the challenges are dedicated to themes (e.g. vintage, spooky, Halloween), concepts (e.g. poverty, trance), or photographic techniques (e.g. rule of thirds, macro, high dynamic range). Semantic categories are present in a smaller amount. In addition, the variety of semantic subjects is limited, as well as the number of images per challenge. Because of these limitations, we used the semantic information present in the form of the 33 textual tags listed in the horizontal axis of Figure 1. On average, 8,000 images are available for each tag.

Aesthetic labels. Each image in AVA is associated with a distribution of scores in a predefined range (1=lowest score, 10=highest score) that we normalized between -1 and 1. We averaged the distributions of scores per semantic tag and obtained the boxplots in Figure 1. As can be seen, such averaged distributions are rather stable across the various semantic tags. However, we are confronted with a fundamental problem: how to represent the aesthetic information compactly and efficiently. The objective is to find a representation suitable for learning different types of statistical models (such as discriminative classifiers or rankers).

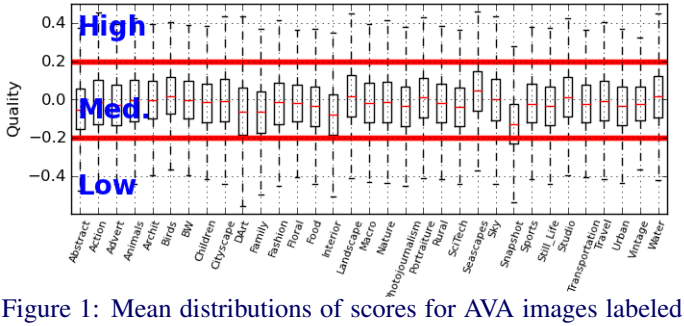

Figure 1: Mean distributions of scores for AVA images labeled with the 33 textual tags. Two thresholds define the aesthetic labels used to train the aesthetic models.

A reasonable representation would be to derive binary labels ($''High-quality''$ and $''Low-quality''$) from the mean scores of images. However, deciding on a threshold for binarization is non-trivial. Following a common approach in computer vision we could interpret classification as a retrieval problem. This decision would ultimately lead to the definition of image ranks as ground truth. Since we have scores distributions associated with each image, a natural approach to derive such ranks would be to sort the images using their mean score. Such a ranking would assume that the difference between the mean scores of a pair of images, termed $\Delta_{i,j}$, is statistically signific

To test the validity of this assumption, we sorted all images in AVA by their mean scores and applied two-sample t-tests to adjacent images. For each pair, the null hypothesis was that the means of the score distributions of the images were equal. We assumed the distributions to be normally distributed, which is a fair assumption as described in [18]. We also assumed that an image's votes are independent of each other, which is also fair as a user is not shown the votes already submitted for an im-

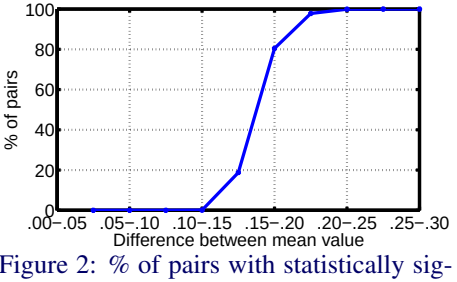

Figure 2: % of pairs with statistically significant differences in mean scores as a function of difference in mean score.

age prior to voting. Lastly, the variances of the distributions were assumed to be unequal. We found that it is not a good option to use ranks derived from sorting mean votes. In fact, none of the $\Delta_{i,j}$ values for adjacent pairs in such a rank are statistically significant at the 10% significance level. As can be seen from Figure 2, $\Delta_{i,j}$ should be set around .20 to generate statistically significant pairs. Therefore, we opted for an annotation strategy involving three labels: $''High-quality''$, $''Medium-quality''$, $''Low-quality''$. A simple thresholding operation is performed on the mean of the original votes to define for each image one of the three labels. A very small amount of image pairs picked around these thresholds are not statistically significant, but this does not impact the performance of our model. We believe that using three labels to represent aesthetic quality is a good compromise between using the mean scores and using binary labels.

2.2 Experimental protocol

We experiment with the images in AVA that are associated with the textual tags listed in Figure 1. These images were split into 5 folds, with images being evenly distributed over the folds according to their semantic tags (training, validation and test lists will be made available on-line for those interested in reproducing our results). Three folds were used for training, one fold was used for validation, and one fold was used for testing. The models were trained 5 times, with folds being switched in a round-robin fashion so that every fold was used as the validation and the test fold exactly once. The results we present are the average over the five folds.

Features. Each image is described using the Fisher Vector (FV) proposed in [19, 20]. Our main motivation for using the FV is that it was shown to yield state-of-the-art results in semantic classification [2] and aesthetic classification [17]. Note however that the models we will benchmark are independent of the image descriptors. The details of the feature extraction are as follows. We extract low-level SIFT descriptors [14] from 32x32 patches on dense grids every 4 pixels at 5 scales. The 128-D SIFT descriptors are reduced with PCA to 64-D. The Gaussian Mixture Model (GMM) is learned using a standard EM algorithm. We experimented with various vocabulary sizes (different numbers of Gaussians, typically between 16 and 256).

Model learning. To learn the semantic and aesthetic models, we employed Stochastic Gradient Descent (SGD) [1] because of its scalability.

Measures of performance. We report the normalized Discounted Cumulative Gain (nDCG), Precision and mean Average Precision (mAP). We focus on nDCG and Precision at 10, 20 and 50 as, in a real world application, it is more important to have accurate results among the top ranked images (typically the ones fitting in the first two or three pages of a search engine result). We also plot mAP calculated on the whole image ranking. We report nDCG@K averaged over all semantic tags. nDCG@K was computed as:

$$nDCG@K = \frac{DCG@K}{IDCG@K};\ DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i}-1}{log_2(1+i)} \tag{1}$$

where rel_i is the relevance level of the image at rank position i and IDCG@K is the DCG@K for a perfect ranking. mAP was computed as the mean, over the semantic tags, of the precision averaged over the set of evenly spaced recall levels $\{0.0, 0.1, 0.2, \ldots, 1.0\}$. To compute mAP, images with a relevance level of 3 (semantically relevant images with high aesthetic quality) were considered relevant.

3 Models for Combined Semantic and Aesthetic Retrieval

We assume that we have a training set of N images $\mathcal{I} = \{(x_i, y_i, z_i), i = 1 \ldots N\}$ where $x_i \in \mathcal{X}$ is an image descriptor, $y_i \in \mathcal{Y}$ is a semantic label and $z_i \in \mathcal{Z}$ is an aesthetic label. In what follows, we assume that $\mathcal{X} = \mathcal{R}^D$ is a D-dimensional descriptor space, $\mathcal{Y} = \{0,1\}^C$ is the space of C semantic labels (where $y_{i,c} = 1$ indicates the presence of semantic class c in image i), and $\mathcal{Z} = \{1, \ldots, K\}$ is the set of K aesthetic labels. In our case we have $K = 3$, where 3="$High-quality$", 2="$Medium-quality$" and 1="$Low-quality$". A major difference between spaces $\mathcal{Y}$ and $\mathcal{Z}$ is that there is a natural order on $\mathcal{Z}$. Given a semantic query specified by a class c (e.g. $c = \{$"Cat"$\}$), a traditional retrieval system would compute and rank the set of image descriptors x according to their relevance $p(y_c = 1|x)$. The problem we are investigating here is the design of a retrieval mechanism returning *high-quality* images which are also *semantically* relevant. We would also like semantically-relevant but medium-quality images to be ranked before low-quality images, as this ordering will be beneficial for classes with few high-quality images. Hence, we want to estimate $p(y_c = 1, z > \theta|x)$, where θ is some threshold on the aesthetic labels. Rather than set θ, we will rank images using ranking functions trained with aesthetic labels.

We first review the approach of [9] which consists of training a single ranker that learns simultaneously the semantics and aesthetics. We outline its limitations and then propose two models which learn separate semantic and aesthetic models.

3.1 The joint ranking model (JRM)

Original model. This approach was first proposed in [9]. Because we do not assume the availability of textual features, the approach of [9] translates to training one ranker per class in our case. Each semantic class is treated independently in which case the label set can be simplified to $\mathcal{Y} = \{0,1\}$, i.e. semantically irrelevant or relevant. A new set of labels denoted u_i is then defined as follows: $u_i = y_i z_i$. We have $u_i \in \mathcal{U} = \{0,1,\ldots,K\}$. Hence $u = 0$ means that the image is irrelevant, $u = 1$ means that the image is relevant and that its quality is the poorest possible and $u = K$ means that the image is relevant and has the highest possible quality. [9] proposes to learn a linear classifier which ranks images according to this new label u. For this purpose they train a ranking SVM as proposed for instance in [10]. Let us denote by (x^+, u^+) and (x^-, u^-) a pair of images together with their semantic and aesthetic labels in $\mathcal{U}$ such that $u^+ > u^-$. **JRM** learns w such that $w^\top x^+ > w^\top x^-$. This can be done by minimizing the following regularized loss function:

$$\sum_{(x^+,u^+),(x^-,u^-):u^+>u^-} \max\{0, \Delta(u^+,u^-) - w^\top(x^+ - x^-)\} + \frac{\lambda}{2}||w||^2 \qquad (2)$$

where $\Delta(u^+,u^-)$ encodes the loss of an incorrect ranking, for instance $\Delta(u^+,u^-) = u^+ - u^-$. One ranker w_c is learned for each class $c = 1,\ldots,C$.

Data rebalancing. **JRM** has an ambitious task: simultaneously learn aesthetics *and* semantics. In this case, the ranker has to deal with 4 relevance levels (the three aesthetic labels, and the semantic irrelevance level). As can be seen in Figure 3, labels are very imbalanced. In particular, for the "Nature" category, the probability of one of the images in a randomly-chosen pair having relevance level 0 is more than 98% (for the other classes we observed similar trends). Therefore, virtually all pairs used to train the **JRM** model encode semantic differences, rather than aesthetic information. Correcting for data imbalances has

been explored extensively for multi class categorization but little, if anything, has been done for data imbalances in ranking problems with multiple relevance levels.

	nDCG(k)			
METHOD	k=10	k=20	k=5	mAP
Semantic class. only	0.230	0.227	0.224	5.810
JRM	0.234	0.228	0.217	5.602
JRM-rebalanced	**0.253**	**0.244**	**0.227**	**6.980**
	Precision(k)			
	10	20	50	
Classification	8.538	8.284	8.270	
JRM	8.760	8.254	7.762	
JRM-rebalanced	**14.272**	**13.104**	**11.574**	

Table 1: Results with and without data rebalancing.

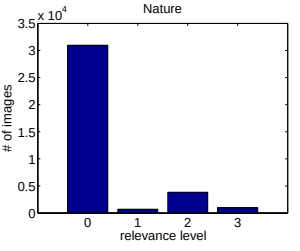

Figure 3: Distribution of relevance levels for the "Nature" category.

We implemented the following rebalancing strategy: first, we randomly draw a pair of images (i, j) subject to $u_i \neq u_j$. Then we simply multiply the probability $p_i(u)$ of drawing an image i with relevance level u_i by the probability of drawing an image j with relevance u_j. The inverse of this value is the weight:

$$\mathcal{W}_{i,j} = [p_i(u = u_i) \cdot p_j(u = u_j)]^{-1} = \left(\frac{N_{u_i}}{N_T} \cdot \frac{N_{u_j}}{N_T - N_{u_i}} \right)^{-1} \tag{3}$$

where N_T is the total amount of *training* images and N_{u_i},N_{u_j} the number of images with relevance level u_i and u_j. At iteration t of the SGD optimization, the $\mathcal{W}_{i,j}$ weight for the sample pair is applied to the update term and suppresses the amount by which the model is updated, for frequently-occuring pairs. With this weighting, highly probable relevance pairs, such as $(0,2)$, are strongly penalized.

Results. In Table 1, we shows precisions at differing ks with and without rebalancing for **JRM**. It is not completely surprising that **JRM** without rebalancing performs similarly to a semantic classifier. In fact, pairs showing the ranker differences between high and low quality images are very rare. Most pairs train the ranker to discriminate between the various semantic classes. With rebalancing we greatly improve the performance since aesthetically relevant pairs are given more importance. These results will serve as a baseline for the two models we introduce in the next subsection.

3.2 Separating semantics and aesthetics

We believe that a major weakness of the **JRM** is that it confounds both sources of variability: semantics and aesthetics. This makes the task of the linear SVM ranker more difficult. Instead, we advocate models which treat semantic and aesthetic separately.

Independent Ranking Model (IRM). The simplest strategy one can think of to model aesthetic and semantic information is the **IRM** of Figure 4. It consists of training a set of semantic classifiers (one per class) and a single class-independent aesthetic ranker capable of learning differences in quality between pairs of images.

The underlying assumption is to consider these two sets of labels as *independent*:

$$p(y,z|x) = p(y|x)p(z|x). \tag{4}$$

For the semantic part, we learn a multi-class classifier. We use the popular strategy which consists of learning a set of one-vs-rest binary classifiers independently. We learn one linear classifier with parameters α_c per class, using the set $\{(x_i, y_i), i = 1 \dots N\}$. We use a logistic loss:

$$-\log p(y_c = 1|x) = \log\left(1 + \exp(-\alpha_c^\top x)\right). \quad (5)$$

The semantic parameters α_c are learned by minimizing the (regularized) negative log-likelihood of the data on the model, which leads to the traditional logistic regression formulation:

$$-\sum_{i=1}^{N} \log p(y_{i,c}|x) + \frac{||\alpha_c||^2}{2}. \quad (6)$$

As a rule of thumb, the logistic loss gives results which are similar to the hinge loss of the SVM but the former option has the advantage that it provides directly a probability estimate.

For the aesthetic part, we learn a class-independent aesthetic ranker on the set $\{(x_i, z_i), i = 1 \dots N\}$. Let us denote by (x^+, z^+) and (x^-, z^-) a pair of images with their aesthetic labels in $\mathcal{Z}$ such that $z^+ > z^-$. We learn the aesthetic parameters β by minimizing the following regularized loss:

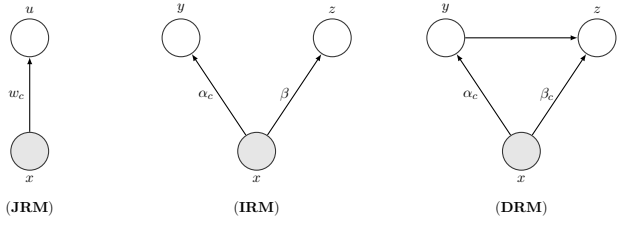

Figure 4: The three learning models we evaluate. **JRM** models semantics and aesthetics **jointly**, whereas **IRM** and **DRM** learn two separate models with different dependence assumptions.

$$\sum_{(x^+,z^+),(x^-,z^-):z^+>z^-} \log[1 + \exp(-\beta^\top(x^+ - x^-))] + \frac{\lambda}{2}||\beta||^2. \quad (7)$$

We then use a sigmoid fit to transform the score into a probability estimate $p(z > \theta|x)$.

Dependent Ranking Model (DRM). In this model, following the lessons of [7, 15] (see also introduction), we introduce an explicit dependence of the aesthetic labels on the semantic labels:

$$p(y, z|x) = p(y|x)p(z|y, x) \quad (8)$$

We train one-vs-rest binary semantic classifiers independently for each class, as was the case for the **IRM** model. However, as opposed to the **IRM**, to model the dependence of aesthetics on semantics, we train one aesthetic ranker per class independently. The loss we optimize is the same of the **IRM** (see equation 7). The only difference is that for class c we learn a ranker with parameters β_c using only the images of this class. As was the case for the **IRM**, we use a sigmoid fit to transform the ranker output score into a probability estimate: $p(z > \theta|y_c = 1, x)$.

Results. Table 2 shows a comparison between the three methods we propose. They measure the performance in terms of nDCG, mAP and Precision at K. The best performance is achieved by **DRM**. **IRM** performs slightly better than **JRM**. The advantage of **DRM** is consistent over the three measures. Worth noticing is that on this database, a baseline implemented using a discriminative semantic classifier, already performs rather well in retrieving relevant high-quality images at the top of the rank. This may be due to the fact that good quality images are highly discriminative for their semantic category.

	Precision(k)				nDCG(k)		
METHOD	k=10	k=20	k=50	mAP	k=10	k=20	k=50
Sem. Class.	8.538	8.284	8.270	5.810	0.230	0.227	0.224
JRM	8.760	8.254	7.762	5.602	0.234	0.228	0.217
JRM-balanced	14.272	13.104	11.574	6.980	0.253	0.244	0.227
IRM	18.128	17.000	15.450	8.806	0.255	0.247	0.236
DRM	**20.992**	**19.912**	**17.444**	**9.726**	**0.295**	**0.285**	**0.265**

Table 2: Comparison between the three learning strategies

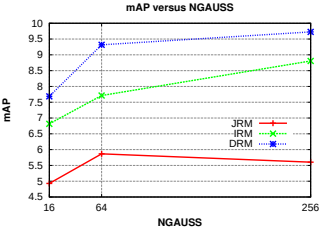

Figure 5: Performance with different visual vocabulary sizes.

However, as the mAP results show, the difference in performance is more marked if the whole rank of images is taken into account for each semantic tag. We also evaluate the impact of the model complexity by varying the visual vocabulary size (number of Gaussians). As can be seen in Figure 5, a good trade-off between computational complexity (at training time) and performance is achieved by selecting $N = 64$ Gaussians. In fact performances reach a plateau after $N = 64$.

In Figure 6 we present a breakdown of the results (nDCG@20) for each semantic tag in order to understand where content-dependence is most beneficial. From this graph we can draw some conclusions. First, **DRM** provides the best results for 15 semantic tags. For most of the other tags it is outperformed only by a small margin. Second, content dependence seems to help more for the semantic tags that are easier for the semantic classifier to learn. Data-rebalancing experiments were also performed for **IRM** and **DRM** but no significant

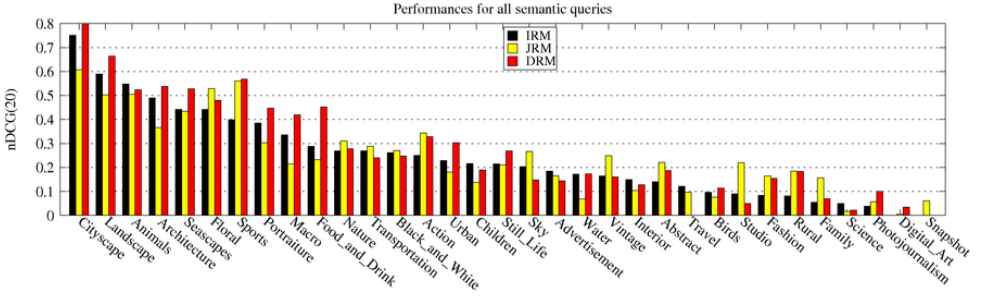

Figure 6: Performances measured with nDCG@20 for all semantic tags for the three models.

difference was found. This is expected because for **IRM** and **DRM**, separate aesthetic ranking models are trained using only relevance levels 1,2 and 3 which are much less unbalanced.

3.2.1 Qualitative analysis

To have a better understanding of the quantitative results outlined above, we also conducted a qualitative analysis. We inspected the ranking results for several semantic queries based on the performances outlined in Figure 6. In particular, we selected ranks with high, medium

and low performance. For each selected rank we plotted the top K images ranked using a semantic 1-vs-all classifier and **DRM**. Part of these qualitative results are reported in the supplementary material. The ground truth relevance levels are represented for each image by a colored image border (green="semantically relevant *and* high quality", yellow="semantically relevant *and* medium quality", red="semantically relevant *and* low quality", black ="semantically non relevant"). The first conclusion that we can draw is that, as expected, using **DRM** we improve the retrieval results for those semantic tags that are easy to learn. Next, it can be noticed that no low quality images are retrieved by **DRM**. This is a positive result since we certainly do not want to return low quality images in the top rank. Another observation is that most of the images with black borders ("semantically non relevant images") have a visual content which is indeed representing the semantic tag for which the image was retrieved (aside from some examples in the "Birds" category). This means that the labels in the AVA database contain many false negatives, and that semantic classification is robust at the top of each rank.

4 Conclusions and Future Work

In this work, we investigate three strategies to rank images by taking into account semantic relevance and aesthetic quality. In particular, we improve state-of-the-art approaches that attempt to learn aesthetic and semantic information jointly. We perform a quantitative and qualitative analysis on a large scale-dataset containing aesthetic and semantic labels. We show that content-dependent rankers combined with semantic classifiers provide the best results, and that data rebalancing is important for improving the ranking performance.

In the future, we would like to investigate the use of other large-scale databases, such as [8, 9, 10], to further improve the performance of our classifiers. We also intend to explore semi-supervised learning techniques that leverage both aesthetic and semantic annotations.

5 Acknowledgements

This work was supported in part by grant TIN 2010-21771-C02-1 of the Spanish Ministry of Science and Innovation.

References

[1] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2007.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[3] R. Datta and J.Z. Wang. Acquine: aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *MIR*, 2010.

[4] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[5] R. Datta, J. Li, and J.Z. Wang. Learning the consensus on visual quality for next-generation image management. In *ACM-MM*, 2007.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[7] S. Dhar, V. Ordonez, and T.L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*. IEEE, 2011.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[9] B. Geng, L. Yang, C. Xu, X.S. Hua, and S. Li. The role of attractiveness in web image search. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011.

[10] T. Deselaers H. Muller, P. Clough and B. Caput. Experimental evaluation in visual information retrieval. the information retrieval series. *Springer*, 2010.

[11] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[12] D. Joshi, R. Datta, E. Fedorovskaya, Q.T. Luong, J.Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*.

[13] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[14] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 2004.

[15] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, 2011.

[16] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008.

[17] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011.

[18] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

[19] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[20] F. Perronnin, J. Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[21] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. 2012.

[22] O. Wu, W. Hu, and J. Gao. Learning to predict the perceived visual quality of photos. In *ICCV*, 2011.

[23] L. Yao, P. Suryanarayan, M. Qiao, J.Z. Wang, and J. Li. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision*, 2012.