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

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.

To develop our approaches we require images with semantic and aesthetic annotations. Recently, a large scale database (AVA, Aesthetic Visual Analysis [3]) containing such annotations was published. AVA contains 33 *semantic labels* in the form of textual tags. It also contains *aesthetic labels* in the form of a distribution of scores in a pre-defined range. We would like to represent the aesthetic information in a manner suitable for learning ranking models. Since we have scores distributions associated with each image, a natural approach would be to represent the information by a ranking, where the ranks would be obtained by sorting the images' mean scores. Such a ranking would assume that the difference between the mean scores of a pair of images, termed $\Delta_{i,j}$, is statistically significant. We tested the validity of this assumption and found that it is not a good one. Instead, we opted for an annotation strategy involving three labels: $''high-quality''$, $''medium-quality''$, $''low-quality''$.

We experiment with the images in AVA tagged with one of the 33 semantic tags. Each image is described using the Fisher Vector (FV) proposed in [4, 5]. To learn the semantic and aesthetic models, we employed Stochastic Gradient Descent (SGD) [1] because of its scalability.

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 K aesthetic labels. In our case $K = 3$, where $3 = ''high-quality''$, $2 = ''medium-quality''$ and $1 = ''low-quality''$.

The joint ranking model (JRM): 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. 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 $w^\top x^+ > w^\top x^-$. A ranking SVM as proposed for instance in [2] may be trained by minimizing the following regularized loss function:

$$\sum_{(x^+,u^+),(x^-,u^-):u^+>u^-} \max\{0, \Delta(u^+,u^-) - w^\top(x^+ - x^-)\} + \frac{\lambda}{2}||w||^2 \quad (1)$$

where $\Delta(u^+,u^-)$ encodes the loss of an incorrect ranking, for instance $\Delta(u^+,u^-) = u^+ - u^-$.

For the JRM, the ranker has to deal with 4 relevance levels (the three aesthetic labels, and the semantic irrelevance level). However, these labels are very imbalanced. As a result, virtually all pairs used to train the JRM model encode semantic differences, rather than aesthetic information. To rebalance the labels 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: $\mathcal{W}_{i,j} = p_i(u = u_i) \cdot p_j(u = u_j)$.

At iteration t of the SGD optimization, the $\mathcal{W}_{i,j}$ weight for the sample pair modulates the degree of change of the model during the update step. With this weighting, highly probable relevance pairs are strongly penalized. 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. For this reason, we advocate models which treat semantics and aesthetics separately.

Independent Ranking Model (IRM): In this simple model, a set of semantic classifiers (one per class) and a single class-independent aesthetic ranker are trained. For the semantic part, we use the popular strategy of learning a set of one-vs-rest binary classifiers independently. We learn one linear classifier with parameters α_c for each class $c = 1, \ldots, C$ using the set $\{(x_i, y_i), i = 1 \ldots N\}$. We use a logistic loss: $-\log p(y_c = 1|x) = \log\left(1 + \exp(-\alpha_c^\top x)\right)$. 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:

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

For the aesthetic part, we learn a class-independent aesthetic ranker on the set $\{(x_i, z_i), i = 1 \ldots N\}$. 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:

$$\sum_{(x^+,z^+),(x^-,z^-):z^+>z^-} \log[1 + \exp(-\beta^\top(x^+ - x^-))] + \frac{\lambda}{2}||\beta||^2. \quad (3)$$

Dependent Ranking Model (DRM): In this model, we introduce an explicit dependence of the aesthetic labels on the semantic labels: $p(y,z|x) = p(y|x)p(z|y,x)$. This model is quite similar to the IRM model, but with one major difference: for class c we learn a ranker with parameters β_c using only the images of this class.

Results As table 1 shows, the best performance for three different measures is achieved by DRM. IRM performs slightly better than JRM.

	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 1: Comparison between the three learning strategies

Therefore we conclude that it is advantageous to learn semantics and aesthetics separately. We also conclude that data rebalancing is an important step to improve the ranking performance.

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

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

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

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

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