

# Learning beautiful (and ugly) attributes

Luca Marchesotti  
luca.marchesotti@xerox.com  
Florent Perronnin  
florent.perronnin@xerox.com

XRCE  
Xerox Research Centre Europe  
Meylan, France

---

## Abstract

Current approaches to aesthetic image analysis either provide accurate *or* interpretable results. To get both accuracy *and* interpretability, we advocate the use of *learned* visual attributes as mid-level features. For this purpose, we propose to discover and learn the visual appearance of attributes automatically, using the recently introduced AVA database which contains more than 250,000 images together with their user ratings and textual comments. These learned attributes have many applications including aesthetic quality prediction, image classification and retrieval.

## 1 Introduction

The amount of visual content we handle on a daily basis has grown exponentially. In this ocean of images and videos, there are many questions that artificial systems could help us answer. In the last decade, the focus of the computer vision community had been on semantic recognition. While this is still a very active research field, new questions are arising. For instance, we might want to predict what people like in an image or a video. Although this is a very challenging question, even for humans, it was shown experimentally that aesthetics/preference can be predicted using data-driven approaches [5, 6, 7, 13, 17, 18, 20, 29].

Early work on aesthetic prediction [5, 13] proposed to mimic the best practices of professional photographers. In a nutshell, the idea was (i) to select rules (*e.g.* “contains opposing colors”) from photographic resources such as [14] and (ii) to design for each rule a visual feature to predict the image compliance (*e.g.* a color histogram). Many subsequent works have focused on adding new photographic rules and on improving the visual features of existing rules [7, 17]. As noted for instance in [7] these rules can be understood as visual attributes [9, 10, 15], *i.e.* medium-level descriptions whose purpose is to bridge the gap between the high-level concepts to be recognized (beautiful *vs.* ugly in our case) and the low-level pixels. However, there are at least two issues with such an approach to aesthetic prediction. Firstly, the hand-selection of attributes from a photographic guide is not exhaustive and does not give any indication of how much, and when, such rules are used. Secondly, hand-designed visual features only imperfectly model the corresponding rules. As an alternative to rules and hand-designed features, it was proposed in [18] to rely on generic features such as the GIST [22], the bag-of-visual-words (BOV) [4] or the Fisher vector (FV) [28]. While it was shown experimentally that such an approach can lead to improved results with respect to hand-designed attribute techniques, a major shortcoming

is that we lose the interpretability of the results. In other words, while it is possible to say that an image has a high or low aesthetic value, it is impossible to tell why. We thus raise the following question: *can we preserve the advantages of generic features and get interpretable results?* In this work, we will address this problem by *discovering and learning attributes automatically*. We note that there is a significant body of work on attribute learning in the computer vision and multimedia literature. This is a cost-effective alternative to hand-listing attributes [10, 15] and to architectures which require a human-in-the-loop [25]. Existing solutions [1, 34, 35] were typically developed for visual object recognition tasks. [34] proposes to mine pre-existing natural language resources. [1] uses mutual information to learn attributes relevant for e-commerce categories (handbags, shoes, earrings and ties) [8] uses latent CRF to discover detectable and discriminative attributes. Moreover, approaches such as [31] use natural language text under the form of caption or surrounding image text. Only [23] takes into text account to devise attributes, but the process is entirely manual.

**Contribution.** Our main contribution is a novel approach to aesthetic image analysis which combines the benefits of “attribute-based” and “generic” techniques. It consists of (i) automatically discovering a vocabulary of visual attributes and (ii) learning their visual appearance using generic features. For this purpose, we leverage the AVA dataset [20] which contains more than 250,000 images together with their aesthetic preference ratings and textual comments. Preference ratings allow us to supervise the creation of the attribute vocabulary (step (i)) and to learn automatically the visual appearance of attributes (step (ii)). Our second contribution is the application of the learned attributes to three different scenarios: aesthetic quality prediction, image classification and retrieval.

The remainder of this work is organized as follows: we first briefly introduce the AVA dataset and explain why it is an appropriate resource for aesthetic attribute learning (section 2). We then introduce the proposed approaches to discover attributes that consist of (i) mining visual attributes using the textual comments and the user ratings (section 3) and (ii) learning the visual appearance of the discovered attributes using generic features (section 4). In section 5, we show practical application of our learned attributes.

## 2 The AVA database

We use AVA, a recently introduced database [20] which contains more than 250,000 images downloaded from WWW.DPCHALLENGE.COM. An interesting characteristic of this dataset is that images are accompanied by natural language text and attractiveness scores. This dataset was assembled for large-scale evaluation of attractiveness classification and regression tasks. But it was also recently used to study the dependence of attractiveness on semantic information [19]. Another peculiarity of this corpus is the organization of photos in *contests*: an equivalent of Flickr groups where images are ranked according to attractive-



Figure 1: Sample photos from the challenge “Green Macro”: images ranked high in the contest (top row) better represent the visual concept “Green Macro”; they have more vivid colors and better technique than the ones at the bottom of the rank (2nd row).

Statistics	During challenge	After challenge	Overall
comments per image ( $\mu$ and $\sigma$ )	9.99 (8.41)	1.49 (4.77)	11.49 (11.12)
words per comment ( $\mu$ and $\sigma$ )	16.10 (8.24)	43.51 (61.74)	18.12 (11.55)

Table 1: Statistics on comments in AVA. On average, an image tends to have about 11 comments, with a comment having about 18 words on average. As the statistics in columns 2 and 3 attest however, commenting behavior is quite different during and after challenges.

ness scores left by users. Consider the sample images in figure 1, they were taken from the contest “**Green Macro**: Get up close and personal with the subject of your choice, using green as your primary color”. Photos in the first row scored highly, the others were ranked at the bottom of the contest. While all six images contain a lot of green, the top ones have brighter, more vivid green elements and the photographic technique “Macro” is much better represented. It is worth noting that more than 1,000 contests such as “**Green Macro**” are available. To give an example of the textual data in AVA, we also report a selection of critiques associated to the top-left photo of figure 1: *scooter88 says..: “Nice leading line. like the placement of grasshopper, well done!”*, *nam says..: “Love the colors, light and depth of field on this, but it’s the perspective that reeled me in 10”*, *Kroburg says..: “Really great picture, love the composition,..Great composition”*. In Table 1, we report statistics about AVA critiques. As can be seen, users tend to comment mainly when the photographic challenge is taking place but on average they tend to leave longer comments when the challenge is over. AVA contains 2.5 million of such textual comments, a veritable gold mine of photographic knowledge aligned with visual data. Another type of annotation which is available in AVA is the set of attractiveness scores given by the users of WWW.DPCHALLENGE.COM. In Figure 2, the dotted line represents the distribution of votes of for all images in AVA. Among the voters, we identified the population of voters that left a comment (commentators) and we plotted their votes distribution. Commentators seem to be the most generous while judging the photos. But the distribution has also higher variance which might imply higher noise or higher divergence of opinion.

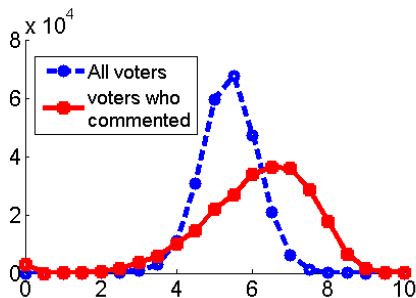


Figure 2: Distribution of scores by population of annotators: participants to challenges give, on average, higher scores to images.

### 3 Discovering beautiful (and ugly) attributes

As mentioned earlier, mining attributes by hand-picking photographic rules from a book is problematic: this is a non-exhaustive procedure and it does not give any indication of how much, and when, these techniques should be used. Therefore, we intend to discover attributes using data. Following [26], “Attributes represent a class-discriminative, but not class-specific property that both computers and humans can decide on”. Such a statement implies that attributes should be understandable by humans. A natural way to enforce inter-

T3:	ribbon, congrats, congratulations, deserved, first, red, well, awesome, yellow, great, glad, fantastic, excellent, page, wonderful, happy
T11:	beautiful, wow, amazing, congratulations, top, congrats, finish, love, stunning, great, wonderful, excellent, awesome, perfect, fantastic, gorgeous, absolutely, capture
T28:	idea, creative, clever, concept, cool, executed, execution, original, well, great, pencil, job, creativity, thought, top, work, shannon, interesting, good
T20:	funny, lol, laugh, hilarious, humor, expression, haha, title, fun, made, oh,love,smile, hahaha, great
T35:	motion, panning, blur, speed, movement, shutter, moving, blurred, abstract, blurry, effect, pan, stopped, sense, camera, fast, train, slow, background, exposure
T27:	colors, red, colours, green, abstract, color, yellow, orange, beautiful, colour, border, vibrant, complementary, composition, leaf, lovely, love, background, bright, purple
T49:	selective, desat, desaturation, red, use, color, works, processing, desaturated, saturation, editing, fan
T8:	portrait, eyes, face, expression, beautiful, skin, hair, character, portraits, eye, smile, nose, lovely, self, girl, look, wonderful, great, lighting, crop
T14:	cat, cats, kitty, eyes, fur, pet
T37:	sign, road, signs, street, stop

Table 2: Sample topics generated by pLSA for  $K = 50$  topics.

pretability is to discover attributes from natural text corpora, as done for instance in [1]. In our case, we use as a textual resource the user comments of the AVA dataset since they contain very rich information about aesthetics. However, such comments are quite noisy: they can be very short as shown in the previous section and they are written in a very spontaneous manner. This makes our task particularly challenging.

This section is organized as follows. We firstly describe how the textual data is pre-processed. We then describe a first approach to attribute discovery which is fully unsupervised as it only relies on comments. We show its limitations and then propose a supervised approach which relies on the user ratings.

### 3.1 Text pre-processing

We merge all the critiques related to an image into a single textual document. Merging the generally very short and noisy comments averages noise and thus leads to a more robust representation. We tokenize and spell-check each document and we remove stop-words and numbers. Each document is represented as a bag-of-words (BOW) histogram using the term frequency-inverse document frequency weighting (tf-idf). Hence, each commented image is associated with a bag-of-words vector.

### 3.2 Unsupervised attributes discovery

As a first attempt to discover attributes, we use the unsupervised Probabilistic Latent Semantic Analysis (pLSA) [11] algorithm on the BOW histograms. The hope is that the learned topics correlate with photographic techniques and therefore they are interpretable as attributes. In Table 2, we report some of the most interpretable topics discovered by pLSA with  $K = 50$  hidden topics. We can see that some topics relate to general appreciation and mood (T3, T11, T28, T20), to photographic techniques and colors (T35, T27, T49) or to semantic labels (T8, T14, T37). Despite the relevance of these topics to visual attractiveness, we cannot directly use them as attributes: they are too vague (i.e. not granular enough) and much manual post-processing would be needed to extract something useful. Experiments with different numbers of topics  $K$  did not lead to more convincing results.

### 3.3 Supervised attributes discovery

We devise an alternative strategy based on the following intuition: we use the attractiveness scores as a *supervisory information* to mitigate the noise of textual labels. The hope is that using attractiveness scores we will be able to identify interpretable textual features that are highly correlated with aesthetic preference and use them to predict aesthetic scores.

**Learning regression parameters.** We mine beautiful and ugly attributes by discovering which terms can predict the aesthetic score of an image. For this purpose, we train an Elastic Net [36] to predict aesthetic scores and, at the same time, select textual features. It is a regularized regression method that combines an  $\ell_2$ -norm and a sparsity-inducing  $\ell_1$ -norm. Let  $N$

<b>UNIGRAMS+</b>	great (0.4351), like (0.3301), excellent (0.2943), love (0.2911), beautiful (0.2704), done (0.2609), very (0.2515), well (0.2465), shot (0.2228), congratulations (0.2223), perfect (0.2142), congrats (0.2114), wonderful (0.2099), nice (0.1984), wow (0.1942), one (0.1664), top (0.1651), good (0.1639), awesome (0.1636),
<b>UNIGRAMS-</b>	sorry (-0.2767), focus (-0.2345), blurry (-0.2066), small (-0.1950), not (-0.1947), don (-0.1881), doesn't (-0.1651), flash (-0.1326), snapshot (-0.1292), too (-0.1263), grainy (-0.1176), meet (-0.1122), out (-0.1054), try (-0.1041), low (-0.1013), poor (-0.0978), distracting (-0.0724),
<b>BIGRAMS+</b>	well done (0.6198), very nice (0.6073), great shot (0.5706), very good (0.3479), great job (0.3287), your top (0.3262), my favorites (0.3207), top quality (0.3198), great capture (0.3051), lovely composition (0.3014), my top (0.2942), nice shot (0.2360), th placing (0.2330), great lighting (0.2302), great color (0.2245), excellent shot (0.2221), good work (0.2218), well executed (0.2069), great composition (0.2047), my only (0.2032)
<b>BIGRAMS-</b>	too small (-0.3447), too blurry (-0.3237), not very (-0.3007), does not (-0.2917), not meet (-0.2697), wrong challenge (-0.2561), better focus (-0.2280), not really (-0.2279), sorry but (-0.2106), really see (-0.2103), poor focus (-0.2068), too out (-0.2055), keep trying (-0.2026), see any (-0.2021), not sure (-0.2017), too dark (-0.2007), next time (-0.1865), missing something (-0.1862), just don (-0.1857), not seeing (-0.1785)

Table 3: Most discriminant unigrams and bigrams with their regression coefficient  $\beta$ . Bigrams are in general more interpretable than unigrams since they can capture the polarity of comments and critiques.

be the number of textual documents. Let  $D$  be the dimensionality of the BOW histograms. Let  $\mathbf{X}$  be the  $N \times D$  matrix of documents. Let  $y$  be the  $N \times 1$  vector of scores of aesthetic preference (the score of an image is the average of the scores it received). We learn:

$$\hat{\beta} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \quad (1)$$

where  $\lambda_1$  and  $\lambda_2$  are the regularization parameters.

**Selecting discriminative textual features.** We first experiment with a vocabulary of  $D \approx 30,000$  unigrams. We cross-validated the regularization parameters using Spearman’s  $\rho$  correlation coefficient and we selected the values of  $\lambda_1$  and  $\lambda_2$  providing highest performances with 1,500 non-zero  $\beta$  coefficients. We analyze the candidate labels by sorting them according to  $|\beta|$  (see Table 3) to verify their interpretability. By inspecting the most discriminant *unigrams*, we can see that the ones at the top of each rank relate to specific visual attributes (e.g. grainy, blurry). But others can be ambiguous (e.g. not, doesn’t, poor) and interpreting them is rather problematic.

These nuances of language can be resolved by looking at n-grams and especially at bigrams. This is a popular choice in opinion mining [24] since bigrams capture *non-compositional* meanings that a simpler feature does not [30]. For instance the word “lighting” does not have an intrinsic polarity while a bigram composed by “great” and “lighting” can successfully clarify the meaning. Hence, we performed regression on a set of  $D = 90,000$  bigrams using the same procedure employed for unigrams. If we look at the bottom rows of Table 3 we can see the bigrams which receive the highest/lowest regression weights. As expected, regression weights implicitly select those features as the most discriminant ones for predicting attractiveness. The highest weights correspond to “beautiful” attributes while the lowest weights correspond to “ugly” attributes. It is noteworthy that we use an Elastic Net to overcome the limitations of other sparsity-inducing norms like LASSO [33] in the feature selection tasks: if there is a group of features among which the pairwise correlations are very high, then the LASSO tends to select only one random feature from the group [36]. In our case, LASSO produces a compact vocabulary of uncorrelated attribute labels, but also a very small number of labeled images. This is problematic because we need as many annotated images as possible at a later stage to train one visual classifiers for each attribute.

**Clustering bigrams.** The effect of the Elastic Net on correlated features can be seen by looking at table 3: as expected, the Elastic Net tolerates correlated features (synonym bigrams) such “well done” or “very nice”, “beautiful colors” and “great colors”. This augments the number of annotated images, but it obliges us to handle synonyms in the vocabulary of at-

tributes. For this reason, we compact the list of 3,000 candidate bigrams (1,500 for Beautiful attributes and 1,500 for Ugly attributes) with Spectral Clustering [21]. We cluster the beautiful and ugly bigrams separately. We heuristically set the number of clusters to 200 (100 Beautiful and 100 Ugly clusters) and we create the similarity matrices with a simple but very effective measure of bigram similarity: we calculate the Levenshtein distance among the second term within each bigram and we discard the first term. This approach is based on the following intuition: most part of the bigrams are composed by a first term which indicates the polarity and a second term which describes the visual attributes *e.g.* “lovely composition”, “too dark”, “poor focus”. What we obtain is an almost duplicate-free set of attributes, and a richer set of images associated with them. Some sample clusters are reported here below:

C18: ['beautiful', 'colors'] ['great', 'colors'] ['great', 'colours'] ['nice', 'colors']

C56: ['challenge', 'perfectly'] ['just', 'perfect']

C67: ['nicely', 'captured'] ['well', 'captured'] ['you', 'captured']

C89: ['excellent', 'detail'] ['great', 'detail'] ['nice', 'detail']).

**Attribute discriminability.** To validate the relevance of the discovered attributes (beyond the qualitative inspection of Table 3), we used them in conjunction with the learned regressors  $\hat{\beta}$  to predict aesthetic preference scores from textual comments. We use Spearman’s  $\rho$  score to measure the correlation between the ground truth image ranking (deduced from the attractiveness scores) and the predicted ranking. We obtain a 0.821 value. These results can be compared to the baseline of [32] which relies on features specifically designed to capture opinions in comments. They report a score of 0.584 which is significantly lower. This shows that our learned attributes can be used to predict attractiveness, thus validating their usefulness for our task.

## 4 Learning the visual appearance of attributes

We randomly draw a bigram from each cluster to name the corresponding attribute. Since we have 200 attributes in total, it is difficult to hand-design a different visual classifier for each attribute. Therefore, we propose to learn such attribute classifiers from generic features. Given the large number of images available in AVA (approx. 250,000) and the large number of attribute classifiers to be learned, it is fundamental to employ a scalable solution. In what follows, we firstly describe the chosen generic features as well as the learning process. We then explain how attributes are re-ranked based on visualness.

**Learning visual attributes.** We extract 128-dim SIFT [16] and 96-dim color descriptors [3] from 24x24 patches on dense grids every 4 pixels at 5 scales. We reduce dimensionality by using a 64-dim PCA. These low-level descriptors are aggregated into an image-level signature using the Fisher Vector which has been shown to be the state-of-the-art for semantic [27] as well as aesthetic tasks [18]. We use visual vocabularies of 64 Gaussians and we employ a three-level spatial pyramid (1x1, 2x2, 1x3). We compute one SIFT and one color FV per image and we concatenate them. This leads to a combined 131,072-dim representation which is PQ-compressed [12] to reduce the memory footprint and to enable all images to be kept in RAM. We learn linear classifiers using a regularized logistic regression objective function and Stochastic Gradient Descent (SGD) [2] learning. Using a logistic loss (rather than a hinge loss for instance) provides a probabilistic interpretation of the classification scores, which is a desirable property since we are training attributes. It is worth noting that by experimenting with several feature configurations we appreciated the importance of



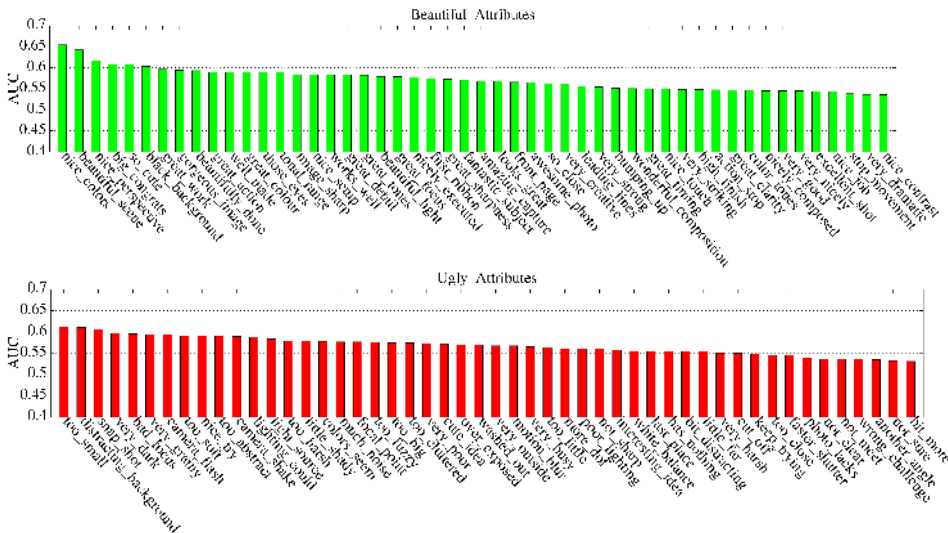


Figure 3: Area Under the Curve (AUC) calculated for the top 50 Beautiful and Ugly attributes.

color features for the classification of attributes. This is not surprising since many attributes are indeed color names or color properties. A second important consideration is that 64 Gaussians is a reasonable trade-off between precision and computational complexity of the features. We also compared the performances of two learning approaches: 1-vs-rest against multi-class classifiers. The former strategy provided better results experimentally.

**Re-ranking attributes.** In the previous section, we enforced interpretability and discriminability of the attribute labels using attractiveness scores as a supervision mechanism. However, this choice does not ensure that all these attributes can be recognized by a computer. This is the reason why we measure “visualness” using Area Under the ROC Curve (AUC) calculated for each individual attribute. In particular, we benchmark the classification performances of each attribute (1-vs-all) and we rank them using AUC. We show the top 50 attributes in Figure 3 for Ugly and Beautiful attributes. Our first observation is that performances of beautiful attributes are higher than ugly attributes. This is not surprising since the latter attributes were trained with fewer images: people prefer to comment on high quality images and as a consequence we are able to discover fewer ugly attribute labels. Second, we notice that attributes which detect lighting conditions and colors (e.g. *too dark*, *great colour*, *too harsh*) perform better than more complex visual concepts such as *interesting idea*, *bit distracting*, *very dramatic*.

## 5 Applications

We now consider three applications of the proposed attributes. **Aesthetic prediction.** In some cases, we might be interested in giving a binary answer regarding the attractiveness of an image: beautiful vs ugly. We therefore propose to use our learned attributes to make such a prediction and compare to the approach of [18] which is based on generic image features and it is to date the most performing baseline on AVA dataset. To make the comparison with [18], we use exactly the same FV generic features in both cases. As can be seen in figure





 <p>great_macro, very_pretty, great_focus, nice_detail, so_cute</p>	 <p>great_capture, great_angle, nice_perspective, lovely_photo, nice_detail</p>	 <p>more_dof, not_sure, too_busy, motion_blur, blown_out</p>	 <p>soft_focus, not_sure, more_light, sharper_focus, more_dof</p>
--	---	---	---

Table 4: Sample results for an image annotation application where the aesthetic quality of each image is described using the 5 most reactive attributes.

4, attributes perform comparably to low-level features, despite the significant difference in dimensionality (131,072 dimensions for the low-level features and 200 dimensions for the attributes). The small price paid in performance (AUC from 0.715 to 0.704) is compensated for the possibility of replacing a single image attractiveness label (good or bad) with the labels of the most responsive attributes.

**Image-tagging.** We now go beyond tagging an image as *beautiful* or *ugly* as such a binary decision can be too aggressive for a subjective problem such as aesthetic quality.

It could form a positive or negative prior in the user’s mind in contradiction to his/her tastes and opinions. To gain user’s consensus we design an application that not only predicts aesthetic quality (*Is this image beautiful or ugly?*) but also produces a qualitative description of the aesthetic properties of an image in terms of beautiful/ugly attributes. As can be seen from the examples of Table 4, this strategy gives the user higher degree of interpretation of the aesthetic quality. For instance, while many users might agree that the leftmost image is a beautiful picture, others might disagree that the yellow flower on the right is ugly:

in general people tend to refuse criticism. Instead, with attributes such as *more light*, *more depth field of view* and *not sure* the application takes a more cautious approach and enables the user to form his/her own opinion. Finally, we realize that these are just plausible hypotheses that should be tested with a full-fledged user study. However such an evaluation is out of the scope of this work.

**Image retrieval.** We now show how the learned attributes can be used to perform attribute-based image retrieval. We display the top-returned results of several queries for Beautiful and Ugly attributes in the mosaic of Figure 5. We notice that the images clearly explain the labels discovered in AVA even for fairly complex attributes such as *too busy*, *blown out*, *white balance* (note the various kind of color casts present in the images of row 6) or *Much noise* in the last row. In the attribute *nice perspective* we can observe what might be a limitation of the presented approach: it can be affected by a semantic bias. In other words, instead of learning the concept nice perspective, we might be learning the concept

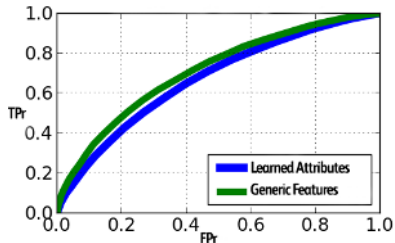


Figure 4: Aesthetic preference prediction: comparison between learned attributes and generic features (SIFT+color [18])



*building*, a semantic concept where, in general we have a great deal of perspective. This limitation can be overcome by designing learning strategies that take into account semantic labels (which are present in AVA).

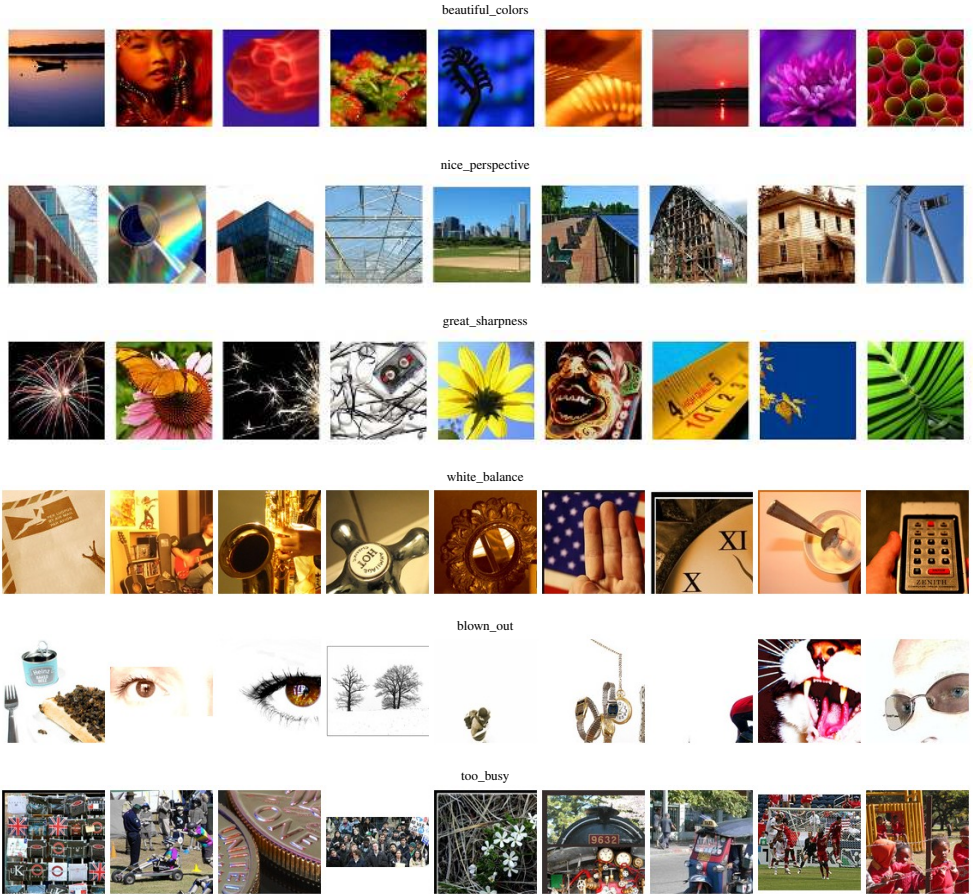


Figure 5: Images with top scores for some representative beautiful and ugly attributes.

## 6 Conclusions

In this paper, we tackled the problem of visual attractiveness analysis using visual attributes as mid-level features. Despite the great deal of subjectivity of the problem, we showed that we can learn automatically meaningful attributes that can be used in various applications such as score prediction, auto-tagging or retrieval. Future work will focus on testing with users the advantage of our beautiful and ugly attributes and on mitigating biases introduced by semantic information.

## 7 Acknowledgements

The authors would like to thank Jean-Michel Renders for the discussions about text analysis and Isaac Alonso for having supported the experimental work of this paper.

## References

- [1] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. *ECCV*, 2010.
- [2] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2007.
- [3] S. Clinchant, G. Csurka, F. Perronnin, and J.-M. Renders. Xrce participation to ImageEval. In *ImageEval Workshop at CVIR*, 2007.
- [4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, 2004.
- [5] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.
- [6] R. Datta, J. Li, and J.Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, 2008.
- [7] S. Dhar, V. Ordonez, and T.L. Berg. High-level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.
- [8] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.
- [9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.
- [10] V. Ferrari and A. Zisserman. Learning visual attributes. *NIPS*, 2007.
- [11] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.
- [12] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 2011.
- [13] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.
- [14] Kodak. *How to take good pictures : a photo guide*. Random House Inc, 1982.
- [15] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.
- [16] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.
- [17] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008.
- [18] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011.
- [19] N. Murray, L. Marchesotti, and F. Perronnin. Learning to rank images using semantic and aesthetic labels. In *BMVC*, 2012.

- [20] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.
- [21] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *NIPS*, 2002.
- [22] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001.
- [23] R. Orendovici and J.Z. Wang. Training data collection system for a learning-based photographic aesthetic quality inference engine. In *ACM-MM*, 2010.
- [24] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2012.
- [25] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.
- [26] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.
- [27] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
- [28] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.
- [29] J. Li R. Datta and J. Z. Wang. Learning the consensus on visual quality for next-generation image management. In *ACM-MM*, 2007.
- [30] E. Riloff, S. Patwardhan, J. Wiebe, et al. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006.
- [31] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where—and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.
- [32] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. In *WWW*, 2012.
- [33] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.
- [34] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.
- [35] K. Yanai and K. Barnard. Image region entropy: a measure of visualness of web images associated with one concept. In *ACM-MM*, 2005.
- [36] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 2005.