Tell Me What You Like and I'll Tell You What You Are: Discriminating Visual Preferences on Flickr Data

Pietro Lovato[1], Alessandro Perina[2], Nicu Sebe[3], Omar Zandonà[1], Alessio Montagnini[1], Manuele Bicego[1], and Marco Cristani[1,4]

[1] University of Verona, Italy
[2] Microsoft Research, Redmond, WA
[3] University of Trento, Italy
[4] Istituto Italiano di Tecnologia (IIT), Genova, Italy

Abstract. The John Ruskin's 19th century adage suggests that personal taste is not merely an absolute set of aesthetic principles valid for everyone: actually, it is a process of interpretation which have also roots in one's life experiences. This aspect represents nowadays a major problem for inferring automatically the quality of a picture. In this paper, instead of trying to solve this age-old problem, we consider an intriguing, orthogonal direction, aimed at discovering how different are the personal tastes. Given a set of preferred images of a user, obtained from Flickr, we extract a pool of low- and high-level features; LASSO regression is then exploited to learn the most discriminative ones, considering a group of 200 random Flickr users. Such aspects can be easily recovered, allowing to understand what is the "what we like" which distinguish us from the others. We then perform multi-class classification, where a test sample is a set of preferred pictures of an unknown user, and the classes are all the users. The results are surprising: given only 1 image as test, we can match the user preferences definitely more than the chance, and with 20 images we reach an nAUC of 91%, considering the cumulative matching characteristic curve. Extensive experiments promote our approach, suggesting new intriguing perspectives in the study of computational aesthetics.

1 Introduction

People often get enjoyment from observing images and express preferences for some pictures over others. Surprisingly, there is as of yet no scientifically comprehensive theory that explains what psychologically defines such preferences [1]. However, certain guidelines which suggest principles of general gratification have been produced: for example, there has been an effort to infer the common aspects determining preference by checking whether average image preference for a group of observers can be reliably predicted from various factors in a test set [2]. Some of these guidelines have roots in the cognitive sciences: facial attractiveness of symmetric faces is one of the most known example [3]. For real-world scenes, there is high agreement in observer's preference ratings: factors such as

K.M. Lee et al. (Eds.): ACCV 2012, Part I, LNCS 7724, pp. 45–56, 2013.

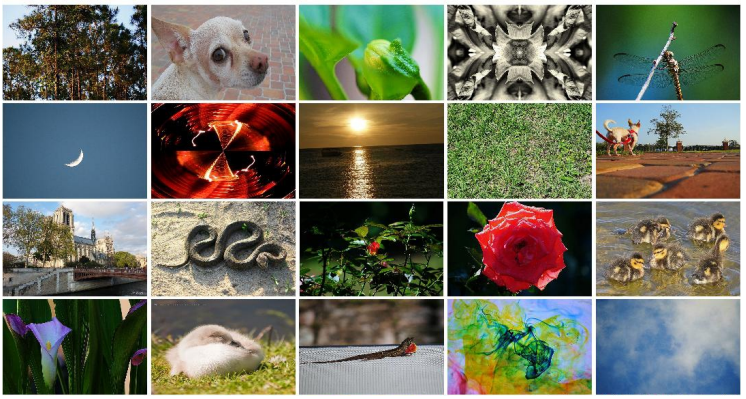

Fig. 1. Example of favourite images taken at random from a Flickr user

naturalness, complexity, coherence, legibility, vista, mystery and refuge seem to produce shared agreement [4], most probably due to the survival utility of a particular environment or viewpoint. Other guidelines are rules-of-thumbs (e.g., "rule of thirds","visual weight balance", etc. [5]) and all of them are modeled in a computational sense by the field of Computational Media Aesthetics (CMA) [6]. Many CMA applications have been developed: from aesthetic photo ranking [7,8] and preference-aware view recommendation systems [9], to picture quality analysis [10,11].

Nevertheless, these technologies ignore the potential role that factors internal to the observer may have on preference, summarized by the old adage "beauty is in the eye of the beholder". Recent studies have shown that preference formation is a result of the interplay between subjective novelty, e.g. how new a visual stimulus seems to an observer, and how well the observer is able to extract the sense of a stimulus and to relate it to previous knowledge, defined as interpretability [12]. Unfortunately, so far no automatic mechanism has been capable to subsume such experiences, limiting the effectiveness of the CMA applications to the sole manipulation of widely-shared preferences. Therefore, in this paper we do *not* propose a strategy for assessing the quality of an image; instead, we consider a brand-new orthogonal direction, learning the "personal aesthetics traits" of people, i.e. those visual preferences that distinguish people from each other. In particular, we take a crowdsearch approach [13] and we focus on Flickr[1], a popular website where every user can select his/her preferred photos, by tagging them as "favorites". This creates, for every user, a set of favorite photos, which is often very heterogeneous and whose modeling/recognition goes beyond standard computer vision tasks such as object/scene recognition (see fig. 1 for an example).

In this paper, we analyze the "favorites set" of 200 users to infer about their *personal* aesthetics traits. To this aim, we characterize each image with different

[1] `http://www.flickr.com/`

features, ranging from low-level color/edge statistics up to more high-level and semantic descriptors such as object detectors and overall scene statistics. LASSO regression is then exploited to learn the most discriminative aesthetic attributes, i.e., the aspects that an user likes that distinguish her/him from the rest of the community: such aspects can be easily recovered and visualized.

In the experiments, we show that personal tastes act like a blueprint for a user, allowing to recognize him with high accuracy; in particular, given just one image of an unknown user (the test samples), you can guess her/his identity more than the chance, and this probability dramatically raises as you consider a higher number of images.

Summarizing, the contributions of this paper are two:

1. A novel research direction: instead of studying which are the commonly liked visual aspects of an image (which is what computational aesthetics does), we explore the opposite direction, i.e., what are those aspects that allow to distinguish different users.
2. An inference method based on LASSO regression applied to heterogeneous visual features, which allows to obtain good recognition scores, and is highly interpretable, giving a clear idea of the aesthetic traits that distinguish an user from the other.

2 The Proposed Approach

Our approach is composed by two steps: feature extraction and feature weighting. In the next sections we will thoroughly detail each one of them.

2.1 Feature Extraction

In this step, the aim is to extract as much information as possible from an image. The idea is that we are not interested in extracting the classic aesthetic qualities of an image, but the aspects that make an image good for particular users; being this last goal slightly different, many factors and dimensions of analysis can be taken into consideration, each one considered for the purpose of describing different aspects of an image. Therefore, we wanted to span our selection from simple and standard image descriptors up to complex and state-of-the art ones.

In the following, we explain the cues we focus on, being aware that the list is neither exhaustive nor the best possible one; actually, our aim is to investigate how they should be properly treated for the task-at-hand.

1. **Color.** We calculated the average intensity of each channel (in the original RGB space).

2. **Edges.** We focused on the presence or absence of edges, as well as their predominant direction. Does a user have a tendency to like images with trees (lots of vertical edges)? Or maybe he is more fascinated by a sunset on the ocean or by a flat landscape, where horizontal edges are in abundance.

We extracted horizontal and vertical edges using the Prewitt filter, whereas the total edges have been computed with the Canny edge detector. We considered the number of horizontal, vertical and total point of edges. To avoid the dependence from the possible different sizes of images, the number of edge's pixels has been normalized by the total image area.

3. **Textures.** The repeated structures, or texture, in an image may be another important aspect; to extract this information we employed MTEX toolbox [14,15]. One of its major function is the suitability in analyzing very sharp texture symmetries. Among the many indices that MTEX computes, we retained the one called "texture index", that summarizes in one value all the information of the model fitted by the algorithm onto the image.

   In addition to that, we calculate the entropy of the image, a statistical measure to characterize the homogeneousness of an image.

4. **Regions.** As shown in the recent work of [16,17], objects and scene semantics are very important to understand the subjective judgement of a picture. Following this, we performed image segmentation collecting some low-order statistics.

   We employed the mean shift segmentation algorithm [18], and in particular the EDISON implementation [19][2]. After segmenting an image we extracted *i)* the number of segments – measuring the regions "density" which characterizes each image – and *ii)* the average extension of the regions. All the values have been normalized w.r.t. the total image area.

5. **Objects.** Once again motivated by [16,17], we employed the Deformable Part Models [20,21] system to detect objects. The algorithm works by detecting and localizing a specific object (for example a plane, a cat, a chair or a person), through the use of a model learned from a set of training examples. The system can detect different objects; in our approach we used as features the number of times every detectable object is present in the image (for a complete list of all detectable objects see [20]); we also retained the average area (the algorithm gives also the bounding box of the detected objects), to guess if objects are more towards the background or the foreground.

6. **Faces.** As a particular class of objects – which detection has been largely studied in the field of biometrics – we extracted the number and sizes of the faces present in the image. We employed the standard Viola-Jones face detection algorithm [22] implemented in the OpenCV libraries[3].

7. **Scenes.** Finally, we focused on describing the semantic of the whole scene, rather than the semantic of the single objects which appears in it. A very

[2] The code is freely available on:
`http://coewww.rutgers.edu/riul/research/code/EDISON/index.html`

[3] `http://opencv.willowgarage.com/wiki/`

powerful scene descriptor is the GIST [23], which, roughly speaking, measures the responses of different Gabor filters. Such filters are built to describe the category of the scene in terms of openness, ruggedness, roughness and expansion[4].

The concatenation of all these descriptors, a vector $\mathbf{x}_m$ of 62 elements, represents the proposed signature for the image m. Since the value ranges are very heterogeneous, each feature/dimension is normalized across the images to have zero mean and unit standard deviation. More details are given in the experimental section.

It is worth noting that for the sake of reproducibility, every parameter of the different off-the-shelf computer vision libraries have been left as the default setting.

2.2 Feature Weighting

The main claim of this paper is that the discriminative aesthetical aspects of each user can be represented by a subset of all the features considered, opportunely weighted.

Given a pool of training images for N users, we perform a sparse regression analysis using Lasso [24]. Lasso is a general form of regularization in a regression problem. In the simple linear regression problem every training image, described by the proposed feature vector and denoted with $\mathbf{x}_m$, is associated with a target variable y_m (in our case, it can be the discrete label representing the user who posted it). Then, we can express the target variable as a linear combination of the image features:

$$y_m = \mathbf{w}^T \mathbf{x}_m \tag{1}$$

The standard least square estimate calculates the weight vector $\mathbf{w}$ by minimizing the error function

$$E(\mathbf{w}) = \sum_{m=1}^{M} \left(y_m - \mathbf{w}^T \mathbf{x}_m \right)^2 \tag{2}$$

where in our case M correspond to the total number of images we have in the training set. The regularizer in the Lasso estimate is simply expressed as a threshold on the L1-norm of the weight $\mathbf{w}$:

$$\sum_j |w_j| \leq t \tag{3}$$

This term acts as a constraint that has to be taken into account when minimizing the error function.

By doing so, it has been proved that (depending on the parameter t), many of the coefficients w_j become exactly zero [24]. Since each component w_j of

[4] code is publicly available on
`http://people.csail.mit.edu/torralba/code/spatialenvelope/`

the weight vector weigh a different feature, it is possible to understand which features are the most important for a given user, and which ones are neglected.

In our particular scenario, where the aim is to capture the facets of the visual aesthetics of a user which discriminate him from other people, we performed Lasso regression for each user separately, considering all training images coming from that user to have positive label. In other words, we have to solve N regression problems, each one returning a weight vector "user-specific" $\mathbf{w}^{(n)}$, $n = 1, \ldots, N$.

This way, by looking at the values in $\mathbf{w}^{(n)}$, we have that only the most important image features that characterize the preferences of the n-th user are retained.

3 Experiments

3.1 Data Collection

To test our approach, we consider a real dataset of 40.000 images, composed by 200 users chosen at random from the Flickr website. For each user, we retained the first 200 favourites[5].

3.2 Testing Protocol and Preliminary Evaluation

After computing all the images' signature, we randomly split the favorite images of each user into a training set and a testing set (here we used a 50%/50% splitting). Then, since the value ranges are very heterogeneous, each feature / dimension is normalized across all training images to have zero mean and unit standard deviation. The testing set is normalized with the constants calculated on the training set.

Then, as explained in the previous section, one Lasso regression is learned on the training set of each user, crossvalidating the parameter t. After estimating $\mathbf{w}^{(n)}$ for the n-th user, we calculated the regression scores $\beta_m^{(n)}$ for each testing image m by simply applying the product described by eq. 1:

$$\beta_m^{(n)} = \mathbf{w}^{(n)T}\, \mathbf{x}_m \tag{4}$$

As a preliminary evaluation, we used these regression scores β of the testing set to calculate a ROC curve for each user; basically, we are building a one-vs-all classifier that highlights the peculiar characteristics that distinguish a specific person. It turns out that Lasso is able to capture the differences between users, obtaining an average AUC of 69.4% with a standard deviation of 8%. Motivated by this promising result, we investigated in detail the issues discussed in the two next sections.

[5] The dataset is available upon request at
`http://profs.sci.univr.it/~{}cristanm/projects/perpre.html`

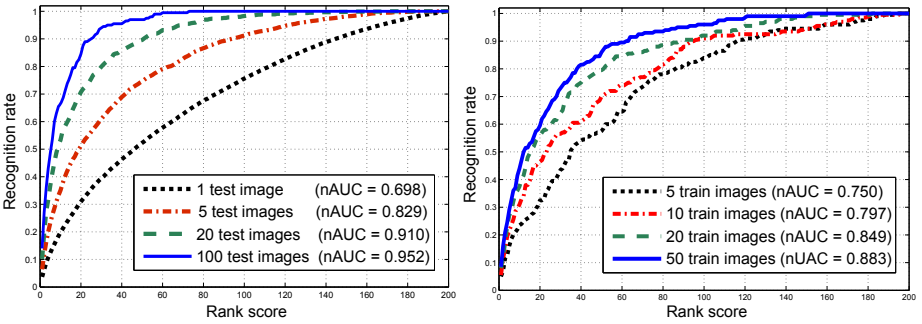

Fig. 2. CMC curves for our dataset. On the left: For each curve, we varied the number of testing images to be considered as a single "set". On the right: For each curve, we varied the number of images used to train Lasso.

3.3 Matching the Personal Aesthetics

In this section, we want to answer this question: how many images are needed to guess the personal aesthetics preferences of a person? Re-formulating the question, we want to prove if we are able to guess the user who tagged an image or a set of images. This task is intrinsically much more difficult than the previous one: instead of testing one user vs all the others, we are trying to predict which one tagged an image as favorite, on the basis of the "subjective" peculiar traits the image contains.

Intuitively, a single image does not contain every facet of the visual aesthetics sense of a person; the idea is to consider a *set* of testing images, and guess if the set contains enough information to catch the preferences of the user, allowing to identify him among all the others.

To do so, we exploit the fact that given a testing image $\mathbf{x}_m$ (or a set of images $\{\mathbf{x}_s\}_{s=1}^{S}$), we can evaluate – for each user n – its corresponding lasso estimate $\beta_m^{(n)}$ (or, in the set case, the mean of the estimates of each image $\frac{1}{S}\sum_s \beta_s^{(n)}$). This value can be used as a score to rank the different users. Hopefully, the user with highest score is the one who originally faved the photo (or group of photos).

To show the results, we build a CMC curve [25], a common performance measure in the field of re-identification [26]: given a test set of images coming from a single user and the membership score discussed, the curve tells the rate at which the correct user is found within the first k matches, with all possible k spanned on the x-axis. Figure 2 shows various CMC curves for our dataset.

On the left, we reported four different CMCs, trying to vary the parameter S, which tells how many images are aggregated to form a single test object.

From the figure it is evident that performing the task of identifying correctly a user with a single image (black dotted line) is very difficult. However, as soon as the number of testing images grouped together increase a little, a consistent improvement can be noted. This is in line with our hypothesis: we are aggregating information from heterogeneous images, each one characterizing only a small portion of the user subjective tastes.

On the right, we assessed the importance of the training set size by keeping the test parameter S fixed to 20. As expected, by lowering the number of training elements, it is more difficult to learn the users' preferences and their aesthetic sense uniqueness. For both figures, the normalized Area Under the Curve (nAUC) has been reported in the legend.

As a final comment, it is also worth noting that, even if at CMC rank 1 we achieve in the worst case a 3.9% rate of correct identification, this is higher that the probability we have to recognize the user by mere chance (which amounts to 0.5%).

3.4 Feature Analysis

This section is aimed at providing a qualitative evaluation of the proposed approach, showing that the regression score $\beta_m^{(n)}$ provides a valid measure of the preferences of an user, while the weight coefficients in the vector $\mathbf{w}^{(n)}$ provides an interpretable description for his visual aesthetic sense. First of all, we performed the following experiment: given a user, we considered all testing images m we have, and we sorted them according to their regression score $\beta_m^{(n)}$. The higher the score, the higher the probability that the user may have actually faved the image. Figure 3 briefly sketches the results; each column correspond to a Flickr user and the first 10 rows (before the white space) are favorite photos chosen from the training set of the user: we computed regression scores on the training set, and in the figure the top 10 are shown. The next images are the 10 testing images (coming from all users) with highest regression score; we highlighted with a blue box the ones that actually belong to the favorites of the user in hand.

The first column reveals some interesting information: although the highest test image for the user is not on his favorites set, it can have some visual appeals reflected on some of the images on his training set (see for example the lines on the spider web, or the red sunsets). It seems that a sort of "internal coherence" starts to show up.

We then looked into the weight coefficients for some users after learning the sparse regression model. For 3 random users, we reported the vector $\mathbf{w}$ in figure 4, on the right of their training and predicted preferred images. For visualization purposes, we labeled the 5 most prominent features (i.e. the ones with highest weight value): actually, the discriminative preferences of a user are determined by some features that have to be present in the pictures (high positive weight in the figure). As an example, by looking at the last bar-plot, it seems that faces, and some peculiarity in the textures are very important to the user; this is verified in his favorites, that contain faces, and gray-level images with a similar textural pattern.

As a final test, we have calculated the mean of the absolute values of the regression scores of all the subjects (Fig.5) This will show in general which are the most discriminative features for all the considered users. Interestingly, it seems that the low level features like color, texture, regions play a primary role compared to high-level cues like the scene or the objects.

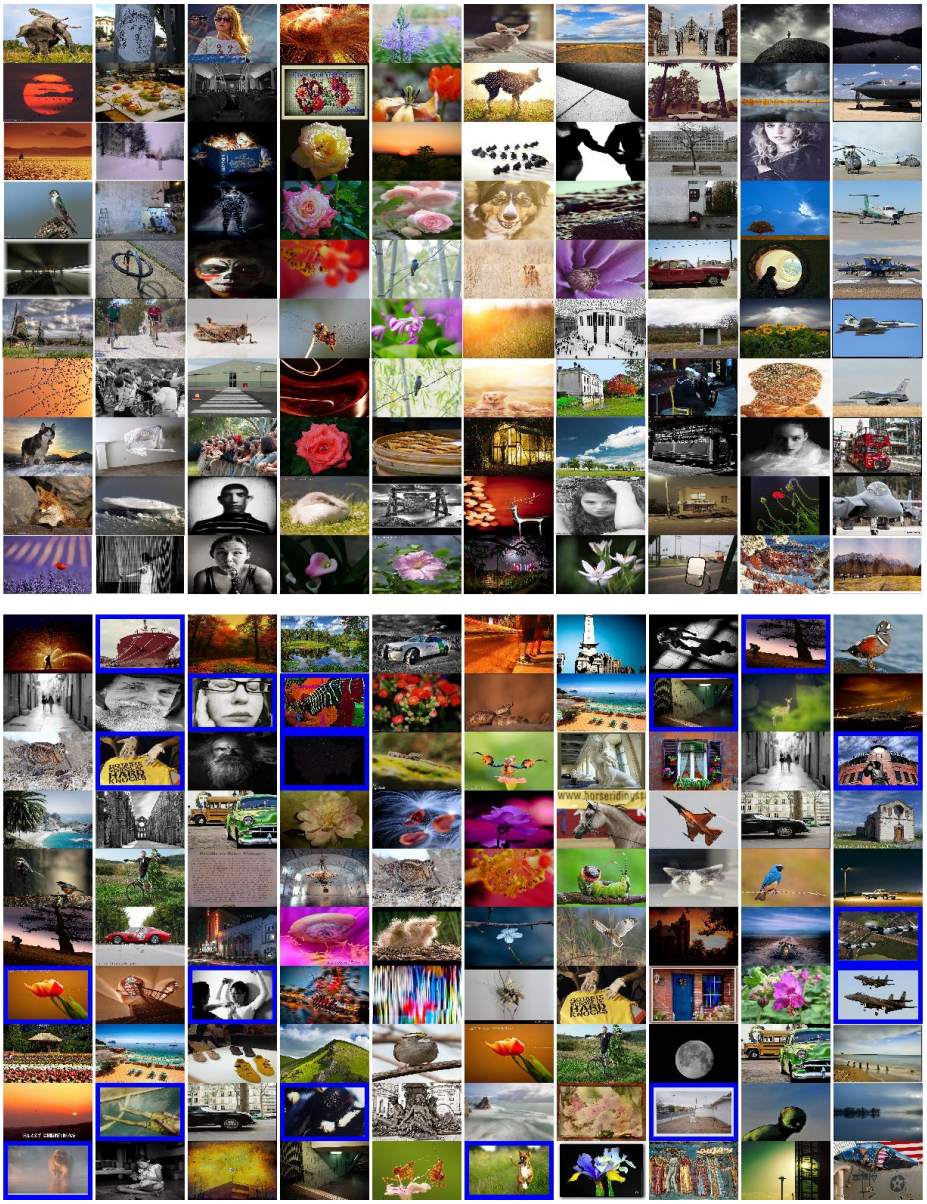

Fig. 3. Training and recognized testing images for different users. Each column is a user, and the first 10 images comes from his training set. In the half-bottom part, we show the first 10 testing images for that user, ranked on the basis of their regression score (the first being the one with highest score). In blue, correct matches are highlighted. A "coherence" between training and testing images can be seen.

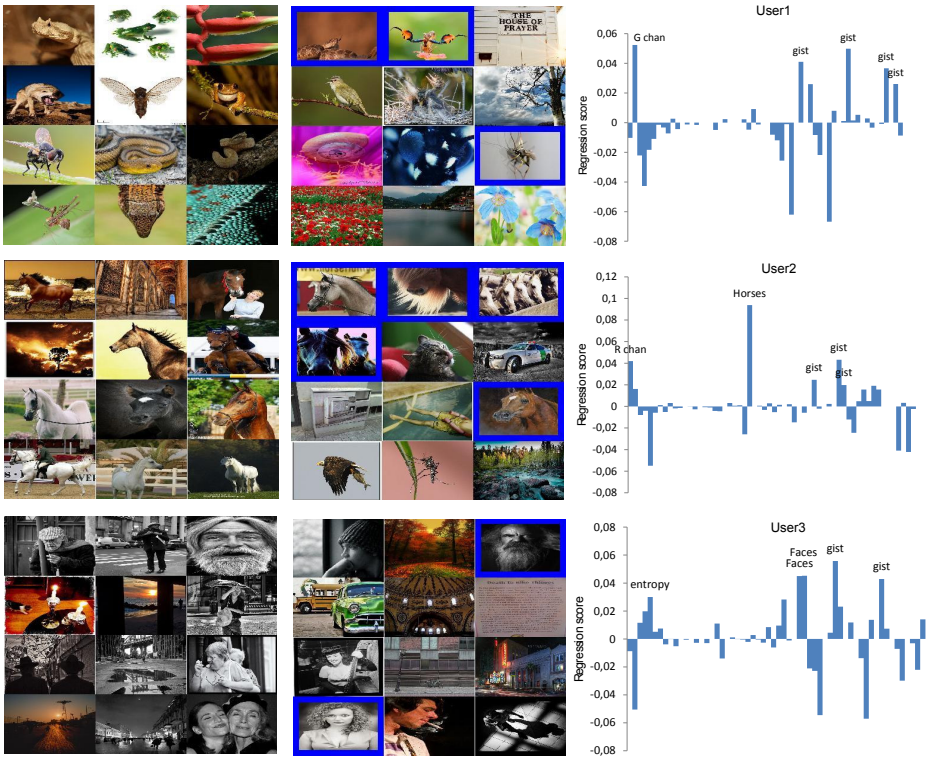

Fig. 4. Most prominent features for 3 users taken from the dataset. In the first two blocks of figures, training and testing elements are shown, in the same fashion of figure 3. On the right, for each user, a bar plot of each feature's importance is shown. The height of each bar represents the value of the corresponding weight.

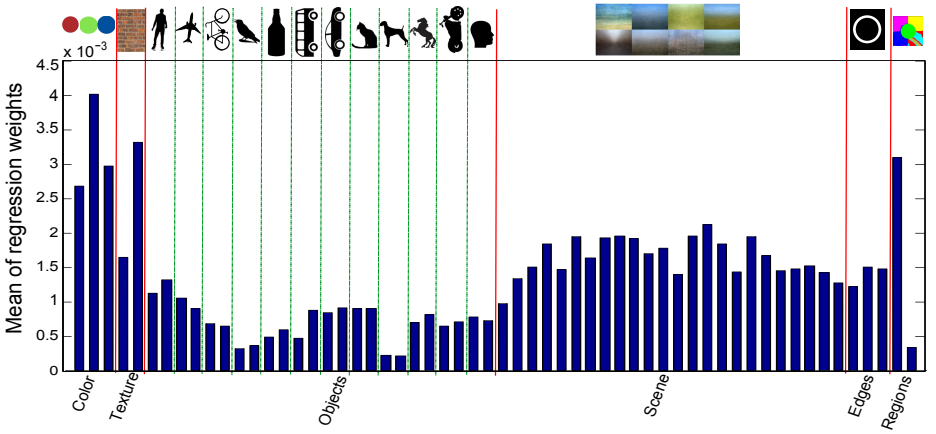

Fig. 5. Mean of the absolute values of the regression weights over all users in the dataset

4 Conclusions

This paper proposed an innovative way to deal with image aesthetics: instead of focusing on the design of common rules of likeliness, we proceeded in an orthogonal direction, modeling the aesthetics differences which characterize many users. This is possible by crowdsearching huge Internet databases as the Flickr repository: we considered 200 users, 40.000 different images. The results may be summarized as follows: different users have different preferences, and these preferences can be employed to identifying the users with high precision; this precision depends on how many images you take into account, but with even an image, you can match the user preferences with an accuracy higher than the chance. This opens up to a set of interesting applications: given a set of images on a publication, you can infer who are the users that most probably will like them all; this could be a novel kind of recommender system, which subsume the aesthetic statistics of a set of images, matching with the personal preferences of each one of us. More intriguing, what if we consider a huge number of users? Imagine to have millions of users, and to apply our framework: the idea is that the classical image aesthetics could be found by checking the features that are not discriminative (i.e., they are liked by everyone), while discriminative aspects could be seen as outlier aspects that make our preferences so unique. These two perspectives are actually under study, with promising preliminary results.

References

1. Leder, H., Belke, B., Oeberst, A., Augustin, D.: A model of aesthetic appreciation and aesthetic judgments. British Journal of Psychology 95, 489–508 (2004)
2. Martindale, C., Moore, K., Borkum, J.: Aesthetic preference: Anomalous findings for berlyne's psychobiological theory. American Journal of Psychology 103, 53–80 (1990)
3. Bronstad, P., Russell, R.: Beauty is in the "we" of the beholder: greater agreement on facial attractiveness among close relations. Perception 36, 1674–1681 (2007)
4. Kaplan, R., Kaplan, S.: The Experience of Nature: A Psychological Perspective. Cambridge University Press (1989)
5. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 271–280. ACM, New York (2010)
6. Adams, B.: Where does computational media aesthetics fit? IEEE Multimedia 10, 18–27 (2003)
7. Yeh, C.H., Ho, Y.C., Barsky, B.A., Ouhyoung, M.: Personalized photograph ranking and selection system. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 211–220. ACM, New York (2010)
8. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying Aesthetics in Photographic Images Using a Computational Approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part III. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
9. Su, H.H., Chen, T.W., Kao, C.C., Hsu, W.H., Chien, S.Y.: Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. IEEE Transactions on Multimedia 14, 833–843 (2012)

10. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR 2006, pp. 419–426. IEEE Computer Society, Washington, DC (2006)
11. Luo, Y., Tang, X.: Photo and Video Quality Evaluation: Focusing on the Subject. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 386–399. Springer, Heidelberg (2008)
12. Biederman, I., Vessel, E.: Perceptual pleasure and the brain. American Scientist 94, 1–8 (2006)
13. Bozzon, A., Brambilla, M., Ceri, S.: Answering search queries with crowdsearcher. In: WWW, pp. 1009–1018 (2012)
14. Hielscher, R., Schaeben, H.: A novel pole figure inversion method: specification of the *mtex* algorithm. Journal of Applied Crystallography 41, 1024–1037 (2008)
15. Bachmann, F., Hielscher, R., Schaeben, H.: Texture analysis with mtex–free and open source software toolbox. Solid State Phenomena 160, 63–68 (2010)
16. Isola, P., Jianxiong, X., Torralba, A., Oliva, A.: What makes an image memorable? In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 145–152 (2011)
17. Curran, W., Moore, T., Kulesza, T., Wong, W., Todorovic, S., Stumpf, S., White, R., Burnett, M.M.: Towards recognizing "cool": can end users help computer vision recognize subjective attributes of objects in images? In: ACM International Conference on Intelligent User Interfaces, pp. 285–288 (2012)
18. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619 (2002)
19. Georgescu, C.: Synergism in low level vision. In: International Conference on Pattern Recognition, pp. 150–155 (2002)
20. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Discriminatively trained deformable part models, release 4 (2010), `http://www.cs.brown.edu/~pff/latent-release4/`
21. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
22. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 511–518 (2001)
23. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42, 145–175 (2001)
24. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
25. Moon, H., Phillips, P.: Computational and performance aspects of pca-based face-recognition algorithms. Perception 30, 303–321 (2001)
26. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of British Machine Vision Conference (2011)