WE LIKE IT! MAPPING IMAGE PREFERENCES ON THE COUNTING GRID

P. Lovato[1] *A. Perina*[2] *D.S. Cheng*[3] *C. Segalin* [1] *N. Sebe* [5] *M. Cristani*[1,4]

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

ABSTRACT

Modeling preferences in photographic images is often reduced to analyzing intermediate explicit representations (e.g. textual tags) as means of capturing the objective and subjective properties of image perception, trying to distill the essence of what gives pleasure. We propose an alternative approach that bypasses the necessity to build an explicit conceptual coding of image preferences, operating directly on the raw properties of the images, extracted with heterogeneous feature descriptors. This is achieved through the *counting grid* model, which fuses together content-based and aesthetics themes into a 2D map in an unsupervised way. We show that certain locations in this map correspond to perceptually intuitive image classes, even without relying on tags or other user-defined information. Moreover, we show that users' individual preferences can be represented as distributions over the map, allowing us to evaluate the affinity between different users' appreciations. We experiment on a large Flickr dataset, clustering users by affinity, and validating these clusters by checking users that belong to the same Flickr photo groups.

Index Terms— Information fusion, image aesthetics, content-based image processing, counting grid.

1. INTRODUCTION

Nowadays, people commonly enjoy watching and sharing images or photos when browsing popular social networks, often expressing preferences for pictures they like. Being fascinated by a picture is clearly the result of a complex interplay between the undeniable aesthetics of an image (*i.e.*, if a photo is objectively beautiful), its content (some people prefer cars over flowers) and the subjective preferences of a person, which are affected by personal interpretation of visual stimuli, experience, mood, and so on [1]. Given these circumstances, can we design a computational model that reliably predicts our image preferences?

Many approaches in the literature only address the "aesthetics" side of the problem: from picture quality analysis [2, 3] to the wide field of computational media aesthetics [4, 5], these solutions mainly focus on discovering universal preferences, and devising objective criteria, inherently marginalizing the role of subjective factors. On the other hand, [6] focuses on characterizing each individual by his personal aesthetic taste, defined by the most discriminative image features that distinguish him from the rest of the community.

A different philosophy, represented recently by [7], relies on a content-based approach, which, from a large collection of tagged images, builds a graph of inter-related visual concepts for image clustering and annotation. While the user-supplied tags provide obvious help in determining objects, locations and situations within images, they scarcely assist in evaluating the (universal or individual) aesthetic value, either because it is highly emotional, or because it is minimally mediated by a semantic connotation.

In this paper, we bypass the problem of finding a comprehensive intermediate coding for conceptual images: instead of processing images to extract explicit aesthetics codes or content-based tags in an independent way, we directly exploit the information within the images to map heterogeneous image features together in a seamless way. In particular, we arrange a wide set of photos taken from the popular Flickr[1] social network in a 2-dimensional map through a *counting grid* [8]: this generative model considers each image as a specific distribution of generic features (color, SIFT features, etc.) and places it alongside similar images.

In our case, we adopt this model in an unconventional and novel way, feeding it with aesthetical and content-based features: the trained counting grid fuses the two worlds in an unsupervised way, thus defining a manifold where local regions mark semantic areas that smoothly transit between each other, and thus allowing fine thematic shifts. Besides revealing image classes, we can use this model to retrieve a Flickr user aesthetic profile, by locating his preferred images in the counting grid and building a map of his specific tastes. We used these profiles to calculate a novel "subjective aesthetics" metric between people, and we tested its usefulness by checking on users that subscribe to the same Flickr photo groups,

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

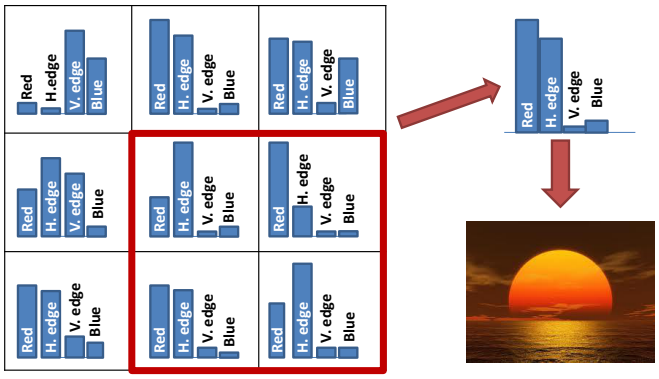

Fig. 1. Generating an image from a simple 3×3 counting grid: given a 2×2 window on the grid, we average the feature counts, obtaining a bag of features which corresponds to the final image.

under the assumption that they indicate shared tastes. In fact, this technique can be used to tell users about other groups they may like. Summarizing, in this paper we present:

- how to fuse in a principled way aesthetics traits and content-based features in a low-dimensional "spatial" latent manifold – the counting grid;
- how to highlight the image preferences of users directly on the counting grid, without the need of intermediate cross-modal representations (*e.g.*, text);
- a way to suggest and encourage groupings, by exploiting a novel distance defined over counting grids.

2. THE COUNTING GRID MODEL

The counting grid (CG) is a generative model recently introduced in [8] for analyzing images collections. It starts by assuming that images are represented as histograms $\{c_z\}$ over unordered bags of features, where c_z counts the occurrences of feature z.

Roughly speaking, a CG is a 2D finite discrete grid where each location $\mathbf{i} = (x, y)$ contains a normalized count of features $\pi_{\mathbf{i},z}$. The underlying generative process draws an image (*i.e.*, its bag of features $\{c_z\}$) by locating a small window in the grid, averaging the feature counts within it to obtain a local probability mass function over the features, and then generating from it an appropriate number of features in the bag (see Fig. 1). In other words, unlike a straightforward embedding (e.g. PCA) that links an image with a point location, the counting grid forces the image to link with a small window of locations. Given that the size $E_1 \times E_2$ of a counting grid is usually small compared to the number of images, this also forces windows linked to different images to overlap, and to co-exist by finding a shared compromise in the feature counts located in their intersection. The overall effect of these constraints is to produce locally smooth transitions between strongly different feature counts by gradually phasing features in/out in the intermediate locations.

Formally, the counting grid $\pi_{\mathbf{i},z}$ is a 2D finite discrete grid, spatially indexed by $\mathbf{i} = (x, y) \in [1 \dots E_1] \times [1 \dots E_2]$, and containing normalized counts of features indexed by z. Thus, we have $\sum_z \pi_{\mathbf{i},z} = 1$ everywhere on the grid. A given bag of features $\{c_z\}$ is generated by selecting a certain location $\mathbf{k}$, calculating the distribution $h_{\mathbf{k},z} = \frac{1}{W_n} \sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z}$ by averaging all the feature counts within the window $W_\mathbf{k}$ (with area W_n) that starts at $\mathbf{k}$, and then drawing features counts from this distribution.

In other words, the position of the window $\mathbf{k}$ in the grid is a latent variable; given $\mathbf{k}$, the likelihood of $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{c_z} = \alpha \prod_z \Big(\sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z} \Big)^{c_z}, \quad (1)$$

where α is a fixed normalization factor.

To learn a counting grid, we need to maximize the likelihood over all training images T, that can be written as

$$p(\{\{c_z^t\}, \mathbf{k}^t\}_{t=1}^T) \propto \prod_t \prod_z \Big(\sum_{\mathbf{i} \in W_{\mathbf{k}^t}} \pi_{\mathbf{i},z} \Big)^{c_z^t}, \quad (2)$$

which is intractable, much like in mixtures. Following [8, 9], we employ a variational EM algorithm that iteratively estimates $\hat{\pi}_{\mathbf{i},z}$ and $\{\hat{\mathbf{k}}^t\}$ (please check the mentioned papers for all the details). For our purposes, the most interesting outputs are the posterior probabilities $p(\mathbf{k}^t|\{c_z^t\})$, which localize each training image in the grid. By construction, similar images will be placed in the same location or nearby.

3. THE PROPOSED APPROACH

3.1. Feature extraction

Our goal is to manage highly heterogeneous image features, letting them smoothly interact in the CG. In particular, we use two families of descriptors: aesthetic and content-based.

Aesthetics has been defined as somewhat underlying the image semantics: we refer to [5, 10] for collecting a set of 16 aesthetic features. Content-based descriptors are essentially objects detectors, the same as those chosen in [6]. Table 1 report them schematically. In the end, each image is described by a count vector of 118 elements, which is adequate, but conceivably neither optimal nor exhaustive.

3.2. Counting grid training

Given the feature vectors, we train a CG. We cannot visualize the resulting 2D map directly (each location contains a bag of features), but we can create an image mosaic using images with the highest posteriors $p(\mathbf{k}^t|\{c_z^t\})$ at each location $\mathbf{k}$ in the map. Fig. 2 shows part of a 70×70 CG (more details later on).

Category	Name	Size	Short Description
Aesthetics	Use of light	1	Average pixel intensity of V channel [5]
	HSV statistics	8	Mean of H,S channels and std.dev of S,V channels; *angular dispersion*, *saturation weighted* and *without saturation*, mean H in IHLS color space [11]
	Emotion-based	3	Amount of pleasure, arousal, dominance [10, 12]
	Colorfulness	1	Level of colorfulness based on Earth Mover's Distance (EMD) [5, 10]
	Color Names	11	Amount of black, blue, brown, green, gray, orange, pink, purple, red, white, yellow [10]
	Textural features	2	Texture index [13, 14] of the image and entropy level
	Wavelet textures	12	Level of spatial smoothness measured with Daubechies wavelets on the three HSV channels [5]
	Tamura	3	Amount of coarseness, contrast, directionality [15]
	GLCM-features	12	Amount of contrast, correlation, energy, homogeneity for each channel HSV [10]
	GIST descriptors	24	Category of the scene in terms of level of openness, ruggedness, roughness and expansion [16].
	Edges	1	Total number of edge points, extracted with Canny
	Level of detail	1	Number of regions after mean shift segmentation [17, 18]
	Regions	1	Average extension of the regions (after mean shift segmentation) [17, 18]
	Low Depth of Field (DOF)	3	Amount of focus sharpness in the central part of the image wrt the overall focus [5, 10]
	Rule of thirds	3	Mean of H,S,V channel in the inner part of the image [5, 10]
	Image parameters	2	Size and aspect ratio of the image
Content	Objects	28	Objects detection [19]: we retained the number of instances and the average area
	Faces	2	Number and size of faces after Viola-Jones face detection algorithm [20]

Table 1. List of the features extracted in the proposed approach, divided between aesthetics and content-based. Each feature is a histogram of counts, or indicates the level of presence of the particular cue.

3.3. User analysis as inference in the CG

Given the trained CG, as a novel contribution for the CG model, we discover the preferences of a subject by aggregating the posteriors of his preferred images. Technically, we calculate the following *user map*

$$\gamma_{\mathbf{i},u} = \sum_{t \in T_u} \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} p(\mathbf{k}^t | \{c_z^t\}), \tag{3}$$

where u is a user and T_u identifies his set of images.In the case of a user from our training set, we already possess the posterior probabilities, while in the case of a new user, we can calculate them using the E-step formula. We can think of $\gamma_{\mathbf{i},u}$ as a summary of the personal preferences of a subject: it is an accumulation of the positions in the grid where the liked images for a unique individual u are placed. In practice, different locations in this user map correspond to different aesthetic characteristics of images: black-and-white photos of faces, sharp textured city landscapes, and so on.

Finally, given the compact representation of the users map, we can estimate the affinity between users by calculating the Euclidean distance of their user maps, and exploit such measures to cluster them together: the goal is to discover groups of people who share visual interests.

4. EXPERIMENTAL EVALUATION

To assess our proposal, we consider 200 users from Flickr, and for each one of them we pick randomly a pool of 200 favorites images. The resulting 40K images dataset has been processed by extracting the features of Table. 1. For the CG learning algorithm, the parameters of the model are the size of the grid and the size of the window. Different parameterizations offer the same conceptual analysis at different resolutions: in our case, we choose a 70×70 grid and 5×5 windows.

Fig. 2 shows a quarter of the trained CG: we can immediately observe that rich semantic groups pop out, highlighted in the lower part of the figure; we created an overlay, tentatively explaining the most evident themes in words, but observation alone gives a better picture of the thematic clusters.

Then, using Eq. 3, we can infer the personal preferences of the Flickr users. Fig. 3 shows three user maps with highly multimodal distributions that reveal personal tastes diversified over several themes. The first two users have similar maps, and a random sample of images shows very similar characteristics. If we were to describe in words their tastes, we could say that they like faces and yellow-tinted photos, but this would be undeniably very reductive. The user maps are a much more holistic representation of their tastes.

To quantitatively evaluate the expressiveness of the user maps, we calculate pairwise Euclidean distances between all users – called *CG distance* from now on. Then, we cluster them hierarchically (by average link), obtaining the dendrogram shown (partially) in Fig. 4 (left). On the right, we show the alternative dendrogram obtained by the Euclidean distance between each user's average feature vector (shortly, *mean feature distance*). In red, we highlight users that share various street and urban Flickr photo groups (mostly in black and white): the CG distance puts them close in the dendrogram, while the mean feature distance is not so informative.

As a further test, given the distance matrix between users (employing the CG distance and the mean feature distance), we computed the mean between every pair of users sharing

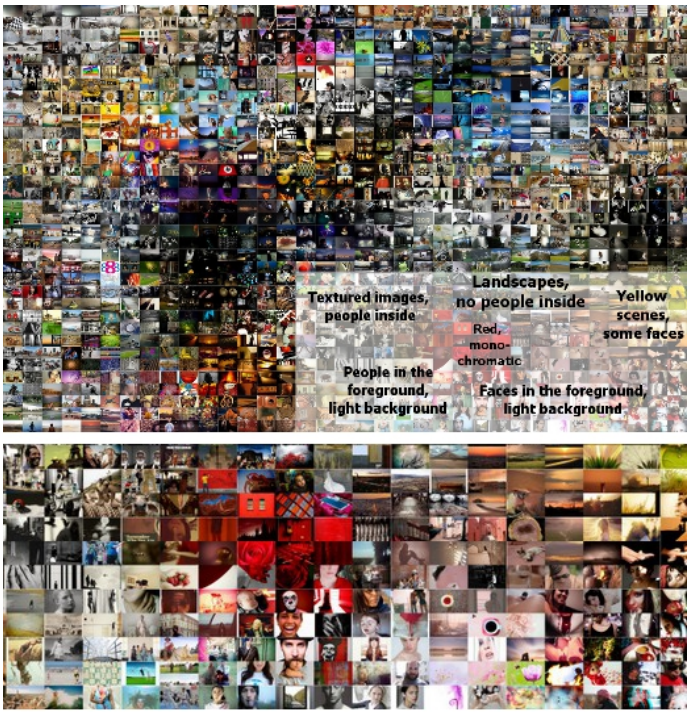

Fig. 2. On the top, part of a trained 70x70 CG: at each location we show the image with the highest posterior. Thematic areas naturally emerge everywhere: on the bottom, we zoom in on a region indicating perceptively coherent groups.

at least 3 groups, assuming this as indication of shared common interests. In addition, we do the same by discarding the content-based features. Of the original 200 users, 183 have a link with each other, and 141 groups have been considered in total. Of course, normalization of distances has been done in order to have a fair comparison. The results are shown in Table 2, revealing three aspects: 1) our approach clusters users that share groups in a more compact way; 2) using the mean feature distance, the inclusion of content-based features do not compact the clusters, while 3) CG distance benefits from the addition of content-based information, compacting more the groups. An experiment considering only the content-based feature cannot be directly performed, as images with detectable objects are very few with respect to the size of the dataset.

Features retained	User maps distance	Feat distance
Non content-based	**0.2400**	0.3500
All features	**0.2125**	0.3552

Table 2. Mean intra-cluster distances.

5. CONCLUSIONS

Our work adopts the counting grid to convincingly combine content and aesthetics for capturing image preferences of users. A novel inference on CG allows to visualize the user tastes as elevation maps over the counting grid manifold: em-

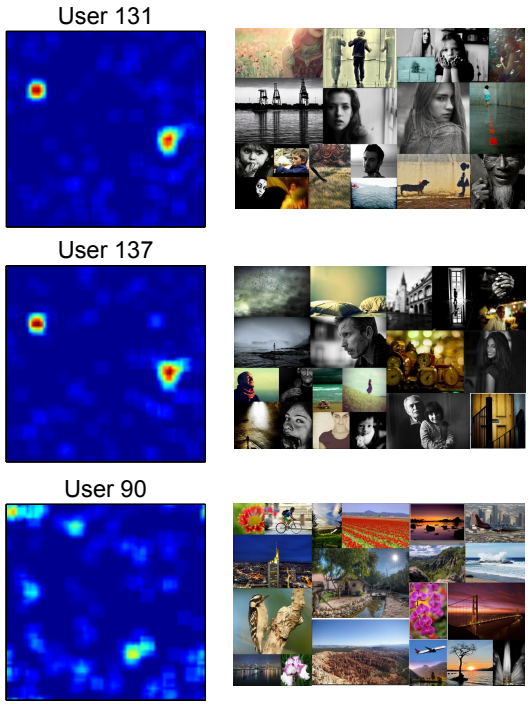

Fig. 3. User maps for 3 users from the dataset: red peaks correspond to clusters of many favorites. The first two users are very similar: indeed, their set of favorites (shown partially on the right) is visually similar. On the contrary, the third one shows a more sparse map, with his preferred images looking very different from the first two.

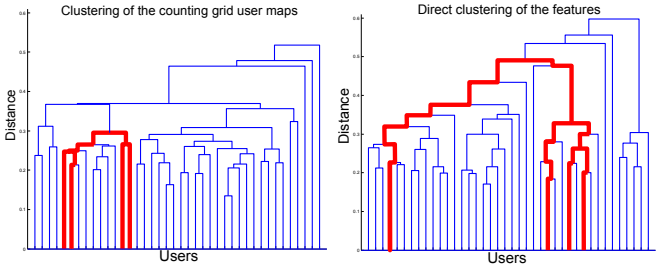

Fig. 4. Clustering of users: on the left, dendrogram based on distances between CG user maps; on the right, based directly on the average feature vectors of his favorites. The highlighted path shows a cluster of 4 users who share lots of groups related to urban street photography.

ploying these maps to evaluate distances among users seems to be fruitful for recommending groups to social networks like Flickr. We intend to develop this aspect in a future work through user tests. Above all, we presented a principled way to holistically represent user tastes that distills, but does not diminish, the wealth and diversity of information contained in image sets. Therefore, we envision further work built on top of this groundwork.

6. REFERENCES

[1] I. Biederman and E. Vessel, "Perceptual pleasure and the brain," *American Scientist*, vol. 94, no. 3, pp. 1–8, 2006.

[2] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," Washington, DC, USA, 2006, CVPR '06, pp. 419–426, IEEE Computer Society.

[3] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," Berlin, Heidelberg, 2008, ECCV '08, pp. 386–399, Springer-Verlag.

[4] Brett Adams, "Where does computational media aesthetics fit?," *IEEE Multimedia*, vol. 10, pp. 18–27, 2003.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision ECCV 2006*, vol. 3953 of *Lecture Notes in Computer Science*, pp. 288–301. Springer Berlin / Heidelberg, 2006.

[6] P. Lovato, A. Perina, N. Sebe, O. Zandonà, A. Montagnini, M. Bicego, and M. Cristani, "Tell me what you like and I'll tell you what you are: discriminating visual preferences on Flickr data," 2012, ACCV '12.

[7] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li, "Flickr distance," in *Proceedings of the 16th ACM international conference on Multimedia*, 2008, MM '08, pp. 31–40.

[8] A. Perina and N. Jojic, "Image analysis by counting on a grid.," in *CVPR*, 2011, pp. 1985–1992.

[9] N. Jojic and A. Perina, "Multidimensional counting grids: Inferring word order from disordered bags of words," in *UAI*, 2011, pp. 547–556.

[10] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the international conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 83–92, ACM.

[11] K.V. Mardia and P.E. Jupp, *Directional Statistics*, Wiley Series in Probability and Statistics. Wiley, 2009.

[12] P. Valdez and A. Mehrabian, "Effects of color on emotions.," *J Exp Psychol Gen*, vol. 123, no. 4, pp. 394–409, Dec. 1994.

[13] R. R. Hielscher and H. Schaeben, "A novel pole figure inversion method: specification of the *mtex* algorithm," *Journal of Applied Crystallography*, vol. 41, no. 6, pp. 1024–1037, 2008.

[14] F. Bachmann, R. Hielscher, and H. Schaeben, "Texture analysis with MTEX–free and open source software toolbox," *Solid State Phenomena*, vol. 160, pp. 63–68, 2010.

[15] H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 6, 1978.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[17] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603 – 619, 2002.

[18] C.M. Georgescu, "Synergism in low level vision," in *International Conference on Pattern Recognition*, 2002, pp. 150–155.

[19] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models, release 4," http://www.cs.brown.edu/~pff/latent-release4/, 2010.

[20] Paul A. Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.