Recognizing People by Their Personal Aesthetics: A Statistical Multi-level Approach

Cristina Segalin[1], Alessandro Perina[2], Marco Cristani[1]

[1]University of Verona, Italy
[2]Istituto Italiano di Tecnologia (IIT), Genova, Italy

Abstract. This paper presents a study on personal aesthetics, a recent soft biometrics application where the goal is to recognize people by considering the images they like. Here we propose a multi-level approach, where each level is intended as a low-dimensional space where the images preferred by a user can be projected, and similar images are mapped nearby, namely a Counting Grid. Multiple levels are generated by adopting Counting Grids at different resolutions, corresponding to analyze images at different grains. Each level is then associated to an exemplar Support Vector Machine, which separates the images of an individual from the rest of the users. Putting together multiple levels gives a battery of classifiers whose performances are very good: on a dataset of 200 users, and 40K images, using 5 preferred images as biometric template gives 97% of probability of guessing the correct user; as for the verification capability, the equal error rate is 0.11. The approach has also been tested with diverse comparative methods and different features, showing that color image properties are crucial to encode the personal aesthetics, and that high-level information (as the objects within the images) could be very effective, but current methods are not robust enough to catch it.

1 Introduction

Understanding the aesthetical preferences of a person, and specifically the images that he/she likes, is a noteworthy ability of the human beings; usually linked to high-level concepts such as the objects being portrayed in an image ("Jeff prefers photos of cars instead of landscapes"), it can be also based to apparent low-level visual properties such as having black/white colors, or full colors.

Having this capability transferred into a machine is without doubts of great benefit for many applications: from recommender systems that suggest images of interest to a particular user, to social aggregators which foster connections among individuals sharing similar aesthetical preferences. Among these applications, a novel one is emerging in these last years in the field of soft biometrics, aimed at identifying or verifying a person given a set of preferred images, dubbed *personal aesthetics* [1]. In general, soft biometric patterns differ from standard biometrics since they do not require a voluntary cooperation of the user in providing identification cues such as the face, the fingerprints etc.; even more notably, the identification or verification operation can be conducted without letting the user know about what is going on [2].

Soft biometrics can be partitioned into *physical/physyiological* (age, gender, ethnicity, height etc.) and *behavioral* biometrics, that is, encoding a characteristic linked to how a person does diverse mental/physical tasks [3]. This last class can be further exploded into *authorship-based* (linked to style peculiarities of the individual - how he/she writes a text), *motor skill-based* (how a person performs a particular physical task), *purely behavioral* (how a person solves a mental task) and *HCI-based* biometrics. HCI-based biometrics assumes that every person has a unique way to interact with an hardware device (a laptop, a smartphone or a simple touchscreen). For example, some approaches use the mouse or keystrokes dynamics to identify an individual [4, 5]; some other more recent methods focused on how Internet applications are utilized, like chatting [6] or browsing histories [7].

Personal aesthetics assumes that, given a set of preferred images of a user, it is possible to extract a set of features which are discriminative for him/her; these patterns can be used as biometric template, and employed for identification and verification. Personal aesthetics fits surely into the behavioral soft biometrics, while at a lower level of specification do not match with any of the previous categories. For these reasons, it could be good to have a "preference-based" category, whose approaches assume that a user may be identified by means of his/her preferences on multimedia data.

The motivations of why focusing on personal aesthetics, and in particular on images, are at least three; first of all, the huge presence of images in Internet: at the moment this article is being written, 55M of new images are daily uploaded on Instagram, with 1.2B of "*likes*" distributed over 16B of globally shared images (See *http://instagram.com/press/*); on Flickr, each of the 87M users has, on average, around 2K views per day (*http://statsr.net/flickr-stats/*); in the past year, 128B of images have been uploaded on Facebook (*http://goo.gl/0tWf.*), accessible to an audience of 1.26B users (*http://expandedramblings.com/.*). The second motivation is the enormous diffusion of the "liking" activities, since liking multimedia material is one of the most common social activities [8].

The third motivation is that, psychology and neuroscience have investigated the interrelation of individual characteristics on aesthetic preferences [9], finding that there are consistent ties between aesthetic appreciation and personality [10]; this last, being a stable characteristic of humans, ensures that personal aesthetics are somewhat *permanent*, a desirable property for soft biometric traits [2].

In this paper, a novel approach for personal aesthetics is proposed, which is based on the projection of the images into different latent spaces, each one of them representing a particular level with which to consider the preferred images. These spaces are 2D Counting Grids (CGs) [11], that is, smooth manifolds where visually similar pictures are mapped nearby. In the details, each CG is characterized by a particular resolution, that in rough words models how much visually similar should be the images in order to be close on the grid: the higher the resolution, the stronger the visual similarity of close images. The presence of multiple resolutions brings to evaluate differently grained similarity relations among images.

The approach assumes to have a set of users and some images preferred by them, which compose a gallery and a probe image set; it consists in a serial pipeline of initialization, enrollment and identification/verification stages.

In the initialization stage, multiple levels correspond to CGs of different resolutions, which are learned with the gallery images of all the users, without using ID labels. In the enrollment, the training data of a single user is projected on the CGs at different resolutions, resulting in different *embedding maps*. These maps are then fed into Support Vector Machines (SVMs), one for each CG. In particular the SVMs are trained as exemplar SVM, that is, using a single map as positive sample, and as negative samples all the maps of the other users at that CG resolution. In the identification/verification stage, probe images are projected into the CGs, forming another set of maps which are then classified by each of the SVMs, and producing a joint prediction; this last is used to provide or verify the identity of the user. It is worth noting that our method works with a varying number of images, both for the enrollment and the identification/verification stage, providing a versatile approach.

Through some explicative experiments, it is easy to capture the advantages of our methods. The use of the 2D CGs allows to see the kind of images liked by some users and disliked by the others; projecting on low-dimensional spaces permits to use any kind and number of counting features for encoding images (see more on this topic later on), contrarily to our previous approaches [1, 12, 13] which are based on an explicit cues weighting; having CGs at multiple resolutions avoids to deal with model selection issues (deciding a "correct" resolution for a CG is a problem [11]). The approach is also effective; in particular, the tests have been performed on the only real dataset currently available in the literature [1], composed by 40000 images which belong to 200 users chosen at random from the Flickr community. For each user, 200 preferred images (his "favorites") have been retained. As identification performance, using 5 preferred test images as biometric signature gives 97% of probability of guessing the correct user (state of the art was 83%); as for the verification capability, an equal error rate of 0.11 (best results was 0.25) is reached. Other than [12], we compared with several other baselines and alternative strategies, including a simple PCA baseline and multidimensional counting grids. Finally, we performed an extensive on the kind of features which can be used to describe an images: overall, the features are grouped into four families (see Table 1), i.e. *color, composition, content* and *textural properties*, according to the taxonomy proposed in [14]. Our experiments showed that using color and composition cues gives the best results, together with some interesting observations about high level features.

The rest of the paper is organized as follows: in Sec. 2 a summarization of the Counting Grid generative model is reported; in Sec. 3 the proposed approach is detailed, explaining how it can be customized for the identification and verification tasks. The approach is thoroughly tested in Sec. 4, and, finally, conclusions future perspectives are given in Sec. 5.

2 Mathematical Background: the Counting Grid Model

The Counting Grid (CG) is a generative model originally aimed at analyzing image collections [11]. It assumes that images are i.i.d. random variables represented as histograms (or bags-of-features) $\{c_z\}_{z=1,\dots,Z}$, where each c_z is a counting variable which enumerates the occurrences of the $z-$th feature.

In its 2D version, a CG π is a 2D finite discrete grid (a flattened torus with wrap-around at its extrema), spatially indexed by $\mathbf{i} = (x, y) \in [1 \dots E] \times [1 \dots E]$, and containing normalized feature counts $\{\pi_{\mathbf{i},z}\}$, indexed by $z = 1, \dots, Z$. Therefore, $\sum_z \pi_{\mathbf{i},z} = 1$ for every location $\mathbf{i}$ on the grid. The generative process underling the CG is as follows: an image (i.e. its BoF $\{c_z\}$) is generated by selecting a certain location $\mathbf{k}$ over the grid, calculating the distribution $h_{\mathbf{k},z} = \frac{1}{S^2} \sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z}$ by averaging all the words counts within the window $W_\mathbf{k}$ (of dimensions $S \times S$ and such that $\mathbf{k}$ is its upper left corner) and then drawing features counts from this distribution. In practice, a small window is located 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 $\{c_z\}$. In simpler terms, a CG could be think as a mixture model, where the components are overlapping windows indexed by $\mathbf{k}$.

This said, it appears clear that the position of the window $\mathbf{k}$ in the grid is a latent variable; given $\mathbf{k}$, the likelihood of $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{c_z} = \frac{1}{S^2} \prod_z \Big(\sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z} \Big)^{c_z}. \tag{1}$$

Given that the ratio between the grid size $E \times E$ of a Counting Grid and the window size $W \times W$, is smaller than number of images, this 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. In practice, local neighborhoods in the grid represent similar concepts and images mapped in close locations are somehow similar.

To learn a Counting Grid, the likelihood over all training images T needs to be maximized, and this can be written as

$$p(\{\{c_z^t\}, \mathbf{k}^t\}_{t=1}^T) \propto \prod_{t=1}^{T} \prod_{z=1}^{Z} \Big(\sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z}^t \Big)^{c_z^t}. \tag{2}$$

The sum over $\mathbf{k}$ makes it difficult to perform assignment to the latent variables (i.e., the components of the mixture) and so to estimate the model parameters and it is necessary to employ an EM algorithm. The procedure is a bit complicated and involves different variational distributions; for this study it is only necessary to quote the posterior distribution, calculated in the E step,

$$p(\mathbf{k}^t|\{c_z^t\}) = q_\mathbf{k}^t \propto \exp \sum_z c_z^t \cdot \log h_{\mathbf{k},z} \tag{3}$$

which is a probabilistic mapping of the t-th bag to the grid windows $\mathbf{k}$. This mapping is usually peaky, i.e. each image tends to map to a few nearby locations

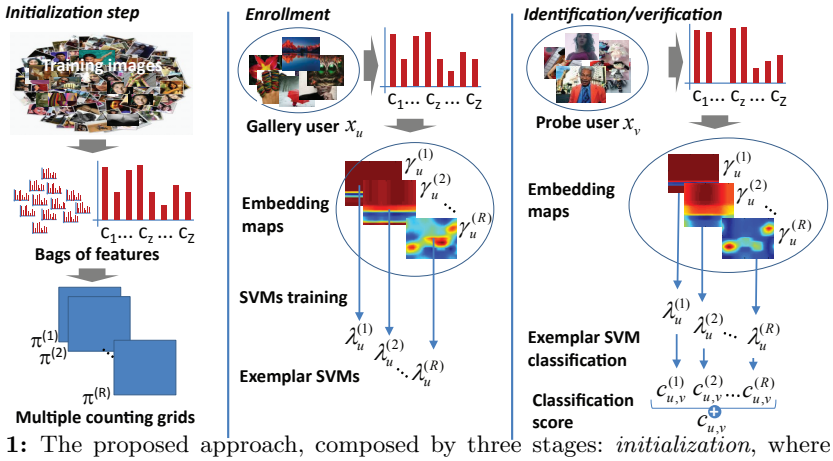

Fig. 1: The proposed approach, composed by three stages: *initialization*, where the multi-resolution Counting Grid is learnt; *enrollment*, where the classifiers for each user are trained, and *identification/verification* stages, where unknown personal aesthetics are matched with the gallery.

in the grid. For details on the learning algorithm and on its efficiency, the reader can refer to the original paper [11].

3 The Proposed Approach

The proposed three-step approach is sketched in Fig. 1. The initialization step is applied on the training image set: it consists on creating a bag of features for each image, and learning a set of Counting Grids, each at a different window size (i.e., the resolution of the CG). In the enrollment stage, the preferred images of each user x_u, $u = 1, \ldots, U$ of the gallery set are mapped on the CG latent spaces, and the resulting maps (one for each CG space) are fed into a discriminative classifier. In the identification/verification stage, the test images of a probe subject are transformed into bags of features, and projected into the CGs; in particular, in the identification scenario, the resulting maps are given as input to all the U gallery classifiers, producing U identification scores. These scores are used to decide the best gallery user. In the case of the verification task, the maps are given to a single gallery classifier (the one which is supposed to match the identity of the probe), which accepts or rejects the signature considering a given threshold.

3.1 Initialization Stage: Creating the Bags of Features

For the sake of comparison, we adopted the dataset used in [12], composed by 40000 images belonging to 200 users, chosen randomly from the Flickr social network. For each user, the 200 last "favored" pictures have been retained, that is, pictures of other photographers that have meet his/her preferences. Repeated favored images across users are less than the 0.05%.

Category	Name	L	Short Description
Color	Use of light	1	Average pixel intensity of V channel [16]
	HSV statistics	3	Mean of S channel and standard deviation of S, V channels [14]
	Emotion-based	3	Amount of *Pleasure*, *Arousal*, *Dominance* [14, 17]
	Circular Variance	1	*Circular variance* of the H channel in the IHLS color space [18]
	Colorfulness	1	Colorfulness measure based on Earth Mover's Distance (EMD) [16, 14]
	Color Name	11	Amount of *Black*, *Blue*, *Brown*, *Green*, *Gray*, *Orange*, *Pink*, *Purple*, *Red*, *White*, *Yellow* [14]
Composition	Edges	1	Total number of edge points, extracted with Canny [1]
	Level of detail	1	Number of regions (after mean shift segmentation) [19, 20]
	Regions	1	Average *size* of the regions (after mean shift segmentation) [19, 20]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [16, 14]
	Rule of thirds	2	Mean of S,V channels in the inner rectangle of the image [16, 14]
	Image parameters	1	Size of the image [16, 1]
Texture	Entropy	1	Image entropy [1]
	Wavelet textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on the HSV channels [16]
	Tamura	3	Amount of *Coarseness*, *Contrast*, *Directionality* [21]
	GLCM-features	12	Amount of *Contrast*, *Correlation*, *Energy*, *Homogeneity* for each HSV channel [14]
Content	Objects	28	Objects detectors [15]: in particular, here are the objects for which detectors are available: *people*, *plane*, *bike*, *bird*, *boat*, *bottle*, *bus*, *car*, *cat*, *dog*, *table*, *horse*, *motorbike*, *chair*. In all the cases we kept the number of instances and their average bounding box *size*
	Faces	2	Number and *size* of faces after Viola-Jones face detection algorithm [22]

Table 1: Summary of all features. The column 'L' indicates the feature vector length for each type of feature.

As for the features, we consider those of [12] for comparability, here reorganized as in the computational aesthetics taxonomy of [14] (see Table. 1); 4 categories are present: *color* (distribution, diversity, purity, emotional content, etc.), *composition* (size and number of homogeneous regions, amount of edges, depth of field, rule of thirds, etc.), *textures* (spatial distribution of visual properties) and *content*, which individuate semantic entities (cars, chairs and the like); in this last case, robust probabilistic object detectors have been employed [15] (for a complete list of all the detectable objects, see Table 1); other than the number of instances of objects in an image, the average area of their bounding boxes is considered.

It is worth noting that each feature extracted in the proposed approach indicates the level of presence of a particular cue, i.e. an enumeration value or an intensity count. This is needed for the modeling with the Counting Grid.

3.2 Initialization Stage: Multi-view Counting Grid Training

Given the bags of features of the training images, the extent E of the Counting Grid and its window size S, a multi-resolution CG is learned. This amounts to learn $R = E - S$ Counting Grids, starting at resolution $r = 1$ (the lowest resolution level) with the window of size $E - 1$, decreasing the window size of one pixel at each time, until the minimum size S (the highest resolution level $r = R$) is reached. At each resolution level r (except the first one), we used the CG learnt at the previous step, i.e., $\pi^{(r-1)}$ as initialization for $\pi^{(r)}$. At the first resolution level, the initialization is random.

Using different windows sizes corresponds to vary the topology of the CG latent spaces: a large window size leads to an embedding map where loosely similar images are near-uniformly distributed over a large area and only the very different images are strongly separated. Conversely, a small window size will create a peaked map, where only highly similar images are projected nearby, and weakly similar pictures are separated. Initializing a model training using the CG of the previous level allows to mitigate local minima problems (as in the case of a too sparse CG, with many images mapped very close) ensuring to use all the CG extent for the mapping. In addition, this initialization strategy permits to show how the mapping evolves at the different resolutions, refining spread and unfocused projections into defined and intuitive thematic regions.

Obviously, the Counting Grids can not be directly visualized (each location contains a distribution of features), but it is possible to create an image mosaic using those images $\{c_z^t\}$ which give the highest posterior at each location $\mathbf{k}$, i.e., $p(\mathbf{k}^t|\{c_z^t\})$, at a given resolution level r. Adopting this visualization strategy, Fig. 2 (left) shows CGs with $E = 45$ at resolutions r=5 (S=40) and 35 (S=10): while going from coarse (top) to finer (bottom) resolutions, the semantics of the CG emerge as peaked regions, where each region carries out a different type of images. In this case a set of images where the orange is predominant are highlighted. As visible, at the coarser resolution the orange images lie in two regions, where other tonalities are also present. Going to the highest resolution has the effect of packing nearby those images into a compact area. On Fig. 2 (center) the CG at the highest resolution is reported.

Such a representation is shown in Fig. 2 (center) for a Counting Grid with $E = 45$ at resolution $r = R$ ($S = 10$, maximum resolution). As visible, close images are visually similar, and semantic topics do emerge.

3.3 Enrollment Stage

Once the different Counting Grids are learnt, the images of each gallery user can be projected within it, obtaining R maps per user, one map for each resolution. The projection corresponds to a generative embedding, calculating a posterior probability at each location $\mathbf{k}$; once we have fixed a user u and a resolution r the posterior is

$$\gamma_u^{(r)} = \sum_{t \in T_u} p(\mathbf{k}^t|\{c_z^t\}, \pi^{(r)}) \tag{4}$$

where T_u identifies the set of images of the user u: T_u can be different, depending on how many gallery images are available for user u. Roughly speaking, the main idea is to sum all the mappings of the images belonging to a given user, thus highlighting the zones of the latent space where the images have been located. The presence of Counting Grids at multiple resolutions allows to map the preferences of the user from a very rough resolution (on the Counting Grids obtained with large windows) until the finest resolution (the Counting Grid being learned with a small sized window), where the map is usually peaked.

A graphical explanation of the mapping process is shown in Fig. 2 and Fig. 3; in Fig. 2, together with the collage of the CG, on the right are reported the embedding maps of a single resolution level (the maximum, i.e., r=R) for three

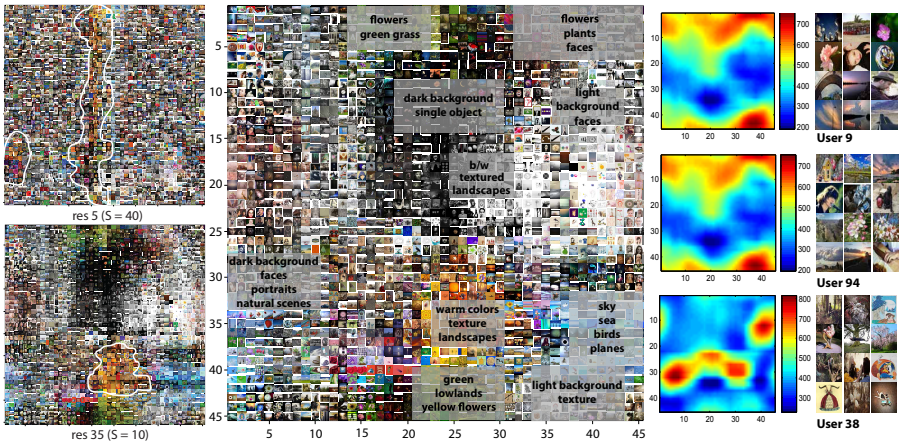

Fig. 2: Visualization of Counting Grids: on the left, CGs with $E = 45$ at resolutions r=5 (S=40, top) and 35 (S=10, bottom). On the center, the $S = 10$ grid is visualized as a collage of images (see the text for the details on how the collage is created). On the right, the embedding maps of a single resolution level (r=R) are reported for three subjects, together with some random images preferred by them (better viewed in colors).

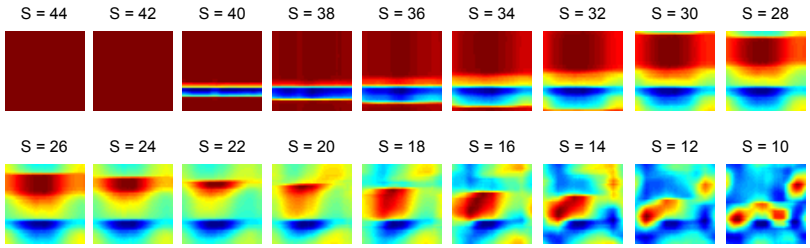

Fig. 3: Embedding maps for user 38 of Fig. 2. Starting from the lowest resolution (r=1, S=44) and going towards higher resolutions, the maps show refined blobs and areas, identifying more precisely semantic areas, easily interpretable, on the grid.

subjects, together with some random images preferred by them. One can notice two facts: 1) given a user, looking at his map and at the CG collage as reference, does allow to easily understand which kind of images are his preferred; 2) comparing the maps of different users, one can understand possible similarities: first two users from the top appear to share much the same preferences, while the third one has radically diverse preferences. This fact is confirmed by checking the random pictures of the users, on the right.

In Fig. 3 are reported the R mappings for the user 38 of Fig. 2. Starting from very blurred and unstructured maps corresponding to the lower resolutions, going toward higher resolution maps, blobs and distinct areas start to emerge, refining the "semantic" knowledge of the preferences a user exhibits.

After the mapping step, the maps $\{\gamma_u^{(r)}\}_{r=1,\dots,R}$ can be used as ID template for user u; to this sake, a battery of exemplar SVMs $\{\lambda_u^{(r)}\}_{r=1,\dots,R}$ are learnt

(one for each resolution), using as positive samples the maps $\gamma_u^{(r)}$ at the different resolutions r (one map for each SVM) and as negative samples the maps of the other users. In this study, Support Vector Machines with radial basis functions have been employed. This step concludes the enrollment stage.

3.4 Identification and Verification Stage

In the identification/verification stage, all the probe images of a user v are first encoded as bags of features. Subsequently, they are mapped on the multi-resolution CG, and the resulting maps $\{\gamma_v^{(r)}\}_{r=1,\ldots,R}$ are used as input of the SVMs related to the gallery user u; they classify the maps producing R scores $\{c_{u,v}^{(r)}\}_{r=1,\ldots,R}$ that, once mediated, provide a single classification score $c_{u,v}$. In other words, each user produces R probe maps; each of them is given as test input to the correspondent SVM of the gallery user, providing a confidence score (the distance from the separating hyperplane). Averaging these scores over all the resolutions gives the final confidence score. In the identification case, a confidence score is associated to each gallery user; this allows to rank the scores, keeping the highest ranked user as the best match with the probe. In the verification of the probe user, assumed to be the $v-$th, the confidence score given by the $v-$th classifier is simply evaluated, accepting o rejecting the signature depending on a threshold opportunely decided.

4 Experimental Evaluation

Several experiments are carried out to understand the potentialities of our approach. First of all, we investigate the ability of the features in capturing what is liked by an user, ensuring the highest identification and verification performance. Then, we compare our approach against a set of competitors, including our previous work [12]: to this sake, the same experiments carried out in [12] have been taken into account. Finally, we analyze how beneficial is to exploit CGs at different resolutions, capturing also how informative is each single resolution.

Identification and verification applications are considered: in the identification task the goal is to select the identity of an individual among a set of gallery users, given a pool of images liked by him/her; the verification task amounts to verify the identity of a particular user by means of his/her preferred images, considering his/her gallery images. In both the cases, the parametrization of the Counting Grids is the same: the size is fixed at $E = 45$ pixels for all of them, while the (smallest) window size is set to $S = 10$; this generates a set of 35 maps per user. The extraction of the image features takes 60 minutes per user (100 images), on a not optimized MATLAB code run on a 3.4 GHz processor with 16 Giga of RAM. The learning of the Counting Grid at a single resolution takes in total 2 minutes, while the mapping + SVM training operation requires 3 seconds for $N = 100$ images of the same user, on the same computer. Regarding the variability of the results in relation to the E and S values, the proposed approach maintains similar performance when the ratio between E and S (also dubbed

category	rank 1	rank 5	rank 20	rank 50	nAUC
color	0.38±0.21	0.65±0.01	0.86±0.01	0.97±0.01	0.96± <0.01
composition	0.11±0.01	0.25±0.02	0.45±0.02	0.69±0.12	0.81±0.01
texture	0.10±0.01	0.21±0.01	0.39±0.03	0.64±0.02	0.79±0.01
content	0.10±0.01	0.20±0.01	0.38±0.03	0.61±0.03	0.78±0.01
all	0.36±0.02	0.64±0.02	0.86±0.01	0.97± <0.01	0.96± <0.01
color + composition + textures	0.37±0.02	0.64±0.02	0.86±0.01	0.98±0.01	0.96± <0.01
color + composition	**0.42±0.02**	**0.71±0.02**	**0.91±0.01**	**0.99±0.01**	**0.97± <0.01**

Table 2: CMC scores for the identification task, 100 images fo reach gallery user and 5 images for the probe user

"capacity" in [11]) is bounded in the interval [3,5]. Even if E and S respect the capacity ratio, performances seem to decrease when $E < 10$ and $E > 70$.

4.1 Feature Analysis

Following the Table 1, we divide the features in four categories: *color*, *composition*, *texture* and *content*. For each category, we instantiate a identification task: given a probe signature built from an image or a set of images, the goal is to guess the gallery user who tagged them; to do that, fixing a gallery user, the average of the confidence scores produced by the exemplar SVMs (one score for each resolution) is calculated. Hopefully, the gallery user with highest averaged score corresponds to the probe user. As identification figure of merits, we use the Cumulative Matching Characteristic (CMC) curve [23]; given a probe signature of a user and the matching confidence score, the curves tells the rate at which the correct user is found within the first k matches, with all possible k spanned on the x-axis (they are also called *ranks*). In all the following experiments CMC plots are obtained averaging the CMC curves of 5 different experiments with different gallery/probe splits. In this experiment, we use 100 images as forming the gallery signatures, and 5 images for the probe signatures.

In Table 2 are reported the CMC values at different ranks, together with the normalized Area Under the Curve (nAUC). As visible, the color category is the most significative, followed by composition, texture and content. The poor performance of the content features, that is, object detectors, is due to the fact that object detectors produce many errors, both in precision and recall: this is due to the nature of the Flickr photos, which are artistic and not reminiscent those of the object recognition benchmark (PASCAL, CALTECH and the like). We evaluate all the possible combinations of group of features (some of them are reported in the table), with the best one formed by color and composition, which will be sued in the following. Interesting, the textures seem to slightly degrade the performances.

4.2 Identification Results

The results of the identification task are carried out following the protocol of [12]. We cross-validate the parameters of the SVM classifier with Gaussian kernel obtaining the best configuration with $C = 1000$ and $g = 0.001$. As competitors,

we report the performance of [12] (with the acronym *LASSO*) and [13] (*PaD*). In addition, we set up some baselines, which may help in motivating some technical choices we have made with our framework. The *Ensemble* method is the same as our proposal, with the only difference that the CGs are learned independently, without sharing their parameters; the *PCA* approach, which actually uses Principal Component Analysis to create a low dimensional space projection space where all the images can be projected. Once the projection of a probe signature is performed, the resulting map containing the projected images (opportunely quantized in order to be of the same dimension irrespective of the nature and cardinalities of the signatures) is fed into the exemplar SVMs. In Fig. 4 (left) the various CMC curves are reported by fixing the number of gallery images to 100, and the number of probe images to 5. As visible, our approach overcomes all the competitors.

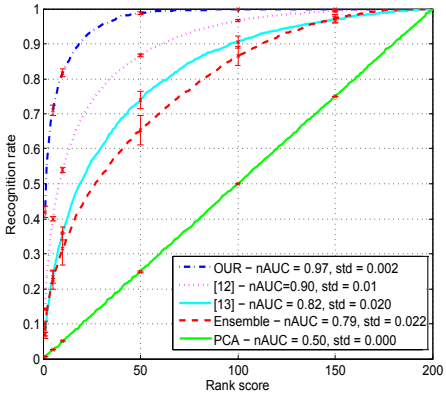

Fig. 4: Comparative results for the identification task, with 5 images for the probe signatures and 100 images for the gallery signatures.

T_{te}	*rank 1*	*rank 5*	*rank 20*	*rank 50*	*nAUC*
1	0.19	0.42	0.66	0.86	0.90
5	0.42	0.71	0.71	0.99	0.97
20	0.63	0.87	0.7	1.00	0.99
100	0.73	0.92	0.98	1.00	0.99
T_{tr}	*rank 1*	*rank 5*	*rank 20*	*rank 50*	*nAUC*
5	0.29	0.59	0.83	0.94	0.95
10	0.46	0.80	0.95	0.99	0.98
20	0.63	0.89	0.98	1.00	0.99
50	0.71	0.93	0.99	1.00	0.99
100	0.71	0.92	0.98	1.00	0.99

Table 3: Identification results, varying the number T_{te}/T_{tr} of images of gallery/probe signatures (and fixing the other cardinality to 100 for each user). All the results are with a variance of less than the 1%.

In Table. 3 we report (in the upper part) the performance of our approach while varying the number of test images used to compose the probe signature of a user, while keeping the number of images used to build the gallery signature fixed to 100; in the lower part we report the analogue figure while varying the cardinality of the gallery signatures and keeping fixed to 100 the cardinality of the probe signature. Intuitively, augmenting the cardinality of the gallery/probe signature does ameliorate the identification performance.

To test the importance of having different CG resolutions, we perform a set of identification trials while using 100 images of gallery and 100 of probe, with 1, 2, 5, 10, 20 and 35 different resolutions (35 is the total number of resolutions employed). In the case of a single resolution, all the S windows size between 10 and $E-1=44$ have been independently evaluated, averaging their recognition performance. For evaluating higher numbers of resolutions, different windows size have been sampled without replacement (depending on the cardinality being evaluated) and ranked in descending order. After that, the window with the

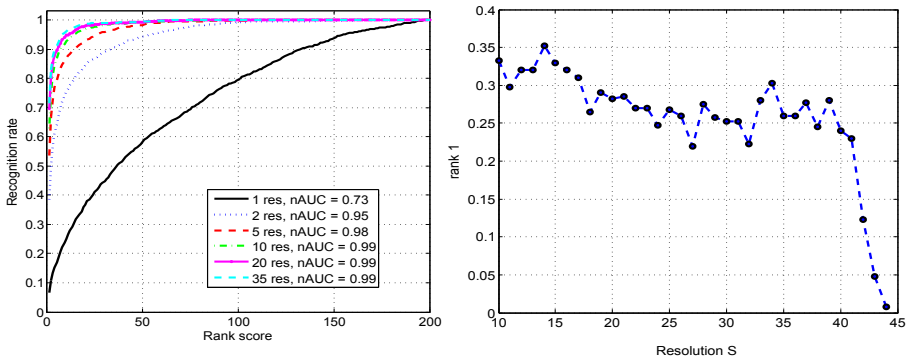

Fig. 5: Identification scores while varying the number of resolution employed, and analysis at rank 1.

largest size has been learned with random initialization; the obtained CG has been used as prior for the second ranked one and so on. Results (averaged over 35 gallery/probe splits) are portrayed in Fig. 5.

As expected, increasing the number of resolution levels does augment the identification capabilities. To better understand the role of each resolution, each one of them has been evaluated independently (under the same experimental protocol, $T_{tr} = T_{te} = 100$, 35 repetitions), reporting in Fig. 5 the rank-1 identification score (standard deviation $< 1\%$ in all the cases). It emerges that performance is better while going toward higher resolutions, even if no one of them can reach the same score one can get when using the joint framework (that is, 0.71, see Table. 3). This means that every resolution level carries out a different complementary analysis of the images.

4.3 Verification Results

In the verification scenario, the capability of the system to verify if a signature matches a given identity is evaluated. For this purpose, a ROC curve is computed for every user u, where client images are taken from the probe set of the user u and impostor images are taken from all the other probe sets. Depending on the number of images taken into account, different kind of client/impostor maps may be built. Given an "authentication threshold", i.e., a value over which the subject is authenticated, sensitivity (true positive rate) and specificity (true negative rate) can be computed. By varying this threshold the ROC curve is finally obtained. In Fig. 6 the authentication ROC curves are portrayed; other than AUC, the equal error rate (EER) is also reported, which models the error when sensitivity and 1-specificity have an equal value.

Even in this case, augmenting the number of test images per signatures increments the performance; as for varying the number of images used for the gallery signatures, and the number of resolutions for producing the multi-scale CG, analogue results than those obtained for the recognition task can be observed, so the results have been omitted.

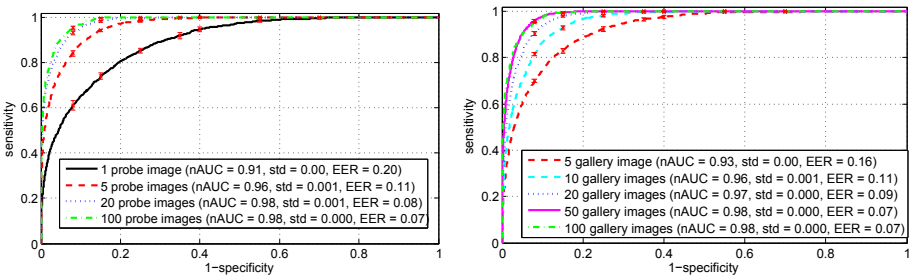

Fig. 6: Verification scores: the ROC curves (together with AUC and EER score) are reported while varying the number of probe (left) and gallery (right) images employed.

5 Conclusions

Personal aesthetics is a recent soft biometrics trait that emerged thanks to the large diffusion of images in Internet and to the possibility of liking them. The idea of capturing the identity of people using their aesthetical preferences underlies the capability of understanding their personal tastes. This approach does both the things in a satisfying fashion: Counting Grids allow to project images in a latent space where similar pictures are mapped nearby, so that semantic areas can emerge and being observed and interpreted. In this respect, future work should focus on the kind of features to use, and in particular on how medium-high level features can be crafted, since object detection have shown to be unreliable; we think that deep learning could be well suited for this aim. On the other side, our method shows that CGs induce latent representations (the embedding maps) very informative for discriminating users by means of kernel machines, especially when multiple images preferred by a single individual are available. In this regard, future work should be spent in testing a real application where personal aesthetics are exploited, encouraging their use in genuine soft/biometric scenarios.

References

1. Lovato, P., Perina, A., Sebe, N., Zandonà, O., Montagnini, A., Bicego, M., Cristani, M.: Tell me what you like and I'll tell you what you are: discriminating visual preferences on Flickr data. In: Computer Vision–ACCV 2012. Springer (2013) 45–56
2. Dantcheva, A., Velardo, C., D'angelo, A., Dugelay, J.L.: Bag of soft biometrics for person identification. Multimedia Tools and Applications **51** (2011) 739–777
3. Yampolskiy, R.V., Govindaraju, V.: Behavioural biometrics: a survey and classification. International Journal of Biometrics **1** (2008) 81–113
4. Pusara, M., Brodley, C.E.: User re-authentication via mouse movements. In: Proceedings of ACM workshop on Visualization and data mining for computer security. (2004) 1–8
5. Rybnik, M., Tabedzki, M., Saeed, K.: A keystroke dynamics based system for user identification. In: Computer Information Systems and Industrial Management Applications, 2008. CISIM'08. 7th, IEEE (2008) 225–230

6. Roffo, G., Segalin, C., Vinciarelli, A., Murino, V., Cristani, M.: Reading between the turns: Statistical modeling for identity recognition and verification in chats. In: 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS (2013) 99–104
7. Olejnik, L., Castelluccia, C., Janc, A., et al.: Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. In: 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs). (2012)
8. Jin, X., Wang, C., Luo, J., Yu, X., Han, J.: Likeminer: a system for mining the power of 'like' in social media networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. (2011) 753–756
9. Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.T., Wang, J., Li, J., Luo, J.: Aesthetics and emotions in images. Signal Processing Magazine, IEEE **28** (2011) 94–115
10. Furnham, A., Walker, J.: The influence of personality traits, previous experience of art, and demographic variables on artistic preference. Personality and Individual Differences **31** (2001) 997–1017
11. Perina, A., Jojic, N.: Image analysis by counting on a grid. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 1985–1992
12. Lovato, P., Bicego, M., Segalin, C., Perina, A., Sebe, N., Cristani, M.: Faved! biometrics: Tell me which image you like and i'll tell you who you are. IEEE Trans. on Information Forensics and Security **9** (2014) 364–374
13. Segalin, C., Perina, A., Cristani, M.: Biometrics on visual preferences: A "Pump and Distill" regression approach. In: IEEE International Conference on Image Processing (ICIP). (2014)
14. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: International Conference on Multimedia, ACM (2010) 83–92
15. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. on Pattern Analysis and Machine Intelligence **32** (2010) 1627–1645
16. Datta, R., Joshi, D., Li, J., Wang, J.: Studying aesthetics in photographic images using a computational approach. In: ECCV. Volume 3953. Springer Berlin / Heidelberg (2006) 288–301
17. Valdez, P., Mehrabian, A.: Effects of color on emotions. Journal of Experimental Psychology: General **123** (1994) 394
18. Mardia, K., Jupp, P.: Directional Statistics. Wiley (2009)
19. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE TPAMI **24** (2002) 603 – 619
20. Georgescu, C.: Synergism in low level vision. In: International Conference on Pattern Recognition. (2002) 150–155
21. Tamura, H., Mori, S., Yamawaki, T.: Texture features corresponding to visual perception. IEEE Trans. on Systems, Man and Cybernetics **8** (1978) 460–473
22. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on CVPR. (2001) 511–518
23. Moon, H., Phillips, P.J.: Computational and performance aspects of PCA-based face-recognition algorithms. Perception-London **30** (2001) 303–322