Personal Aesthetics for Soft Biometrics: a Generative Multi-resolution Approach

Cristina Segalin
Dept. of Computer Science
University of Verona, Italy
cristina.segalin@univr.it

Alessandro Perina
Dept. of Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia(IIT), Italy
alessandro.perina@iit.it

Marco Cristani
Dept. of Computer Science
University of Verona, Italy
marco.cristani@univr.it

ABSTRACT

Are we recognizable by our image preferences? This paper answers affirmatively the question, presenting a soft biometric approach where the preferred images of an individual are used as his personal signature in identification tasks. The approach builds a multi-resolution latent space, formed by multiple Counting Grids, where similar images are mapped nearby. On this space, a set of preferred images of a user produces an ensemble of intensity maps, highlighting in an intuitive way his personal aesthetic preferences. These maps are then used for learning a battery of discriminative classifiers (one for each resolution), which characterizes the user and serves to perform identification. Results are promising: on a dataset of 200 users, and 40K images, using 20 preferred images as biometric template gives 66% of probability of guessing the correct user. This makes the "personal aesthetics" a very hot topic for soft biometrics, while its usage in standard biometric applications seems to be far from being effective, as we show in a simple user study.

Categories and Subject Descriptors

K.6.5 [**Computing Milieux**]: Security and Protection—*authentication, unauthorized access*; I.5.4 [**Computing methodologies**]: Pattern Recognition—*pattern analysis*

Keywords

Soft biometrics, computational aesthetics, Counting Grid

1. INTRODUCTION

Soft biometric traits are human characteristics that provide some information about the identity of an individual, that differ from standard biometric patterns since they are not intrusive and do not require explicit cooperation for their extraction - they can be fully imperceptible [2].

Soft biometrics can be divided into *physical/physyiological* (age, gender, ethnicity, height, EEG signals etc.) and *behavioral* biometrics, that is, encoding a characteristic linked to the behavior of a person [9]. This last class can be further partitioned into *authorship-based* (linked to style peculiarities of the individual - how she/he writes a book), *motor skill-based* (how a person performs a particular physical task), *purely behavioral* (how a person solves a mentally demanding task) and *HCI-based biometrics* [24].

HCI-based biometrics are based on the idea that every person has a unique way to interact with a personal computer. For example, some methods investigated the possibility of identifying a person considering mouse or keystrokes dynamics [17, 19]; some other approaches focused on how people use Internet, like chatting [18] or browsing histories [14].

Very recently, a brand-new HCI-based biometric trait emerged, exploiting the "personal aesthetics" of people, that is, those image preferences that distinguish people from each other [9]. The approach assumes that, given a set of preferred images, it is possible to extract a set of features individuating discriminative visual patterns; these patterns can be used as biometric template, and employed for identification.

The motivations of why focusing on pictures to encode the identity of an individual are many: from one side, taking pictures is the action most commonly performed with mobile phones (82% of the users from USA), followed by exchanging text messages (80% of the users) and accessing the Internet (56% of the users) [4]. Furthermore, 56% of the American Internet users either post online original pictures and videos (46% of the total Internet users) or share and redistribute similar material posted by others (41% of the total Internet users). In this scenario, the use of the *liking* mechanisms, that is, online actions allowing users to publicly express preference for a given picture, has become pervasive and massive, becoming a social mass phenomenon [20]. On the other side, psychology and neuroscience have investigated the role of personal characteristics on aesthetic preferences [8], finding that there are remarkable ties between aesthetic appreciation and personality [6]. This latter, being a stable characteristic of humans, ensures that personal aesthetics are somewhat *permanent*, a desirable property for soft biometric traits [2].

In this paper, this novel promising direction is followed, proposing a generative embedding approach for managing the personal aesthetics soft biometrics. The general assumptions of the approach are that, for a given set of users, we have a pool of images preferred by each one of them; we

also suppose that these images have been chosen from a potentially infinite set of images and that the amount of liked images shared by more than a user is minimal. These assumptions are crucial for the effectiveness of the approach, but reasonable, as we will see in the following.

The approach consists in an initialization stage, followed by the enrollment stage and finally the identification stage. In the initialization stage, a low-dimensional multi-resolution latent space is learned, consisting of a set of Counting Grids (CGs) [16]: each CG is a 2D space (a flattened torus), where visually similar pictures are mapped nearby; 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. Having multiple resolutions means to evaluate differently grained similarity relations among images.

In the enrollment stage, a collection of preferred pictures of an individual in the gallery set is mapped into the multi-resolution CG, resulting in an ensemble of *embedding* maps, one for each resolution. These embedding maps are then fed into discriminative classifiers: in particular, for each resolution grain (that is, for each embedding map) a classifier is instantiated, learned in one-VS-all modality.

In the identification stage, test images (one, or more than one) are projected into the CGs, forming another set of embedding maps which are then classified, producing a joint prediction; this last is used to provide the identity of the user. Please note that the approach allows to use a varying number of images both for the enrollment and the identification stage, providing a flexible soft-biometrics mechanism.

The proposed approach is very expressive: it allows to understand "visually" the kind of images liked by a user, and how such images distinguish him from other persons, in a very compact and economic way, using directly the CGs. This overcomes one of the limitations of [9], where the images are treated directly in the original feature space, and analyzed through LASSO regression as classifier; in the case of high number of features, this would lead to overfitting issues. As second, methodological contribution, an intrinsic limitation of Counting Grids has been circumvented, that is, their model selection (how to select a particular resolution [16]). This problem has been faced by considering various CG resolutions, capturing the diverse mappings they generate; this results in a multi-resolution image analysis, which has proven to be better than focusing on a single resolution. More importantly, the CG modeling allows to get impressive identification performances, definitely beating the state of the art. Comparative tests have been performed on the only real dataset currently available in the literature [9], 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 20 preferred test images as biometric signature gives 66% of probability of guessing the correct user (state of the art was 25%), promoting personal aesthetics as a promising soft biometric modality, with performances quite close to what one can expect from a classic biometric signature. Anyway, the assumption of having an infinite number of images that the user can select from (that is, the entire Flickr repository) is an essential hypothesis for making our approach effective; in order to demonstrate this fact, we set up a user study which analyzes a random subset of users; each user is asked to select a number of preferred images from a finite pool of available test images, as it could happen in a standard biometric approach where the user has to select a signature from a finite number of alternatives (a password). In this case, the available aesthetical variability diminishes dramatically, and as a consequence, identification performances drop.

Summarizing, the contributions of this work are

- a novel approach for personal aesthetics, which is a very recent soft biometric trait;
- a novel methodology for Counting Grids modeling, solving the problem of the model selection;
- new state of the art and impressive results, doubling in some cases the state of the art performances;
- a critical study of the limitations of the personal aesthetics.

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 tasks. The approach is thoroughly tested in Sec. 4, and, finally, conclusions are given and future perspectives are envisaged in Sec. 5.

2. MATHEMATICAL BACKGROUND: THE COUNTING GRID MODEL

The Counting Grid (CG) is a recent generative model [16] aimed at analyzing image collections. 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 two-dimensional version[1], a CG π is a 2D finite discrete grid (a flattened torus), spatially indexed by $\mathbf{i} = (x, y) \in [1 \dots E] \times [1 \dots E]$, and containing normalized counts of features $\{\pi_{\mathbf{i},z}\}$, indexed by $z = 1, \dots, Z$. Therefore, $\sum_z \pi_{\mathbf{i},z} = 1$ for every $\mathbf{i}$ on the grid. Under this model, an image (i.e. its BoF $\{c_z\}$) is generated by selecting a certain location $\mathbf{k}$, 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 other words, unlike a straightforward embedding (e.g. PCA) that links an image with a point location, the CG forces the image to link with a small window of locations. Simply speaking, 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

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

[1]N-dimensional in general, here we focus on 2 dimensions.

Given that the size $E \times E$ 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. 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

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

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; this is the same that happen with mixtures of Gaussians, hidden Markov models etc.; therefore, 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,

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

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 in the grid. For details on the learning algorithm and on its efficiency, the reader can refer to the original paper [16].

3. THE PROPOSED APPROACH

The proposed three-stage approach is sketched in Fig. 1. The initialization step is applied on a training set made by generic images: it consists on creating a bag of features for each image, and learning the multiscale Counting Grid. 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 space, and the resulting maps (one for each CG scale) are fed into a discriminative classifier. In the identification stage, the test images of a probe subject are transformed into bags of features, and embedded into the CGs; 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.

3.1 Initialization Stage: Creating the Bags of Features

For the sake of comparison, in this work the dataset used in [9] has been considered; it is composed by 40000 images belonging to 200 users, chosen at random from the Flickr website. For each user, the 200 last "favored" pictures have been retained (the act of favoring an image consists in clicking on a specific icon close to the liked image in the main Flickr interface). Repeated images across users are less than the 0.05%.

From each image $\mathbf{x}_t$, the same set of cues of [9] has been extracted, composed by 19 types of features resulting in a 111-dimensional real vector (see Table 1); the goal is to manage highly heterogeneous image features, letting them smoothly interact in the CG.

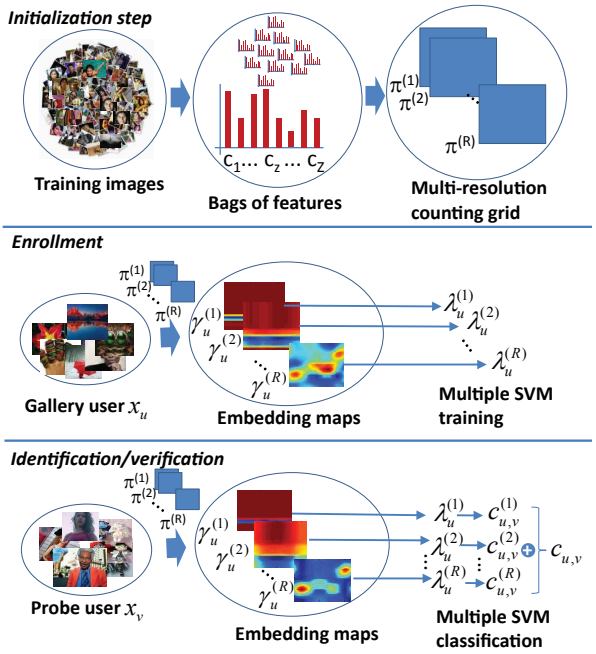

Figure 1: The proposed approach, composed by three stages: *initialization*, where the multi-resolution Counting Grid is learnt; *enrollment*, where the classifiers for each user are trained, and *identification* stages, where unknown personal aesthetics are matched with the gallery.

The features are organized in two families: on one side, are reported the cues that focus on aesthetic aspects [3, 11]: in practice, they encode low-level global image properties. On the other side, there are the *content-based* features, which individuate local image patterns representing semantic entities (cars, chairs and the like); to this end, robust probabilistic object detectors have been employed [5] (for a complete list of all the detectable objects, see Table. 1); other than the number of objects in an image, the object area (that is, the bounding box of the detected object) is also retained. In the case of multiple instances of the same object, the average area is considered. Faces have been extracted adopting the standard Viola-Jones face detection algorithm [23]. Finally, the GIST descriptor [15] for scene categorization has been 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 intensity count. This is needed for the modeling with the Counting Grid. For this reason, features working on angular measures (as those modeling the Hue channel in the HSV color space) have been discarded. For more details on the features, see [9]. Since the range values are very heterogeneous, each feature is normalized across all training images to have zero mean and unit standard deviation. The same normalization is then applied to the features extracted from test data.

3.2 Initialization Stage: Multi-resolution Counting Grid Training

Given the bags of features, the extent E of the classical Counting Grid and its window size S, a multi-resolution CG

Category	Name	L	Short Description
Perceptual	Use of light	1	Average pixel intensity of V channel [3]
	HSV statistics	3	Mean of S channel and standard deviation of S, V channels [11]
	Emotion-based	3	Amount of *Pleasure*, *Arousal*, *Dominance* [11, 22]
	Circular Variance	1	*Circular variance* of the H channel in the IHLS color space [12]
	Colorfulness	1	Colorfulness measure based on Earth Mover's Distance (EMD) [3, 11]
	Color Name	11	Amount of *Black*, *Blue*, *Brown*, *Green*, *Gray*, *Orange*, *Pink*, *Purple*, *Red*, *White*, *Yellow* [11]
	Entropy	1	Image entropy [10]
	Wavelet textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on the HSV channels [3]
	Tamura	3	Amount of *Coarseness*, *Contrast*, *Directionality* [21]
	GLCM-features	12	Amount of *Contrast*, *Correlation*, *Energy*, *Homogeneity* for each HSV channel [11]
	Edges	1	Total number of edge points, extracted with Canny [10]
	Level of detail	1	Number of regions (after mean shift segmentation) [1, 7]
	Regions	1	Average *size* of the regions (after mean shift segmentation) [1, 7]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [3, 11]
	Rule of thirds	2	Mean of S,V channels in the inner rectangle of the image [3, 11]
	Image parameters	1	Aspect ratio of the image [3, 10]
Content	Objects	28	Objects detectors [5]: 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 [23]
	GIST descriptors	24	Level of openness, ruggedness, roughness and expansion for scene recognition [15].

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

is learned. In practice 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), the prior parametrization for $\pi^{(r)}$ is the CG learnt at the previous step, i.e., $\pi^{(r-1)}$. At the first resolution level, the initialization is random. In simple words, the size of the window S determines how smoothed is the latent space where the images coexist: the larger is the window, the smoother is the mapping, and the larger is the neighborhood where similar images could be mapped. In practice, operating with a large window size corresponds to heavily smooth the images, capturing solely their main characteristics, while at the smaller scale, all the details of an image contribute to determine a precise location in the grid. Therefore, having a set of multi-resolution CGs corresponds to analyze the images at a different level of detail, from a coarse to a fine grain.

For the sake of visualization, 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. In Fig. 2, on the left, the Counting Grid with $E = 45$ at resolution $r = R$ ($S = 10$, maximum resolution) is reported. As visible, close images are visually similar, and semantic topics do emerge[2].

3.3 Enrollment Stage

Once the multi-resolution Counting Grid is learnt, the images of each gallery user can be mapped within it, obtaining R maps, one for each resolution. The generative embedding occurs by calculating a posterior probability at each location $\mathbf{k}$; once we have fixed a user u and a resolution r the posterior is

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

where T_u identifies the set of images of the user u: T_u can have different cardinalities, 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 scale 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.

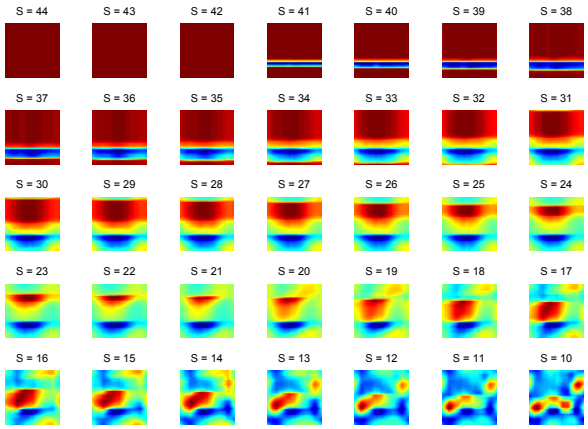

Figure 3: Embedding maps for user 38 of Fig. 2. Starting from the lowest resolution (r=1, S=44) and going towards higher resolutions, the maps show refined blobs and areas, identifying more precisely semantic areas, easily interpretable, on the grid.

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 left are reported the embedding maps of a single resolution level (the maximum, i.e., r=R) for three 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 tastes. This fact is confirmed by checking the random pictures of the users, on the right[3].

In Fig. 3 are reported the R mappings for the user 38 of Fig. 2. Starting from very blurred and unstructured maps

[2]A larger figure is reported in the additional material, also depicting different CGs at different resolution

[3]The complete set of images of these users are reported in the additional material.

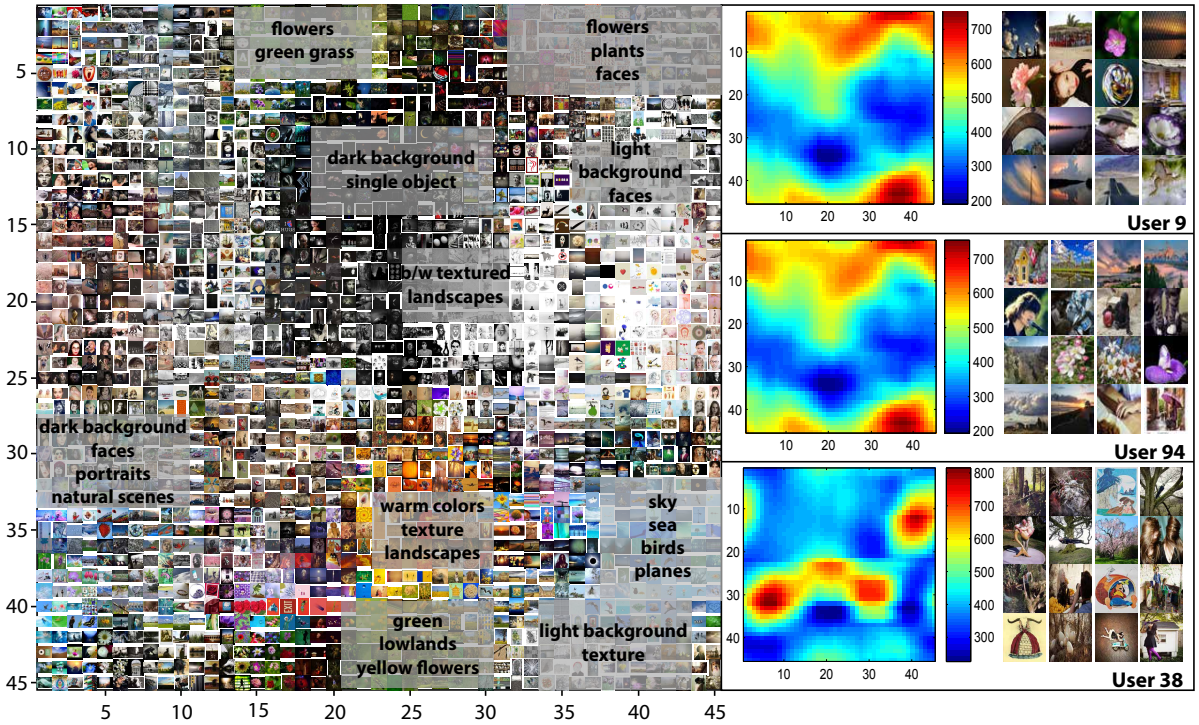

Figure 2: Counting grid at resolution $r = R$: on the left, the grid is visualized as a collage of images that, at a given location k, exhibit the highest posterior probability $p(\mathbf{k}^t|\{c_z^t\})$. On the right, the embedding maps of a single resolution level (r=R) are reported for three subjects, together with some random images preferred by them.

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 discriminative classifiers $\{\lambda_u^{(r)}\}_{r=1,\dots,R}$ are learnt (one for each resolution). In this study, Support Vector Machines with radial basis functions have been employed: in particular, SVMs take as positive samples the maps $\{\gamma_u^{(r)}\}_r$, while as negative samples the maps of all the other gallery users. This step concludes the enrollment stage.

3.4 Identification Stage

In the identification 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 $\{\gamma_v^{(r)}\}_{r=1,\dots,R}$ are used as input of the SVMs related to the gallery user u; they classify the maps producing R scores $\{c_{u,v}^{(r)}\}_{r=1,\dots,R}$ 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.

4. EXPERIMENTAL EVALUATION

The general aim of the experiments is to explore to which extent the personal aesthetics signature is effective in a soft-biometrics context; to this aim, and for the sake of comparability, the same experiments carried out in [9] have been taken into account, for what concerns the identification task[4]. In addition, several experiments have been performed, to investigate the peculiarities of our proposal. In all the experiments, the Flickr dataset has been divided into a training and testing partitions, each composed by 100 preferred images. The training partition has been used to learn the Counting Grid, to produce the embedding maps of the gallery users and to learn the gallery SVMs. The testing partition has been used to select the probe images, mapping them into the CG and producing the probe embedding maps.

The first experimental scenario considers an identification task: here the soft-biometric system wants to recognize the user of a Flickr account, given a pool of unknown preferred images - all liked by the same individual - against a set of gallery users. The second set of experiments is aimed to individuate the limitations of personal aesthetics, and in particular what happens when the users are forced to select from a relatively small number of test images and not from a potentially infinite set as the Flickr repository could be considered.

In both the cases, the parametrization of the multi-resolution Counting Grid is the same: the CG size has been fixed at $E = 45$ pixels, while the (smallest) window size has been set to $S = 10$; this generates a set of 35 maps per user. The extraction of the image features takes 60 minutes per user

[4]The verification experiments in [9] have also been taken into account, not reported here for the lack of space, essentially confirming the superiority of our approach.

(100 images), on a not optimized MATLAB code run on a 3.4 GHz processor with 16 Giga of RAM (the long time is due to the object detectors). 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 "capacity" in [16]) 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 Identification Results

The results of the identification tasks have been carried out following the protocol of [9], for a fair comparison; cross-validation has been performed using 2-fold scheme, repeating each experiment 20 times and shuffling the gallery/probe partition of each user. We also crossvalidated the parameters C and g of the SVM classifier obtaining the best configuration with $C = 1000$ and $g = 0.001$.

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, the SVM classification (averaged) confidence score produced by each of the gallery classifiers has been analyzed. Hopefully, the gallery user with highest score is the one related to the user who originally selected as favorite the photo or the set of photos. To evaluate the recognition capability the Cumulative Matching Characteristic (CMC) curve has been computed [13]; the CMC is a widely-known performance measure in the field of person recognition/identification. 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. Fig. 4 shows various CMC curves for the dataset, where the curves have been obtained by averaging the CMC curves of the 20 different experiments with different gallery/probe splits. In particular, four different CMCs are reported, 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 practice, it is assumed to have 1, 5, 20 and 100 images which have been faved by the same unknown user.

Table 2 reports the CMC mean values (plus standard deviations) for the different ranks, for a quantitative analysis. We also reported the normalized area under the CMC curve (nAUC) as global performance measure.

As visible, the proposed approach definitely overcomes the performance of [9], at every signature cardinality, at each rank. As expected, having more images as test does improve systematically the identification performances. In particular one can note that, having 20 images as test signature allows to reach an average probability of guessing the correct user at the first rank of 0.66, which is definitely above the chance. The probability raises at 0.88 if we check the event of having the correct user within the first 5 ranked users.

Another experiment consisted in fixing the number of test images to 100 and varying the cardinality of the gallery signature, that is, the number of images used to compose the embedded maps fed into the SVM classifiers.

The curves are reported in Fig. 5, and the diverse rank values are listed in Table. 3.

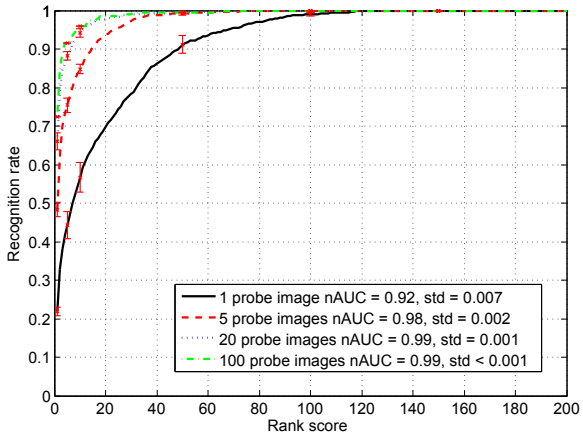

Figure 4: Cumulative Matching Characteristic curve for our approach, while varying the number of test images used to compose the probe signature. The normalized area under the curve (nAUC) is also reported.

T_{te}	Met.	rank 1	rank 5	rank 20	rank 50	nAUC
1	[9]	0.06	0.18	0.40	0.82	0.76
	our	**0.22±0.01**	**0.44±0.04**	**0.70±0.04**	**0.91±0.02**	**0.92±0.007**
5	[9]	0.14	0.39	0.68	0.96	0.89
	our	**0.48±0.02**	**0.75±0.02**	**0.94±0.01**	**0.99±<0.01**	**0.98±0.002**
20	[9]	0.25	0.62	0.88	0.99	0.96
	our	**0.66±0.02**	**0.88±0.01**	**0.98±<0.01**	**1.00±<0.01**	**0.99±0.001**
100	[9]	0.35	0.79	0.97	0.99	0.98
	our	**0.73±<0.01**	**0.92±<0.01**	**0.98±<0.01**	**1.00±<0.01**	**0.99±0.000**

Table 2: Recognition results, varying the number T_{te} of images that compose the *probe* signatures (and fixing the number of gallery images T_{tr} to 100 for each user). The rank numbers are the x-axis values of the CMC curve we focus on. In practice, the reported values represent the average probability of having the correct match within the first 1-5-20-50 signatures, considering different number of probe images.

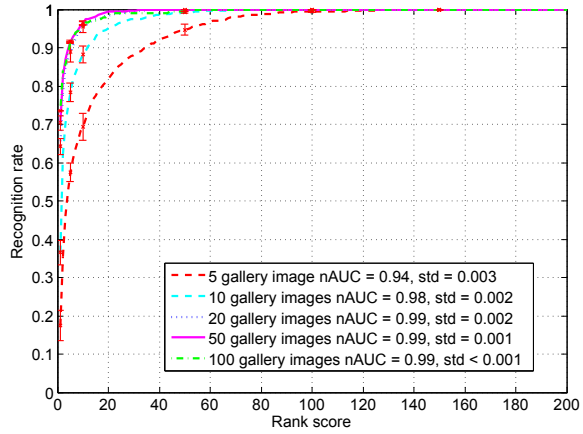

Figure 5: Cumulative Matching Characteristic curve for our approach, while varying the number of train images used to compose the gallery signature. The normalized area under the curve (nAUC) is also reported.

Even in this case the proposed approach sets the new best performance, once again exhibiting higher performance when increasing the number of images adopted to build the gallery signature.

T_{tr}	Met.	rank 1	rank 5	rank 20	rank 50	nAUC
5	*our*	**0.17±0.04**	**0.57±0.02**	**0.82±0.02**	**0.95±0.01**	**0.94±0.003**
	[9]	0.07	0.23	0.49	0.88	0.81
10	*our*	**0.37±0.03**	**0.78±0.02**	**0.95±0.01**	**0.99±<0.01**	**0.98±0.002**
	[9]	0.11	0.32	0.62	0.94	0.87
20	*our*	**0.64±0.02**	**0.89±0.02**	**0.98±0.01**	**1.00±<0.01**	**0.99±0.002**
	[9]	0.15	0.44	0.74	0.97	0.91
50	*our*	**0.70±0.02**	**0.99±<0.01**	**1.00±<0.01**	**0.99±0.001**	**0.92±<0.01**
	[9]	0.22	0.57	0.83	0.99	0.94
100	*our*	**0.73±<0.01**	**0.92±<0.01**	**0.98±<0.01**	**1.00±<0.01**	**0.99±<0.01**
	[9]	0.35	0.79	0.97	0.99	0.98

Table 3: Recognition results, varying the number T_{tr} of images that compose the *gallery* signatures (and fixing the number of probe images T_{te} to 100 for each user). The rank numbers are the x-axis values of the CMC curve we focus on.

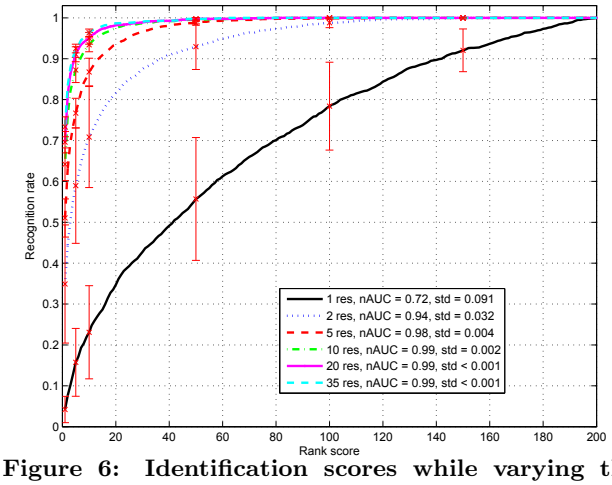

Figure 6: Identification scores while varying the number of resolution employed.

To explore the identification performance while diminishing the number of resolutions employed to learn the CG representation, another test has been performed. In this case 1, 2, 5, 10, 20 and 35 different resolutions have been considered (keep in mind that 35 is the number of resolutions used so far for producing the above results). In the case of a single resolution, all the S windows size between 10 and $E-1=44$ have been independently evaluated, calculating the recognition performance while using 100 images of gallery and 100 of probe. The resulting identification scores have been averaged, providing also the standard deviation values. 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 largest size has been learned with random initialization; the obtained CG has been used as prior for the second ranked one and so on. The experiments have been repeated 35 times, reporting the average recognition score. Results are portrayed in Fig. 6.

As expected, increasing the number of resolution levels does augment the identification capabilities.

4.2 Limitations of our approach

So far, all the works on personal aesthetics did the general assumption that all the images selected from the users, both of training and testing, were not overlapping, that is, no common preferred images are shared among users. In the dataset used so far this hypothesis holds, being the number of repeated images less than the 0.2%. But this is not always the case, especially when a much larger number of users is occurring, or when the images to select come from a restricted number of available pictures. In this last experiment we take into account this situation with a user study: as first operation we build a "reduced" test dataset, by sampling one image from each pool of the 200 originally liked images of all the 200 users, clearly avoiding repeated images. These 200 images have been organized on a web interface, where the users can select them. Then, 16 users of the original dataset have been asked to select from this interface 5, 10, 20 images. After that, the selected images have been used as a test signature for our approach, and compared with the gallery signatures (which for simplicity have been kept equal to the experiments of the previous section), generating three different CMC curves. As comparison, we use the test signatures coming from the original test images of the dataset, and not from the reduced set. The results are shown in Fig.7, which show that the images coming from the reduced dataset have obviously less discriminative power than the ones coming from the original one: this is because of the less aesthetical variability contained within, and to the possible number of images which have been selected by more than one user. In this sense, it is interesting to note that, while increasing the number of images used for building the signature from 5 to 20, the performance of the reduced dataset slightly diminish, while in the original case they obviously augment.

5. CONCLUSIONS

So far, aesthetical preferences have been considered in computer science as underlying a common sense of beauty; in fact, the literature focused on the design of methods aimed at evaluating a general "computational aesthetics" of pictures. With the approach of [9], personal aesthetic preferences started to emerge as peculiar features capable of characterizing the identity of a person. In this paper, a second approach in this brand new field is proposed, which consists in a generative embedding strategy: a generative step (the mapping on the Counting Grids) is followed by a discriminative step (the SVM training). This way, exploiting the advantages of hybrid generative/discriminative approaches, the compact and interpretable CG representation becomes a feature for a discriminative classifier, resulting in the new state of the art on a dataset of 200 users and 40K images. This papers presents also one of the main limitation of our approach, that is, the images selected by the user have to come from a large (potentially infinite) set of images, with no images shared among users; when this assumption holds no more, we have two problems emerging: the first is that the variability of the test signatures irremediably diminishes, as the number of images to select from is smaller, and the second is that the number of images which can be chosen by more than a user is no more negligible. This suggests that, for a valid biometric application (and not solely a soft biometric one), different (and more structured) multimodal interaction paradigms would be necessary.

6. REFERENCES

[1] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603 – 619, 2002.

[2] A. Dantcheva, C. Velardo, A. D'angelo, and J.-L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.

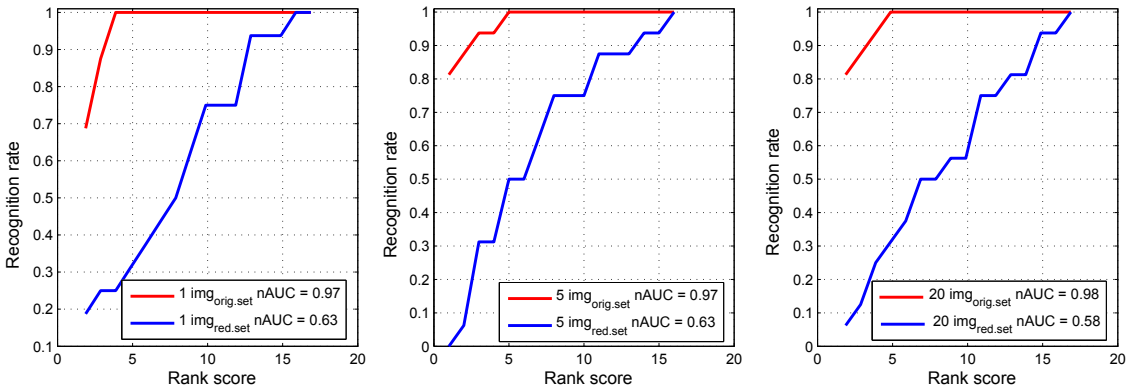

Figure 7: Identification performance while using the original test dataset ($img_{orig.\ set}$) and the reduced dataset ($img_{red.\ set}$).

[3] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, volume 3953, pages 288–301. Springer Berlin / Heidelberg, 2006.

[4] M. Duggan and L. Rainie. Cell phone activities 2012. *Pew Research Center*, 2012.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[6] A. Furnham and J. Walker. The influence of personality traits, previous experience of art, and demographic variables on artistic preference. *Personality and Individual Differences*, 31(6):997–1017, 2001.

[7] C. Georgescu. Synergism in low level vision. In *International Conference on Pattern Recognition*, pages 150–155, 2002.

[8] D. Joshi, R. Datta, E. Fedorovskaya, Q. T. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[9] P. Lovato, M. Bicego, C. Segalin, A. Perina, N. Sebe, and M. Cristani. Faved! Biometrics: Tell Me Which Image You Like and I'll Tell You Who You Are. *IEEE Trans. on Information Forensics and Security*, 9(3):364–374, March 2014.

[10] 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. In *Computer Vision–ACCV 2012*, pages 45–56. Springer, 2013.

[11] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *International Conference on Multimedia*, pages 83–92. ACM, 2010.

[12] K. Mardia and P. Jupp. *Directional Statistics*. Wiley, 2009.

[13] H. Moon and P. J. Phillips. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception-London*, 30(3):303–322, 2001.

[14] L. Olejnik, C. Castelluccia, A. Janc, 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)*, 2012.

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[16] A. Perina and N. Jojic. Image analysis by counting on a grid. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1985–1992. IEEE, 2011.

[17] M. Pusara and C. E. Brodley. User re-authentication via mouse movements. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 1–8. ACM, 2004.

[18] G. Roffo, C. Segalin, A. Vinciarelli, V. Murino, and M. Cristani. Reading between the turns: Statistical modeling for identity recognition and verification in chats. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 99–104. IEEE, 2013.

[19] M. Rybnik, M. Tabedzki, and K. Saeed. A keystroke dynamics based system for user identification. In *Computer Information Systems and Industrial Management Applications, 2008. CISIM'08. 7th*, pages 225–230. IEEE, 2008.

[20] J. Suler. Image, word, action: Interpersonal dynamics in a photo-sharing community. *CyberPsychology & Behavior*, 11(5):555–560, 2008.

[21] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Trans. on Systems, Man and Cybernetics*, 8(6), 1978.

[22] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[24] R. V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, 1(1):81–113, 2008.