BIOMETRICS ON VISUAL PREFERENCES: A "PUMP AND DISTILL" REGRESSION APPROACH

C. Segalin[1] *A. Perina*[2,3] *M. Cristani*[1]

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

ABSTRACT

We present a statistical behavioural biometric approach for recognizing people by their aesthetic preferences, using colour images. In the enrollment phase, a model is learnt for each user, using a training set of preferred images. In the recognition/authentication phase, such model is tested with an unseen set of pictures preferred by a probe subject. The approach is dubbed "pump and distill", since the training set of each user is pumped by bagging, producing a set of image ensembles. In the distill step, each ensemble is reduced into a set of surrogates, that is, aggregates of images sharing a similar visual content. Finally, LASSO regression is performed on these surrogates; the resulting regressor, employed as a classifier, takes test images belonging to a single user, predicting his identity. The approach improves the state-of-the-art on recognition and authentication tasks in average, on a dataset of 40000 Flickr images and 200 users. In practice, given a pool of 20 preferred images of a user, the approach recognizes his identity with an accuracy of 92%, and sets an authentication accuracy of 91% in terms of normalized Area Under the Curve of the CMC and ROC curve, respectively.

Index Terms— behavioral biometrics, image preferences, bagging, LASSO regression

1. INTRODUCTION

Biometric systems have become of paramount importance in the last years [1], with several and heterogeneous traits taken into account. Among these, *behavioral biometric traits* encode a characteristic linked to the behavior of a person [2]. In particular, the so-called HCI-based behavioral biometrics [2] are based on the idea that every person has a unique way to interact with a personal computer, considering for example keystrokes or mouse dynamics [3, 4]. This paper focuses on a novel biometrical trait which exploits the "personal aesthetics" traits of people, i.e., those visual preferences that distinguish people from each other [5]. A personal aesthetics biometric system works in the following manner: in the enrollment phase, the "preference model" of a user is learnt from a set of *preferred images*; in the recognition/authentication phase, such model is tested with an unseen set of favorites preferred by a probe subject. In [5], a binary LASSO-based classifier is employed to learn the user preferences, encoding images as arrays of 34 different visual features, and considering each image as an independent entity. At the best of our knowledge, this is the unique approach which works on "personal aesthetics" biometrics.

In this paper we present a novel framework for personal aesthetics biometrics, based on a statistical classification, where the training stage is characterized by a "pump and distill" strategy. In the "pump" step, the training set of each user (a set of liked images) is augmented by bagging, generating a set of ensembles of preferred images. In the "distill" step, each image of an ensemble is associated to a single *thematic exemplar*, chosen in a set of exemplars, learned beforehand by clustering. For example, we could have in principle clusters of cars, humans, etc., depending on the nature of the clustering procedure, and a thematic exemplar is the centroid of a cluster. All the images of an ensemble linked to the same exemplar are fused together, by averaging them, thus obtaining a surrogate. In practice, a surrogate encodes a customized version of a cluster of a user, capturing what kind of cars, humans, etc. a user likes. Finally, LASSO regression is performed on these surrogates; the resulting regressor, employed as a classifier, takes simple test images belonging to a single user, predicting his identity.

Experiments have been performed on a set of 200 Flickr users, considering 200 preferred images per user. Our approach overcomes definitely the previous method [5] on recognition and authentication tasks in average, promoting our idea as an effective strategy for learning the personal aesthetics.

The rest of the paper is organized as follows: in Sec. 2 we detail our approach, focusing on the "pump and distill" training set generation strategy, and specifying how user recognition and authentication can be performed. The tests in Sec. 3 will explore carefully different configurations of our approach, suggesting how to instantiate the system to produce the best performance. Finally, in Sec. 4, conclusions are given and future perspectives are envisaged.

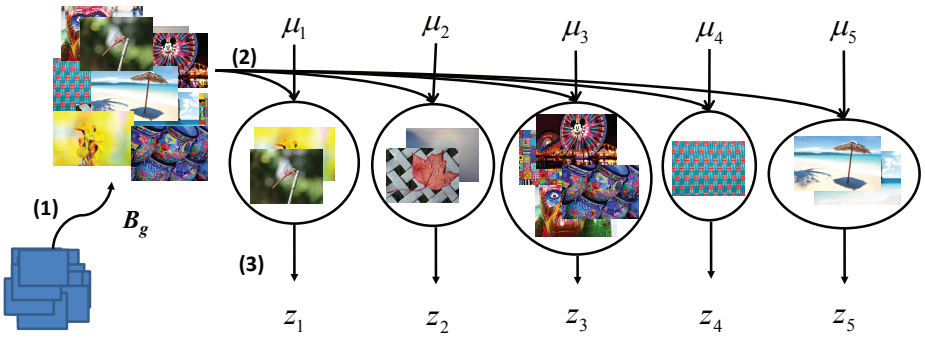

Fig. 1: "Pump and distill" approach; (1) "pumping" a bag B_g from the original image training set; (2) image assignation to a thematic exemplar μ_i; "distilling" all the assigned images to the same exemplar into a surrogate $\mathbf{z}_i$.

2. PROPOSED APPROACH

Our approach focuses on a dataset comprising $V = 200$ Flickr users, each one of them associated with 200 preferred images, that is, images that he likes, for a total of 40000 images. Each image $\mathbf{x}$ is composed by a 62-dimensional real vector, obtained by concatenating the same 34 features employed in [5], ranging from simple color statistics to image aesthetics cues to object detections. In the following, with the term "image" we imply its feature description. For more details on the features, see [5].

The dataset is divided in two partitions, a *gallery* set used for training and *probe* set for testing, both formed by 100 preferred images per user.

We assume that our approach uses the results of a clustering algorithm operating on the 62-dimensional image representations, and in particular we suppose to have K cluster centers, here called "thematic exemplars" μ_k, $k = 1, ..., K$. These exemplars represent different image typologies, depending on the clustering approach employed. For simplicity, here we adopt a simple K-means on the raw 62-dimensional space fed with the training samples, but other, more advanced, partitioning techniques can be applied.

Our approach can be dubbed "pump and distill" to highlight two important steps which characterize the training of the classifiers, one for each user.

In the "pump" step, we augment the training set by employing bagging [6], a strategy designed to improve the stability and accuracy of machine learning algorithms. More in the detail, for each user v, $v = 1, ..., V$, bagging generates G new training sets, the "bags" ${}^{(v)}B_g$, $g = 1, ..., G$, each obtained by sampling uniformly with repetition M times from his training images. In practice, a bag represents a small exert of what is liked by the user.

In the "distill" step, for each bag ${}^{(v)}B_g$ we want to generate a set of *surrogates* $\{\mathbf{z}\}$. In practice, we assign each image $\mathbf{x} \in {}^{(v)}B_g$ to the nearest thematic exemplar μ_k, adopting whatever plausible distance. In our case, we use the simple Euclidean distance. After the assignation, all the images $\mathbf{x} \in {}^{(v)}B_g$ that have been associated to a given cluster are fused together by averaging their feature vectors, creating a surrogate $\mathbf{z}$. For each bag, the number of surrogates may vary from 1 (a single cluster is associated to all the images in the bag) to K (all the clusters have at least one image in the bag associated with them), the last case obviously holding only if M, the number of images per bag, is $>= K$. In this last case, with G bags we end up with $G \times K = N^+$ surrogates per user. In practice, a surrogate encodes a user-customized representation of a cluster, a sort of epitomic representation of a bunch of similar images liked by a user.

At this point we use the pool of surrogates of all the users as distilled training set, to learn a per-user classifier. For this sake, we perform a sparse regression analysis using Lasso [7], assigning to all the training (positive) surrogates $\{\mathbf{z}_n\}$, $n = 1, ..., N^+$ of a user the $y_n = +1$ label, and -1 to all the other N^- surrogates; that is, the negative samples are the positive surrogates of all the other users. With Lasso, we can regress a label assuming it as a linear combination of the image features:

$$y_n = {}^{(v)}\mathbf{w}^T \mathbf{z}_n \tag{1}$$

where ${}^{(v)}\mathbf{w}$ is the weight vector, that is, a linear classifier for the user v, obtained by minimizing the error function with the standard least square estimate:

$$E({}^{(v)}\mathbf{w}) = \sum_{n=1}^{M} \left(y_n - {}^{(v)}\mathbf{w}^T \mathbf{z}_n \right)^2 \tag{2}$$

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

$$\sum_j |{}^{(v)}w^{(j)}| \leq \alpha \tag{3}$$

where ${}^{(v)}w^{(j)}$ is the $j-$th component of the linear classifier. This term acts as a constraint that has to be taken into account

when minimizing the error function, and which enforces the linear classifier to have many coefficients set to 0.

In the testing step, we want to match the probe images of the user u (that is, their 62-dimensional real-valued feature vectors) with the gallery biometrical traits of the user v, represented by his positive surrogates $\{^{(v)}\mathbf{z}_n\}$, $n = 1, ..., N^+$.[1] Clearly, a single image scarcely represent the visual aesthetics sense of a person; therefore, the idea is to consider a *pool* of testing images TE as test biometrical trait, and guess if the pool contains enough information to catch the preferences of the user, allowing to identify him among all the others. In particular we can perform two different operations, that is, user recognition and user authentication.

2.0.1. User recognition

In the user recognition, given the classifier template $^{(v)}\mathbf{w}$ of the user v, the matching score is aimed at measuring how likely the set $\{^{(u)}\mathbf{x}\}$ of the user u contains images which are in accord with the surrogates $\{^{(v)}\mathbf{z}\}$ by the user v. In order to determine it, we compute for every image $^{(u)}\mathbf{x}_m$ in the testing pool TE of user u the regression score $\beta_m^{(u,v)}$, as described by Eq. 1:

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

Then, the final matching score for the whole pool is determined as the averaged regression scores of the images belonging to it, i.e.:

$$\beta^{(u,v)} = \frac{1}{M_{\mathrm{TE}}} \sum_{m=1}^{M_{\mathrm{TE}}} \beta_m^{(u,v)} \tag{5}$$

where M_{TE} is the cardinality of the testing pool.

For the recognition, we compute the matching score of the probe image (or pool) using all the classifiers $\{^{(v)}\mathbf{w}\}$. Hopefully, the gallery user with the highest score is the one who originally faved the photo (or pool of photos). In order to evaluate the recognition rate, we built a CMC curve [8], a common performance measure in the field of person recognition: given a probe set of images coming from a single user and the matching score previously defined, the curve tells the rate at which the correct user is found within the first r matches, with all possible r spanned on the x-axis (in our case, $r = 1, ..., 200$).

2.0.2. User authentication

In this case the system is tested in a authentication scenario: 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. In particular, different kinds of client/impostor signatures may be built, depending on the number of images we take into account as testing pool. Matching a signature composed by more than one image occurs by following what is described previously, i.e., roughly speaking, by averaging the matching scores derived from the set of probe images. 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.

[1] Experimentally, we observed that applying the pump and distill processing to the testing images leads to inferior performance.

3. EXPERIMENTS

The experiments focus on evaluating how effective is our surrogate representation in capturing the personal aesthetics. For each user, we have 200 images: we partition them into a training and a testing set (100 images each), crossvalidating using a 2-fold scheme as in [5], and repeating each experiment 5 times, shuffling the the partitions. To explore different ways of creating the surrogates, for processing the training set we fix $G = 50$ bags and we vary the number of images used to fill them. In particular, we use $M = 5, 10, 20, 50, 100$ images per bag. As number of clusters, we fix $K = 6$. Later in this section, we discuss about varying G and K. Given the surrogates, to learn a LASSO classifier, we decide the best α by a 10-fold cross-validation on a subset of the training set (Eq. 3). As comparison, we consider the approach of [5], which essentially can be thought as having 100 surrogates for training, each formed by a single image. For a fair comparison with the present approach, we run the code of [5] following the same protocol, that is, repeating each experiment 5 times, shuffling train and testing sets. Actually, in [5], experiments were run for a single 2-fold cross-validation run, and authentication results were run for a single test image.

3.1. Identification results

Table 1 shows the recognition results in terms of normalized area under the CMC curve (nAUC CMC). Other than changing the number M of images per bag, we also vary M_{TE}, i.e., the cardinality of the testing set. Please remember, the approach in [5] can be thought as having bags formed by one image. Many observations can be made: 1) the "pump and

M (=$\left\lvert {}^{(v)}B_g \right\rvert$)	nAUC CMC			
	M_{TE}=1	M_{TE}=5	M_{TE}=20	M_{TE}=100
5	**0.69**	**0.83**	**0.92**	**0.96**
10	0.69	**0.83**	**0.92**	**0.96**
20	0.68	0.82	0.91	**0.96**
50	0.68	0.80	0.90	0.95
100	0.66	0.79	0.89	0.95
[5]	0.66	0.77	0.84	0.88

Table 1: nAUC values of the CMC curves varying the num. of images per bag (M), and the num. of test images (M_{TE}). The last row shows the performance of [5].

distill" approach overcomes the approach based on simple images of [5]: this suggests that pooling together images into surrogates produces more discriminative information, which is successfully exploited by LASSO; 2) in general, the "pump and distill' gives its best with testing pools composed by multiple images: the more the images, the higher the nAUC, and this is intuitive. Considering the increment in performance wrt [5], the highest difference holds for $M_{\text{TE}} = 20$, increasing the nAUC of 15%; 3) changing the number of images per bag is not very important, especially when having a high number of test images. Anyway, there is a tendency of having higher results with less images per bag. More into detail, having bags formed by 5 images creates in average 166 surrogates that are formed by more that 1 image (otherwise, we have simple images) in the 38% of the cases, and in particular, 73% of these *proper* surrogates are formed by 2 images, 22% by 3 images, and 5% by 4 images, respectively. In practice, coupling just 2 images into a surrogate substantially ameliorates the classifier.

In Fig. 2 we show the CMC curve of our approach using $M = 5$ images per bag, and $\text{TE} = 20$ test images, against the [5] approach. In this way, we can detail how or approach overcomes the competitor. Actually, the highest difference in terms of recognition rate (the probability of having the correct match in the first r ranked positions) is localized in the first ranks (see the table in the figure), which is very beneficial in terms of a real biometric system, where is expected to find the correct match in the top ranked positions. For example, the probability of having the correct match in the first 10 position, in our approach, is 62%, against the 46% of the competitor.

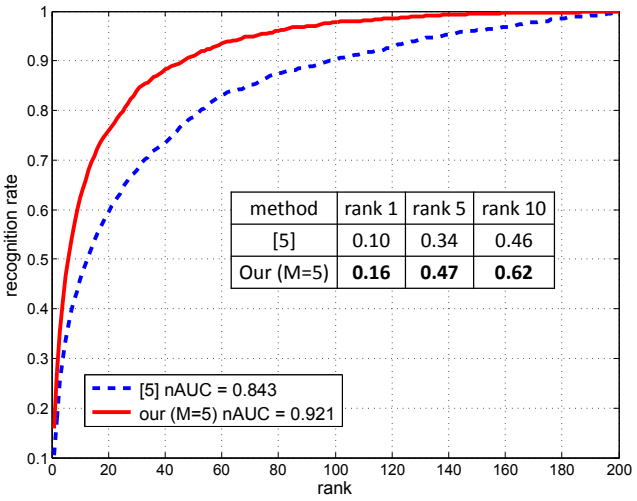

Fig. 2: Recognition results: CMC curves and recognition rates at rank 1,5,10.

3.2. Verification results

Following what described in Sec. 2.0.2, we show in Table 2 the nAUC values related to the ROC curves. In a similar way of the previous section, here we vary the number M of images per bag, and the number M_{TE} of test images.

M (=$\left\|{}^{(v)}B_g\right\|$)	nAUC ROC			
	M_{TE}=1	M_{TE}=5	M_{TE}=20	M_{TE}=100
5	**0.70**	**0.83**	**0.91**	**0.95**
10	0.67	**0.83**	**0.91**	**0.95**
20	0.67	0.82	0.90	0.94
50	0.67	0.80	0.90	0.94
100	0.66	0.79	0.88	0.94
[5]	0.65	0.76	0.82	0.87

Table 2: nAUC values of the ROC curves varying the number of images per bag (M), and the number of test images (M_{TE}). The last row shows the nAUC scores of the ROC curves related to the [5] approach.

Here, similar considerations to those of the recognition case can be carried out, so having few images per bag (5,10) gives the best performance.

As for changing the values of the number of clusters K, we observe on most of the results that the performance (on both recognition and authentication) is similar for $K = 5, 6, 7, 8$, while start decreasing for smaller and bigger values of K. We monitored also the performance while changing the number of bags G. In this case, in average, the results exhibit a slight increase in the interval $[20, 50]$ bags, decreasing for a lower number of bags, and keeping constant after 50 bags until 200, where overfitting starts to reduce the performance.

4. CONCLUSIONS

Behavioral biometrics is a novel and promising trend of the last years, and in this paper we show a inherent brand-new scenario, that is, guessing the identity of a user by looking to the images he/she prefers; in other words, modeling its personal aesthetics. In this paper we show an effective yet simple method to increase the performances of recognition and authentication, using a "pump and distill" strategy, which creates informative image surrogates as training units. A surrogate is the fusion of multiple images which have similar features, encapsulating the personal aesthetics of a person in a more effective way than using the simple images as basic entities. As future perspectives, we plan to use different features; most interesting, the clustering step will be explored in more deep, accounting for grouping strategies expressively based on content-based arguments (that is, pooling together all the images representing particular objects or scene) or focusing on perceptual cues. As a matter of fact, the clustering performed in this paper produces clusters with no clear semantic meaning, even in some case some images of natural scenery, or objects, seem to be consistently grouped.

5. REFERENCES

[1] A.K. Jain, P. Flynn, and A.A. Ross, *Handbook of Biometrics*, Springer, 2008.

[2] R.V. Yampolskiy and V. Govindaraju, "Behavioural biometrics: a survey and classification," *Int. J. Biometrics*, vol. 1, no. 1, pp. 81–113, 2008.

[3] M. Pusara and C.E. Brodley, "User re-authentication via mouse movements," in *ACM workshop on Visualization and data mining for computer security*. 2004, pp. 1–8, ACM.

[4] M. Rybnik, M. Tabedzki, and K. Saeed, "A keystroke dynamics based system for user identification," in *Int. Conf. on Computer Information Systems and Industrial Management Applications*, 2008, p. 225230.

[5] Pietro Lovato, Alessandro Perina, Nicu Sebe, Omar Zandon, Alessio Montagnini, Manuele Bicego, and Marco Cristani, "Tell me what you like and ill tell you what you are: Discriminating visual preferences on flickr data," in *Computer Vision - ACCV 2012*, KyoungMu Lee, Yasuyuki Matsushita, JamesM. Rehg, and Zhanyi Hu, Eds., vol. 7724 of *Lecture Notes in Computer Science*, pp. 45–56. Springer Berlin Heidelberg, 2013.

[6] Leo Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[8] H. Moon and P.J. Phillips, "Computational and performance aspects of pca-based face-recognition algorithms," *Perception*, vol. 30, pp. 303 – 321, 2001.