# A Fast and Robust Negative Mining Approach for Enrollment in Face Recognition Systems

Samuel Botter Martins
Institute of Computing
University of Campinas (Unicamp)
Email: sbmmartins@ic.unicamp.br

Giovani Chiachia
Institute of Computing
University of Campinas (Unicamp)
Email: chiachia@ic.unicamp.br

Alexandre Xavier Falcão
Institute of Computing
University of Campinas (Unicamp)
Email: afalcao@ic.unicamp.br

***Abstract*—Consider a face image data set from clients of a company and the problem of building a face recognition system from it. Video cameras can be used to acquire several images per client in order to maximize the robustness of the system. However, as the data set grows huge, the accuracy of the system might be seriously compromised since the number of negative samples for each user is increasing. We propose here a first solution for this problem, which (i) limits the number of negative samples in the training set for preserving responsiveness during user enrollment, (ii) selects the most informative negative samples with respect to each user for preserving accuracy, and (iii) builds a user-specific classification model. We combine a high-dimensional data representation from deep learning with a method that selects negative samples from a large mining set and builds, within interactive times, effective user-specific training set and classifier, using linear support vector machines. The method can also be used with other feature extractors. It has shown superior performance as compared to five baseline methods on three unconstrained data sets.**

## I. Introduction

Over the past two decades, face recognition has been a key research area. Such an effort may be explained by the wide range of applications that require face recognition, such as video surveillance, access control, and on-line transactions. As a consequence, many systems and approaches have been proposed for face recognition, and some of them have achieved state-of-the-art performances in specific applications [1]–[3].

An important component in typical biometric systems, such as face recognition, is *user enrollment*, which is responsible for capturing appropriate biometric readings of a new user to be enrolled in the system and for storing this data either in raw format or as feature vectors or user models. This process is directly related to the approach used to match biometric samples and ultimately recognize the users. For example, a common matching approach consists of computing pairwise distances from a probe sample to gallery samples. The enrollment process in this case essentially consists of storing valid gallery samples — or their corresponding feature vectors — in the system database for later distance computation. While pairwise-matching approaches have been largely used in biometrics, their performance might be seriously compromised as the data set grows large with new users enrolled in the system.

We may divide the design of a biometric system into two strategies: User-Independent (UI) and User-Specific (US). UI

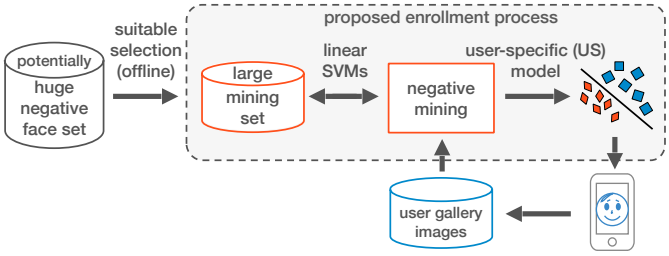



Fig. 1. User enrollment. From a potentially huge data set of negative face images, the algorithm relies on a suitable (under the time constraints) selection of samples to create a large mining set. Then, for a given user, it creates a small training set by identifying the most informative negative samples and builds an effective US model.

models do not require access to gallery samples for their training and therefore can be built offline, even prior to system deployment. Time and memory requirements to learn these models are usually not a matter of concern for the system operation, since the learning task is decoupled from the operation. Principal Component Analysis (PCA) [4] and Linear Discriminant Analysis (LDA) [5] applied on face data sets available at development time are common examples of UI models. US models incorporate gallery samples into the learning task and are usually built with discriminative techniques executed during user enrollment [6] or at matching [7]. One of such approaches is to learn a discriminative binary classifier that assumes the enrolling user as the positive class and a set of face images from unrelated individuals — e.g., from other individuals in a previously curated large face data set — as the negative class. In this case, pairwise matching is replaced by predicting the class to which a probe sample belongs according to the discriminative US model.

US models [3], [8], [9] are usually more effective than UI models and huge annotated data sets, with many individuals and images per individual, can be created by video cameras for the design of robust face recognition systems. However, US models demand critically higher time and memory to be trained as the number of negative samples for a user under enrollment increases, making the choice of the negative training samples very important in number and quality to preserve the responsiveness and accuracy of the system.

We propose here a first solution for the above problem —

a method that, during user enrollment, (i) selects a limited number of the most difficult (informative) negative samples from a large mining set, with respect to the positive samples of that user, and (ii) builds an effective US model within interactive time. Most US models combine a same feature extractor with some US classifier, such as a Support Vector Machine (SVM) [10]. User-specific feature extraction is also an alternative [3], but for the sake of efficiency, our method relies on a same deep learning architecture [8] for feature extraction to build US training sets and linear SVM classifiers.

Figure 1 illustrates the user enrollment process of the proposed approach. The algorithm has shown to be robust in iteratively mining a much smaller and effective subset of negative training samples, according to a criterion based on distances to SVM decision boundaries, under different time constraints. Our solution also exploits the increasing importance of high-dimensional feature spaces in face recognition [3], [11], [12] for *unconstrained* scenarios, where the face images present a large range of the variation in pose, lighting, expression, among others.

We evaluate the method on three unconstrained data sets, namely PubFig83 [2], Mobio [13], and Casia-WebFace [14] and conduct an array of experiments by increasingly mining thousands of available images. Results show that the proposed approach can attain significantly superior performance with respect to five other baselines — which rely on the same classification scheme — within interactive times without negatively affecting the responsiveness and accuracy of the system. Moreover, given that it can be split into client and server tiers — requiring low bandwidth between the tiers — it is also well suited to systems that operate on budgeted devices.

The remainder of this paper is organized as follows. In Section II, we present the related work on user enrollment in face recognition systems and negative mining. The considered feature extraction method is presented in Section III. We introduce the proposed approach in Section IV. The experimental setup is then described in Section V, while in Section VI we present and discuss the results. Concluding remarks are stated in Section VII.

## II. Related work

The vast majority of face recognition systems operate with user-independent (UI) models, previously built without regarding particularities in the appearance of the individuals to be recognized [8], [11], [15]. While such strategy may avoid the burden of using complicated learning tasks in the operational scenario — and it is well aligned with the evaluation protocol of a number of public face recognition benchmarks [16], [17] — it completely disregards the opportunity of leveraging gallery samples to build better models and improve the overall system performance.

In this spirit, several works have proposed robust face recognition systems based on US models that operate in *open-* and *closed-set* scenarios [2], [3], [6], [9], [12]. However, works based on US models are usually targeted at recognition performance and often employ time and memory demanding procedures to build them. Therefore, they do not assume user enrollment as a fundamental time constrained process in user interactive face recognition systems, and may be impractical for real applications whose databases have the potential to become huge.

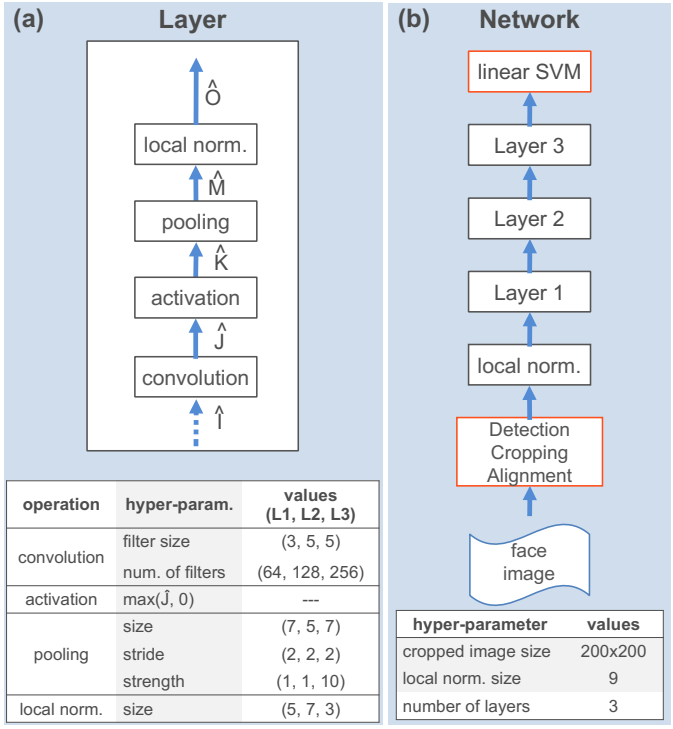
To our knowledge, this is the first work to propose negative mining for user-specific (US) gallery model building at enrollment time. Perhaps the most related work to ours is [6], where the authors propose the use of partial least squares (PLS) [18] to build US models. Nevertheless, our work differs from this one in at least two fundamental ways. First, we mine negative *samples* instead of negative *individuals*. Second, and more importantly, we do not build US models against gallery samples in a closed-set scenario. Instead, we rely on a previously curated large face data set to mine negative samples. This not only avoids the burden of gallery maintenance, which is the focus of [6], but it is also more realistic, since it is aligned with face recognition in the open-set scenario [19].

In turn, negative mining has been extensively used in Computer Vision, especially for object detection [20]–[23]. The simplest strategy is to sample the data set randomly, which is a clearly sub-optimal approach. A more satisfactory strategy consists of applying *bootstrapping* techniques.

Essentially, a common negative mining approach consists of two steps. First, a binary classifier is trained using the positive samples and an initial random subset of negative samples. The second step is inspired on the *bootstrapping* procedure [24], and consists of *mining* negative samples by giving more importance to the "hard" ones — i.e., the incorrectly classified negative examples — thereby improving the training set. A new classifier is then trained and this procedure may be repeated a few times.

Felzenszwalb et al. [21] present a general negative mining method for object detection systems that uses classical SVMs and latent SVMs. The method iteratively solves a sequence of training problems using a relatively small number of hard examples from a large training set. Its novelty is a theoretical guarantee that it leads to the *exact* solution of the training problem defined by the large training set. As critical shortcoming, the intermediary training sets may grow considerably, requiring a high processing time, since the number of negative training samples is not limited. This makes it impractical for the problem addressed in this work.

Our approach has similarities with [21], which also uses SVMs as basis for mining negative samples, but has also differences, because we conceived it to operate with a few positive samples and to not allow the negative set to grow arbitrarily. Both of these characteristics were incorporated to make the approach suitable for the enrollment process in face recognition systems. In fact, in terms of time and memory concerns, our work is more aligned with [22], but while the latter is targeted at pedestrian detection, here our target is robust face recognition.




| operation | hyper-param. | values (L1, L2, L3) |
|---|---|---|
| convolution | filter size | (3, 5, 5) |
| | num. of filters | (64, 128, 256) |
| activation | max($\hat{J}$, 0) | --- |
| pooling | size | (7, 5, 7) |
| | stride | (2, 2, 2) |
| | strength | (1, 1, 10) |
| local norm. | size | (5, 7, 3) |

| hyper-parameter | values |
|---|---|
| cropped image size | 200x200 |
| local norm. size | 9 |
| number of layers | 3 |

Fig. 2. The schematic diagram illustrates how the main operations are combined within each layer (a) of three feed-forward layers (b) in the HT-L3-1st model. The tables show (a) the values of all hyper-parameters of each layer (L1, L2, and L3) and (b) the global-hyper-parameters for the adopted network architecture. Face images are previously detected, cropped, and aligned by the position of the eyes.

## III. High-Dimensional Feature Extraction

We use a ConvNet, Convolutional Network [25], named HT-L3-1st [8], for feature extraction. This choice is justified by its success in face recognition problems [2], [9] where images are acquired with no control over illumination, facial expression, pose — the unconstrained scenario. Note, however, that the proposed negative mining method independs of the feature extractor.

A ConvNet is composed of non-linear and linear image processing operations, stacked to extract deep representations, called *multiband images*, whose pixel attributes are concatenated into **high-dimensional** feature vectors for pattern recognition. The set of its hyper-parameters is called architecture. Figure 2 illustrates how the main operations are combined within each layer (a) of three feed-forward layers (b) in the HT-L3-1st model. The hyper-parameters size and stride of a given value $b$ must be understood as $b \times b$. The faces are first detected, cropped, and aligned by the position of the eyes. For a given US training set, where the positive and negative face samples are feature vectors, the system trains a linear SVM classifier.

This section describes ConvNets from the image processing perspective. Let $\hat{I} = (D_I, \vec{I})$ be a multiband image, where $D_I \subset Z^2$ is the image domain and $\vec{I}(p) = \{I_1(p), I_2(p), \ldots, I_m(p)\}$ is the attribute vector of a pixel $p = (x_p, y_p) \in D_I$. ConvNets use box adjacency relations $A$ of $D_I$ of size $b \times b$, i.e., a pixel $q \in A(p)$, if $|x_q - x_p| \leq \frac{b}{2}$ and $|y_q - y_p| \leq \frac{b}{2}$.

### A. Filter Bank Convolution

Let $\Phi_i = (A, \vec{W}_i)$ be a multiband filter with weight vector $\vec{W}_i(q) = \{w_{i,1}(q), w_{i,2}(q), \ldots, w_{i,m}(q)\}$, where $q \in A(p)$ is adjacent to the origin $p$ of the filter. A multiband filter bank $\Phi = \{\Phi_1, \Phi_2, \ldots, \Phi_n\}$ is a set of filters $\Phi_i = (A, \vec{W}_i)$, with $i = \{1, 2, \ldots, n\}$. The weights of a filter $\Phi_i$ are randomly generated from a uniform distribution, and normalized to zero mean and unit norm. The convolution between an input image $\hat{I}$ and a filter $\Phi_i$ produces a band $i$ of the filtered image $\hat{J} = (D_J, \vec{J})$, where $D_J \subset D_I$ and $\vec{J} = (J_1, J_2, \ldots, J_n)$.

### B. Activation

Activation creates an image $\hat{K} = (D_J, \vec{K})$ by $K_i(p) = \max(J_i(p), 0)$, where $p \in D_J$ and $i = \{1, 2, \ldots, n\}$ are the image bands. This definition combined with random filters of zero mean and unit norm has the purpose of creating a sparse code to improve the effectiveness of feature extraction.

### C. Spatial Pooling

Spatial pooling is a very important operation that aims at bringing small translation invariance by aggregating activations from the same filter within a given region. Let $D_M = D_K/s$ be a regular subsampling of every $s \times 1$ pixels in an adjacency $B$. The value $s$ is the stride of the pooling operation, which results into a spatial resolution reduction when $s > 1$. The pooling operation creates the image $\hat{M} = (D_M, \vec{M})$, defined by

$$M_i(p) = \sqrt[\alpha]{\sum_{\forall q \in \mathcal{B}(p)} K_i(q)^\alpha} \qquad (1)$$

where $p \in D_M$, $i = 1, 2, \ldots, n$ are the image bands, and $\alpha$ controls the pooling sensitivity. The values $\alpha = (1, 1, 10)$, strength in Figure 2a, indicate additive pooling for L1 and L2, and max-pooling for L3.

### D. Divisive Normalization

The last operation is the divisive normalization, which is based on gain control mechanisms found in cortical neurons [26]. It is also applied with adjacency size $9 \times 9$ to the input image. It creates an output image $\hat{O} = (D_O, \vec{O})$, $D_O \subset D_M$, where $\vec{O}(p) = \{O_1(p), O_2(p), \ldots, O_n(p)\}$ and

$$O_i(p) = \frac{M_i(p)}{\sqrt{\sum_{j=1}^{n} \sum_{\forall q \in \mathcal{C}(p)} M_j(q) M_j(q)}} \qquad (2)$$

for some adjacency $C$. It promotes a competition among pooled filter bands, such that high responses prevail over low ones. The output feature vector results from the concatenation of $\vec{O}(p)$ for all $p \in D_O$ when $\hat{O}$ is the output of the last layer.

## IV. Proposed Approach

We propose a negative mining approach based on linear Support Vector Machines (SVMs) [10] with the following motivations. First, the ability to perform well with small sample sizes, especially in the case where the samples are represented by high-dimensional feature spaces, and second, we can train linear SVMs quite fast under these circumstances.

**Algorithm 1.** Proposed SVM-based Negative Mining

| | |
|---|---|
| Input: | Positive set $P$, large mining set $N$, maximum processing time $max\_time$, and number of negatives to be mined $c$. |
| Output: | Best model $\beta_{out}$ for the positive set $P$. |
| Auxiliary: | Sets $N_t$, $N_v$, lists $L_t$, $L_v$, variables $\beta$, $swaps$, $stop$, $s$, $t$, $ds$, $dt$, $pt$, $proc\_time$. |

1. $N_t$ ← random selection of $c$ samples from $N$
2. $N_v$ ← $N \setminus N_t$
3. $proc\_time$ ← 0
4. $\beta_{out}$ ← NIL
5. **While** $proc\_time < max\_time$
6. $\quad pt$ ← $point\_time()$
7. $\quad \beta$ ← linear SVM trained on $P \cup N_t$
8. $\quad proc\_time$ ← $proc\_time$ + ($point\_time()$ − $pt$)
9. $\quad$ **If** $proc\_time > max\_time$
10. $\quad\quad$ **Return** $\beta_{out}$
11. $\quad \beta_{out}$ ← $\beta$
12. $\quad pt$ ← $point\_time()$
13. $\quad L_t$ ← empty list, $L_v$ ← empty list
14. $\quad$ **For each** $s \in N_t$ not support vector
15. $\quad\quad$ insert $(s, \beta(s))$ into $L_t$
16. $\quad$ **For each** $t \in N_v$
17. $\quad\quad$ insert $(t, \beta(t))$ into $L_v$
18. $\quad L_t$ ← sort $L_t$ by $\beta(.)$ in increasing order
19. $\quad L_v$ ← sort $L_v$ by $\beta(.)$ in decreasing order
20. $\quad swaps$ ← 0, $stop$ ← 0
21. $\quad$ **While** $L_t \neq \emptyset$ and $L_v \neq \emptyset$ and $stop \neq 1$
22. $\quad\quad$ remove $(s, ds)$ from $L_t$ head
23. $\quad\quad$ remove $(t, dt)$ from $L_v$ head
24. $\quad\quad$ **If** $dt < ds$
25. $\quad\quad\quad N_t \leftarrow (N_t \cap s) \cup t$
26. $\quad\quad\quad N_v \leftarrow (N_v \cap t) \cup s$
27. $\quad\quad\quad swaps \leftarrow swaps + 1$
28. $\quad\quad$ **Else**
29. $\quad\quad\quad stop \leftarrow 1$
30. $\quad$ **If** $swaps = 0$
31. $\quad\quad$ **Return** $\beta_{out}$
32. $\quad proc\_time$ ← $proc\_time$ + ($point\_time()$ − $pt$)
33. **Return** $\beta_{out}$

A pseudocode of the proposed negative mining is presented in Alg. 1. The algorithm considers gallery images of the individual being enrolled as the positive set $P$ and a much larger negative mining set $N$ from which a small set of $c$ informative images must be iteratively mined within a given maximum processing time $max\_time$. The mining set is split into a negative training set $N_t$ ($|N_t| = c$) and a negative validation set $N_v$.

A linear SVM $\beta$ is trained at each iteration by taking $P \cup N_t$ as input. If the algorithm processing time $proc\_time$ right after the SVM training exceeds $max\_time$, the algorithm terminates and returns the model $\beta_{out}$, which is either nil — no linear SVM could be trained within $max\_time$ — or points to the model trained at the previous iteration. Otherwise, the algorithm saves the newly trained model in $\beta_{out}$ and continues its execution. For the sake of clarity, the function $point\_time()$ points the current running time during the execution of the algorithm.

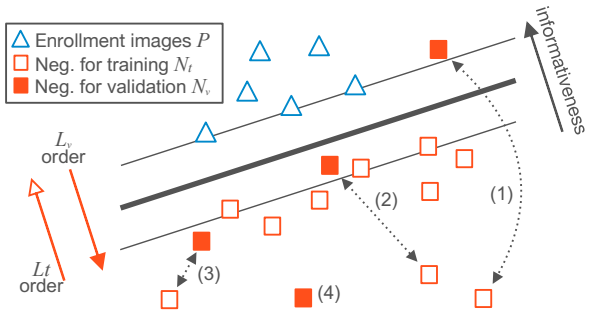



Fig. 3. Mining process in a given iteration. The least informative samples in $N_t$ that are *not* support vectors are swapped with the most informative ones in $N_v$, as indicated by the swapping sequence (1), (2), and (3). Swapping occurs no matter each side of the margin the negative samples are, which increases the ability of the method to operate well even in unbalanced learning scenarios. In (4), no swap occurs because such validation sample is less informative than any other available for swapping in $N_t$.

The signed distances to the SVM hyper-plane of all samples in the negative training set — except support vector — and in the validation set are computed and inserted into the lists $L_t$ and $L_v$. These lists are then sorted, according to the signed distance $\beta(.)$, for subsequent sample swapping.

Images are swapped between $N_t$ and $N_v$ according to a criterion based on an "informativeness" degree, which is exactly the signed distance $\beta(.)$ to the SVM hyper-plane of the given iteration.

Given a sample $s \in N_t \cap N_v$, the assumption is that the greater $\beta(s)$ is, the more informative for the gallery model $s$ will be. Therefore, the least informative samples in $N_t$ that are *not* support vectors are swapped with the most informative ones in $N_v$. If no improvement in the overall informativeness of $N_t$ is observed in a given iteration — i.e., no swaps occurred — or if the maximum processing time $max\_time$ is reached, the algorithm terminates and returns the current valid model for the individual being enrolled.

An important property of the approach as compared to [21] is that correctly classified negative samples may also be swapped, which enables it to mine negative samples even in extremely unbalanced learning scenarios. Moreover, we can see from Alg. 1 that its running time is dominated by the SVM training in Line 7, which can ranges from quadratic to cubic on the size of the input training set, depending on the regularization constant $C$ [27]. Given that the number of negative samples predominates over the number of positive samples, our expectation is that learning gallery models by iterating a few times the mining process with $c << |N|$ will probably

speedup the enrollment process while not compromising the recognition performance.

## V. Experimental Setup

### A. Data Sets

The experiments used three unconstrained data sets, namely PubFig83 [2], Mobio [13], and Casia-WebFace [14]. Fig. 4 presents some images of each data set.

PubFig83 is a collection of real-world face images of 83 celebrities collected from the Internet with at least 100 considerably different, "in the wild" images available per individual. Each image has originally 100x100 pixels in size. In addition to representing a modern and challenging problem, such a remarkable number of diverse images per individual allows for effective evaluations of user-specific (US) models in unconstrained facial recognition.

The Mobio data set used in this work is precisely the same used in the competition on unconstrained face recognition in mobile platforms, organized as part of the *Intl. Conf. on Biometrics*, ICB'13 [9]. This data set has 150 people with a female-male ratio of nearly 1:2 and contains images recorded across 12 sessions with mobile phones and without any control over illumination, and facial expression and pose. We believe that using Mobio is appropriate because it represents a challenging and emergent problem, whose operational requirements can be well addressed by the proposed approach.

Finally, the Casia-WebFace is a large scale data set containing $10,575$ subjects and $494,414$ images collected from the Internet through the website IMDb [1] — a well structured website containing rich information of celebrities, such as name, gender, and photos. Each celebrity has an independent page on this website with a specific "id" which in turn is used to label the subjects from the data set. The number of images per subject varies from 2 to 804 images. As far we know, Casia-WebFace is the largest data set publicly available in the literature.

From all data sets, we extract visual representations using the HT-L3-1st descriptor (see Section III) of [8], obtained by deep learning — a technique that has been successfully applied in several computer vision problems, including face recognition on PubFig83 [2] and Mobio [9]. All faces were previously detected, cropped, and aligned by the position of the eyes. In common with other current top-performing visual representations for face recognition [3], [11], HT-L3-1st has the property of outputting high-dimensional feature vectors, with 25K elements.

### B. Evaluation Protocol

Since the size of the PubFig83 and Mobio data sets is small, we consider that the *mining set* of Fig. 1 is already built for both data sets. Evaluations are then carried out in a realistic *open-set* scenario [19], in that no information of other gallery individuals is used for building US models of new individuals at enrollment time.

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

The Mobio protocol [9] naturally addresses this scenario, and hence we report results using the union of its original *training* and *development set* as the mining set of Fig. 1 (a total of $14,010$ images) and its *evaluation set* as containing images of individuals under enrollment (gallery), i.e., the users of the system for whom the False Acceptance Rates (FAR) and Correct Acceptance Rates (CAR) are calculated.

PubFig83 original evaluation protocol, however, is designed for *closed-set* face recognition. Therefore, we extended the protocol by further splitting the data set into two subsets: one that simulates the mining set, containing images of 58 individuals chosen at random (a total of $5,800$ images), and the other equivalent to the evaluation set (gallery), containing images of the remaining 25 individuals to report FAR and CAR values. Each individual from the evaluation set has 90 training images and 10 probe images. Thus, we can simulate a scenario wherein each user is enrolled with a considerable number of images (e.g., social networks).

For the Casia-WebFace data set, there is no such a established protocol. The data set was originally proposed to train deep convolutional networks for later assessment in other well established benchmarks such as the Labeled Faces in the Wild (LFW) [16]. Therefore, here we propose a protocol for the Casia-WebFace suited for user-specific model training in the open-set scenario.

Firstly, $50$ individuals from the ones that contain at least $50$ images ($2,550$ individuals) are chosen randomly, becoming the *gallery users*. We then choose $50$ images randomly of each one in order to build a balanced set, simulating a system that operates with a doable number of images per user. All images of the other individuals from the data set form the *huge negative set* (Fig. 1). Since that the number of images per individual from the Casia-WebFace is not balanced, the size of the huge negative set may vary between $468,025$ and $491,914$ images depending on the chosen individuals. The mining set of Fig. 1 is then simulated by randomly selecting $25,000$ samples from the huge negative set due to our processing constraints. Random selection allows unbiased evaluation of the methods that take the mining set as input.

Each experiment is repeated ten times and results are reported in terms of mean values of CAR and FAR with their respective standard errors. Prior to the execution of the experiments, individuals from the mining set and evaluation set of each data set are previously chosen and fixed, following the strategy adopted by the Mobio protocol in [9].

Execution times of all experiments were obtained in a same Intel I7-3770k PC with 32GB of RAM, and no memory swapping. We used LIBSVM [28] via Scikit-learn [29] to train the SVMs (Alg. 1, Line 7) with the regularization constant $C$ fixed at $10^5$ as in [2], [3], [9].

### C. Compared Methods

We compared our approach with five others. The first two are User-Independent (UI) models built with PCA [4] and LDA [5], both methods applied in the entire mining set. These techniques are widely used to build offline face recognition

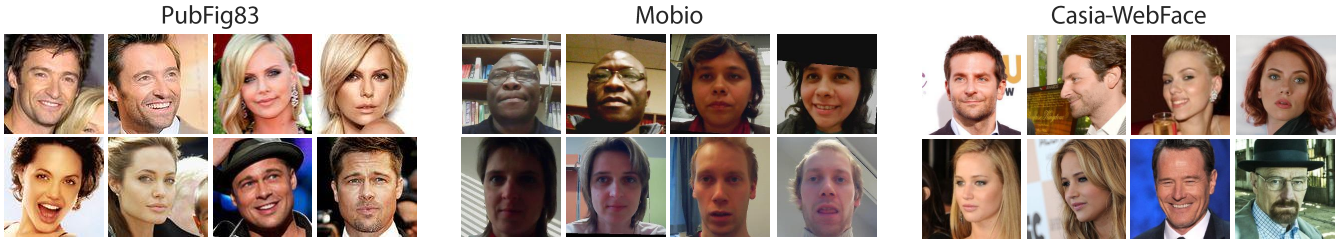



Fig. 4. Images of four individuals of PubFig83, Mobio, and Casia-WebFace. Due to their unconstrained nature, we can observe that all data sets present factors of variation in face appearance: pose, expression, illumination, occlusion, hairstyle, aging, among others.

models, during the conception of the recognition system. This comparison aims to show the superiority of US over UI models. Both PCA and LDA implementations are from Scikit-learn [29], the number of retained projection vectors was according to the rank of the input covariance matrices, and the matching between face samples were done via cosine similarity.

The other compared methods were based on user-specific (US) models. We started by comparing US models built with linear SVMs also using the entire mining set, as in PCA and LDA. Given that this approach is also based on linear SVMs, but uses all negative samples at disposal for learning (no negative mining), we may say that it represents a statistical upper bound for the proposed approach, the last being based on a considerably smaller training set. Therefore, for clarity, we call it *expected upper bound*.

We then evaluate two negative mining approaches, one consisting of a *random selection* of the negative samples — and serving as baseline and sanity check for the proposed approach — and the other implementing the well known SVM-based negative mining criterion of [21].

The processing time of each negative mining method corresponds to the sum of the time spent during the user enrollment to mine the mining set and train the final linear SVM classifier which will be used to assess the final recognition performance.

## VI. Results

Initially, we compared the performance between UI and US models with no negative mining. That is, all methods — PCA, LDA, and linear SVM (our *expected upper bound*) — were trained using the entire *mining set* as input. The results are shown in Table I.

TABLE I
CAR and time in secs (between parentheses) as obtained with UI (PCA and LDA) and US models (expected upper bound) in all data sets for a fixed FAR at 0.01%

| | PubFig83 | Casia WebFace | Mobio Male | Mobio Female |
|---|---|---|---|---|
| PCA | 8.80 (0.00) | $11.92 \pm 0.12$ (0.00) | 32.58 (0.00) | 13.33 (0.00) |
| LDA | 5.20 (0.00) | $1.52 \pm 0.22$ (0.00) | 23.98 (0.00) | 6.05 (0.00) |
| exp. up. bound | 70.40 (12.32) | $46.56 \pm 0.55$ (161.38) | 42.71 (52.25) | 27.14 (49.15) |

The enormous difference in CAR — for a FAR fixed at $0.01\%$ — confirms the superiority of US over UI models and the effectiveness of linear SVMs to deal with high-dimensional feature spaces in unconstrained face recognition scenario. PCA and LDA dismiss learning during user enrollment, which explains the zeros in their learning times and might also explain their poor performance.

The linear SVM with no mining, on the other hand, can negatively affect the responsiveness of the system, since it requires $161.38$ seconds for Casia-WebFace, for instance. The larger the base is, the higher the processing time will be. Thus, negative mining methods are crucial to attain the "ceiling" CAR of the *expected upper bound* within an interactive time, without affecting the responsiveness.

Table II presents the experimental results (also in terms of CAR) of the evaluated mining approaches, all them set to operate with a FAR of $0.01\%$. The considered maximum processing times were chosen based on the time demanded to train the *expected upper bound* (Table I), so that mining methods are evaluated covering different levels of responsiveness and so that maximum allowed times in fact represent time constraints.

We consider a negative training set with $5\%$ of the mining set (parameter $c$ in Algorithm 1) for PubFig83 and $1\%$ for Mobio and Casia-WebFace. These values were chosen based on the data set sizes and our memory and processing limitations. Thus, for each gallery individual being enrolled in the system, all negative mining methods use the same initial negative training set built randomly.

Given that the spent time by *random selection* for outputting the final classifier is less than the smallest time constraint, its CAR value is repeated for all constraints. Indeed, *random selection* is the most efficient approach, but its ability to select informative negative samples for the training set is poor.

The method of [21] presents CAR values similar to *random selection* for the smallest processing time constraint in PubFig83 and Mobio and for all cases in Casia-WebFace. This is a consequence of its mining criterion, which may allow the number of negatives in the training set to grow arbitrarily, resulting in the execution of only a few iterations.

Table II clearly shows that our negative mining approach is able to attain superior recognition performance within a fraction of the time required by the *expected upper bound*, especially in the Mobio and Casia-WebFace data sets. As

TABLE II
CAR AS OBTAINED WITH THE NEGATIVE MINING APPROACHES FOR A FAR FIXED AT 0.01%.

**PubFig83**

| max. time (secs) | Random selection | | Felzenszwalb et al. | | proposed in this paper | |
|---|---|---|---|---|---|---|
| 2.00 | 41.36 | 2.02 | 41.36 | 2.02 | 68.12 | 0.87 |
| 4.00 | | | 41.36 | 2.02 | **70.75** | **0.16** |
| 6.00 | | | 62.76 | 1.32 | 70.16 | 0.19 |
| 8.00 | | | 66.44 | 1.46 | 70.21 | 0.24 |

(a) PubFig83

**Casia-WebFace**

| max. time (secs) | Random selection | | Felzenszwalb et al. | | proposed in this paper | |
|---|---|---|---|---|---|---|
| 4.00 | 19.38 | 0.68 | 19.38 | 0.68 | 19.38 | 0.68 |
| 8.00 | | | 19.38 | 0.68 | 40.74 | 0.52 |
| 12.00 | | | 19.38 | 0.68 | **47.52** | **0.36** |
| 16.00 | | | 19.38 | 0.68 | 46.75 | 0.34 |

(b) Casia-WebFace

**Mobio Male**

| max. time (secs) | Random selection | | Felzenszwalb et al. | | proposed in this paper | |
|---|---|---|---|---|---|---|
| 4.00 | 34.66 | 1.38 | 34.66 | 1.38 | **45.00** | **0.27** |
| 8.00 | | | 37.47 | 1.11 | 43.72 | 0.12 |
| 12.00 | | | 42.25 | 0.22 | 43.81 | 0.05 |
| 16.00 | | | 42.19 | 0.32 | 43.83 | 0.09 |

(c) Mobio Male

**Mobio Female**

| max. time (secs) | Random selection | | Felzenszwalb et al. | | proposed in this paper | |
|---|---|---|---|---|---|---|
| 4.00 | 21.91 | 0.52 | 21.75 | 0.52 | 27.73 | 0.42 |
| 8.00 | | | 27.65 | 0.32 | **27.93** | **0.12** |
| 12.00 | | | 27.28 | 0.13 | 27.65 | 0.05 |
| 16.00 | | | 26.75 | 0.41 | 27.69 | 0.06 |

(d) Mobio Female

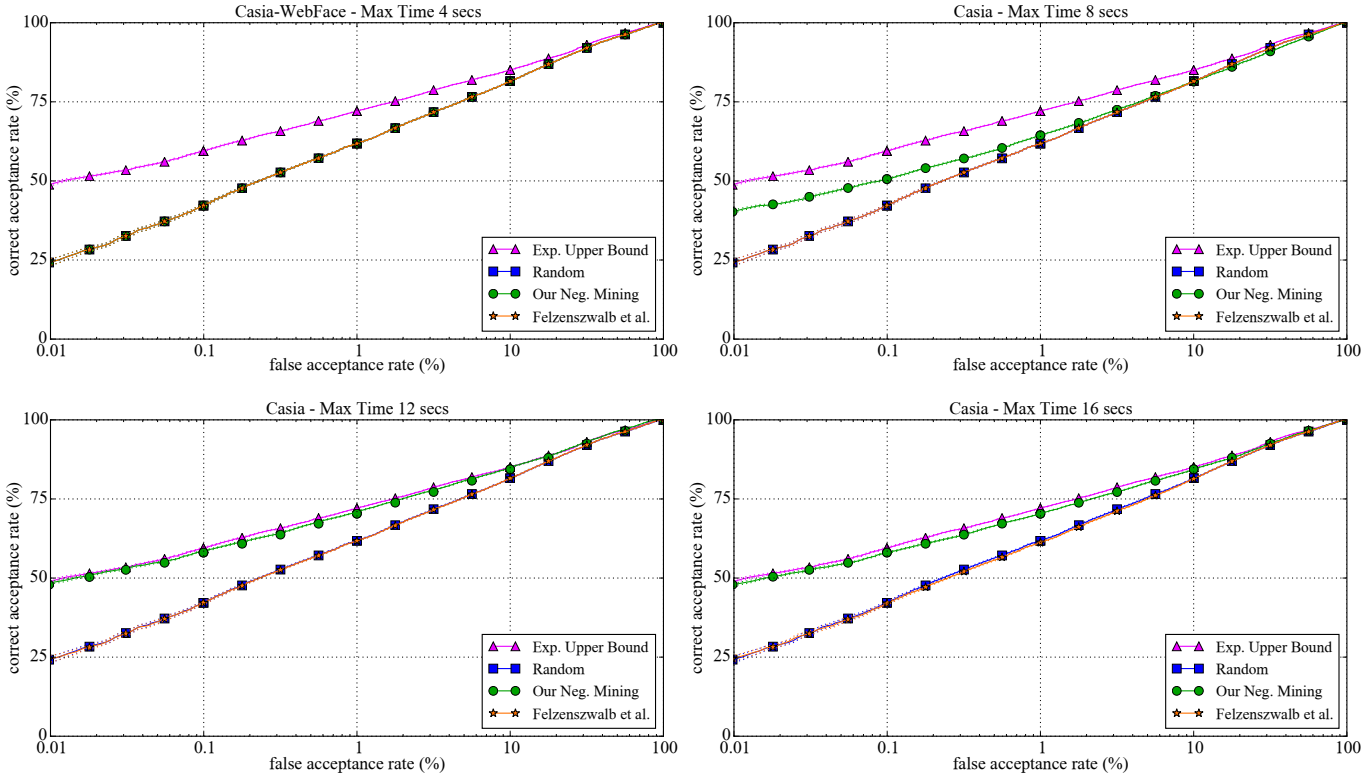



Fig. 5. System performances of the negative mining approaches in Casia-WebFace for the considered maximum processing times and for different points of FAR. Intervals correspond to standard errors.

compared to the Felzenszwalb et al.'s method, our negative mining is also preferred in both aspects, recognition and time performance. We believe that our mining criterion is more robust to critical negative samples, since we may also mine important correctly classified negative samples outside the SVM margin (see Fig. 3), while these samples are ignored in the mining process of [21]. Moreover, the processing time of each iteration from our method tends to be approximately constant, since the negative training set size is fixed. The numbers in bold in Table II show that, for both data sets, our method can achieve the *expected upper bound* recognition performance (Table I) without affecting the responsiveness of the system.

In Figure 5, we present a Receiver Operating Characteristic (ROC) curve (mean error values with standard errors) of the Casia-WebFace for each considered maximum processing

time. These curves illustrate the behavior of *random selection*, Felzenszwalb et al.'s method, and our negative mining approach at different operating points (as in Table II).

## VII. Conclusion

We have presented a first solution for the design of user-specific (US) classification models during user enrollment in a face recognition system that does not affect its speed and accuracy. The method can mine informative negative samples from a large data set, obtain US training sets and use them to build effective US classifiers within a few seconds. It is robust to the random choice of different inputs and has shown significantly superior performance with respect to five baselines.

Given that the algorithm is application-independent, we may conclude that it is a relevant contribution for biometric systems that aim to maintain robustness as the number of users increases. Our future work concentrates on new applications and suitable techniques to reduce huge data sets into representative large mining sets.

## Acknowledgment

The authors thank Samsung (under the terms of Brazilian federal law (8.248/91 20132015), CNPq (302970/2014-2), and FAPESP (2014/12236-1) for the financial support.

## References

[1] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[2] N. Pinto, Z. Stone, T. Zickler, and D. D. Cox, "Scaling-up biologically-inspired computer vision: A case-study on facebook," in *IEEE Computer Vision and Pattern Recognition (CVPR), Workshop on Biologically Consistent Vision*, 2011.

[3] G. Chiachia, A. X. Falcão, N. Pinto, A. Rocha, and D. Cox, "Learning person-specific representations from faces in the wild," *IEEE Trans. on Inf. Forens. Security (TIFS)*, vol. 9, no. 12, pp. 2089–2099, 2014.

[4] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1991.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection," *IEEE Trans. on Pattern Analysis Machine Intelligence (TPAMI)*, vol. 19, no. 7, pp. 711–720, 1997.

[6] G. P. Carlos, H. Pedrini, and W. R. Schwartz, "Fast and scalable enrollment for face identification based on partial least squares," in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition (FG)*, 2013.

[7] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *European Conf. on Computer Vision (ECCV), Workshop in Faces in Real-Life Images*, 2008.

[8] N. Pinto and D. Cox, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition (FG)*, 2011.

[9] M. Gunther, A. Costa-Pazo, C. Ding, E. Boutellaa, G. Chiachia *et al.*, "The 2013 face recognition evaluation in mobile environment," in *IEEE/IAPR Intl. Conf. on Biometrics (ICB)*, 2013.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[12] W. R. Schwartz, H. Guo, J. Choi, and L. S. Davis, "Face identification using large feature sets," *IEEE Trans. on Image Processing (TIP)*, vol. 21, no. 4, pp. 2245–2255, 2011.

[13] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2012.

[14] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[15] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. of Massachusetts Amherst, Tech. Rep. 07-49, 2007.

[17] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "The good, the bad, and the ugly face challenge problem," *Image and Vision Computing*, vol. 30, no. 3, pp. 177–185, 2012.

[18] V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang, *Handbook of Partial Least Squares*. Springer, 2010.

[19] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. on Pattern Analysis Machine Intelligence (TPAMI)*, vol. 35, no. 7, pp. 1757–1772, 2013.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[21] P. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. on Pattern Analysis Machine Intelligence (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[22] J. Valmadre, S. Sridharan, and S. Lucey, "Learning detectors quickly using structured covariance matrices," *CoRR*, vol. 1403.7321, 2014.

[23] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] W. S. Geisler and D. G. Albrecht, "Cortical neurons: Isolation of contrast gain control," *Vision Research*, vol. 32, no. 8, pp. 1409–1410, 1992.

[27] L. Bottou and C.-J. Lin, "Support vector machine solvers," in *Large Scale Kernel Machines*. MIT Press, 2007, pp. 1–28.

[28] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.