This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

Learning Person-Specific Representations from Faces in the Wild

Giovani Chiachia, Alexandre X. Falcão, *Member, IEEE,* Nicolas Pinto, Anderson Rocha, *Member, IEEE,* and David Cox, *Member, IEEE*

Abstract—Humans are natural face recognition experts, far out-performing current automated face recognition algorithms, especially in naturalistic, "in the wild" settings. However, a striking feature of human face recognition is that we are dramatically better at recognizing highly familiar faces, presumably because we can leverage large amounts of past experience with the appearance of an individual to aid future recognition. Meanwhile, the analogous situation in automated face recognition, where a large number of training examples of an individual are available, has been largely underexplored, in spite of the increasing relevance of this setting in the age of social media. Inspired by these observations, we propose to explicitly learn enhanced face representations on a *per-individual* basis, and we present two methods enabling this approach. By learning and operating within person-specific representations, we are able to significantly outperform the previous state-of-the-art on PubFig83, a challenging benchmark for familiar face recognition in the wild, using a novel method for learning representations in deep visual hierarchies. We suggest that such person-specific representations aid recognition by introducing an intermediate form of regularization to the problem.

Index Terms—Face Recognition, Face Information Modeling, Representation Learning, Deep Learning, Biologically-Inspired Computer Vision, Partial Least Squares, Support Vector Machines.

1 Introduction

The notion of creating a face "representation" tailored to the structure found in faces is a longstanding and foundational idea in automated face recognition research [1], [2], [3]. Indeed, a multitude of face recognition approaches employ an initial transformation into a *general* representation space before performing further processing [4], [5], [6], [7]. However, while the resulting face representation naturally captures structure found in common with all faces, much less attention has been paid to exploring the possibility of face representations constructed on a per individual basis.

Several observations motivate exploring the problem of *person-specific* face representations. First, intuitively, different facial features can be differentially distinctive across individuals. For instance, a given individual might have a distinctive nose, or a particular relationship between face features. Meanwhile, in realistic "in the wild" environments, these features might undergo significant variation due to changes in lighting, viewing angle, expression, occlusion, *etc.* Exploring feature extraction that is robust to these variations and tailored to specific individuals of interest is a potentially promising approach to tackling unconstrained face recognition.

- *Giovani Chiachia, Alexandre X. Falcão, and Anderson Rocha are with the Institute of Computing, University of Campinas (Unicamp), Campinas, Brazil. E-mails: {chiachia, afalcao, anderson.rocha}@ic.unicamp.br.*
- *Nicolas Pinto and David Cox are with Harvard University, Cambridge, USA. E-mails: pinto@alum.mit.edu and davidcox@fas.harvard.edu.*

In addition, the task of learning specialized representations in a per-individual basis has a natural relationship to the notion of "familiarity" in human face recognition, in that the brain may rely on enhanced face representations for familiar individuals [8], [9]. If we consider that humans are generally excellent at identifying familiar individuals even under uncontrolled viewing conditions [10], face familiarity is a specially relevant notion to pursue in the design of robust face recognition systems [11].

Finally, we argue that exploring this approach is especially timely today, as cameras become increasingly ubiquitous, recording an ever-growing torrent of image and video data. While to date much of face recognition research has focused on matching (*e.g.*, same/different) paradigms based on image pairs, the sheer volume of image data, in combination with user-driven cooperative face labeling, makes "familiar" face recognition increasingly relevant. One context where such an approach is especially attractive is in social media, where the problem is often to recognize an individual belonging to a limited, fixed gallery of possible friends, for whom many previous labeled training examples are frequently available. More generally, the ability to leverage a large number of past examples of specific individuals is a potential boon any time multiple examples of some finite number of persons of interest are available.

In Fig. 1, we present two distinct pipelines illustrating how our approach compares with methods most commonly found in the literature. As a first step, both pipelines (a) and (b) transform the input images into a feature set where the faces are described by the same, general attributes. Well-known techniques to derive this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

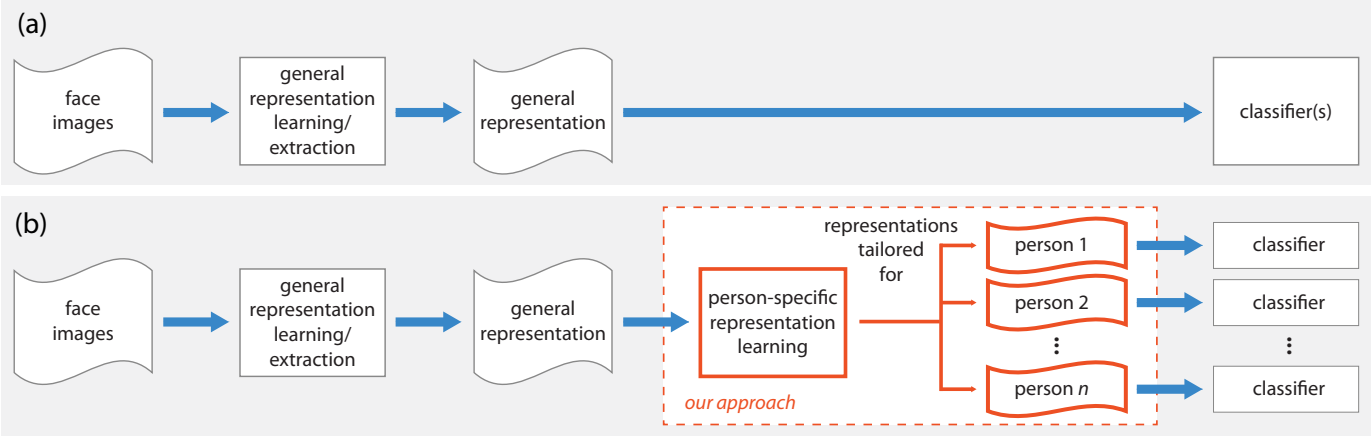

Fig. 1. Pipelines illustrating how methods can be regarded with respect to the employed face representation approach. Both pipelines (a) and (b) transform the input images into a feature set where the faces are described by the same, general attributes. Common techniques to derive this representation are Eigenface [1], Gabor wavelets [2], Local Binary Patterns [3], Fisherface [4], among others. On top of general face representations, methods following pipeline (a) directly perform learning tasks. In contrast, as presented in pipeline (b), our approach is to explicitly cast these general representations in person-specific ones by means of intermediate learning tasks that are based on domain knowledge, and are aimed at emphasizing the most discriminant face aspects of each individual.

representation are Eigenface [1], Gabor wavelets [2], Local Binary Patterns [3], Fisherface [4], Scale-Invariant Feature Transform [12], among others. On top of general face representations, face recognition methods following pipeline (a) directly perform learning tasks such as training one or multiple binary classifiers [13], [14], [15], learning similarity measures [6], [16], or learning sparse encodings [7]. In contrast, as presented in pipeline (b), our approach is to explicitly cast these general representations in person-specific ones by means of *intermediate* learning tasks that are based on domain knowledge, and are aimed at emphasizing the most discriminant face aspects of each individual.

From a machine learning perspective, we believe that these enhanced intermediate representations might alleviate the problem, allowing the subsequent classifiers to generalize better. From the perspective of psychological research, there are similarities between the proposed approach and the face recognition units (FRUs) of Bruce & Young's theoretical model of human face recognition [17], in which FRUs can be thought of person-specific units that retain visual structural information of familiar faces.

While several previous studies have used some form of person-specific representation in face recognition [18], [19], [20], the study of person-specific feature learning is still in its infancy. Here we validate the concept of person-specific face representations, and describe two approaches based on subspace learning and deep convolutional networks to building them from face images in the wild. Taken together, we argue that these techniques show that the person-specific representation learning approach holds great promise in advancing face recognition research, including psychological research in the vein of [21].

1.1 Contributions

We present and evaluate two methods for person-specific representation learning with the goal of validating the overall approach. Our methods and experiments consider the unconstrained, "in the wild" face recognition scenario that here is represented by the PubFig83 [15] dataset, described in Sec. 2.

The first method is presented in Sec. 3 and is based on a person-specific application of partial least squares (PS-PLS) to generate per-individual subspaces from any kind of visual representation in $\mathbb{R}^d$ [22]. A key motivating insight here is that a person-specific subspace, due to its supervised nature, can capture both aspects of the face that are good for discriminating it from others, as well as natural variation in appearance that is present in the unconstrained images of that individual. We show that operating in person-specific subspaces yields significant improvements in face recognition performance as compared to either "general" subspace learning approaches or classic supervised learning alone. Further, we show that such subspace method, when applied atop a deep convolutional neural network representation, can achieve recognition performance that exceeds the previous state-of-the-art on PubFig83.

Therefore, in Sec. 4 we introduce a second method to incorporate person-specific learning directly into a deep convolutional neural network. We demonstrate that, as long as we observe a few key principles in the network information flow, it is possible to learn discriminative filters at the topmost convolutional layer of the network with an approach based on SVMs. The inspiration to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

this approach comes from the assumption that class-specific transformations might be learned at the top of the human ventral visual stream hierarchy [23], and that neurons responding to specific faces might exist in the brain at even deeper stages [24]. We compare our method with other approaches and demonstrate that the proposed learning strategy produces an additional and significant performance boost on the PubFig83 dataset, for both *identification* and *verification* paradigms.

2 PubFig83: Dataset and Protocol

The PubFig83 dataset [15] is a subset of the PubFig dataset [25], which is, in turn, a large collection of real-world images of celebrities collected from the Internet. This subset is the result of a series of processing steps aimed at removing spurious face samples from PubFig, *i.e.*, non-detectable, near-duplicate, *etc.* In addition, only persons for whom 100 or more face images remained were considered, leading to a dataset with 83 subjects.

PubFig83 was established and released to promote research on familiar face recognition in the wild [15], in the setting where multiple training examples are available per individual (*e.g.*, in contrast to face matching protocols). To our knowledge, this is the only *publicly-available* face dataset with this many unconstrained, diverse images per individual, making it especially well-suited to the study of person-specific representations.

We aligned the images by the position of the eyes and followed the original protocol of [15], where the dataset was randomly split into ten pairs of training and test sets considering, for each individual, 90 images for training and 10 for test. In identification mode, performance is measured in the "closed-set" scenario and is reported in terms of accuracy, standing for the proportion of times that the system correctly predicts test images from the set of 83 individuals. In verification mode, performance is reported via receiver operating characteristic (ROC) curves and is measured both in the closed-set and in the "open-set" scenarios, where the later is devised by further splitting the dataset at random into 50 known and 33 unknown individuals.

In Fig. 2, we present images of three individuals in a given split of PubFig83. Here we show five (out of 90) training images and five (out of 10) test images of each. We can observe that this dataset presents many factors of variation in face appearance: aging, pose, illumination, expression, occlusion, hairstyle, among others. Extracting representations from these images in a way that such intrapersonal variation is alleviated, while extrapersonal variation is emphasized, is the foundational purpose of automatic face representation research [26]. Another challenging aspect of the dataset is that images are originally 100×100 pixels in size.

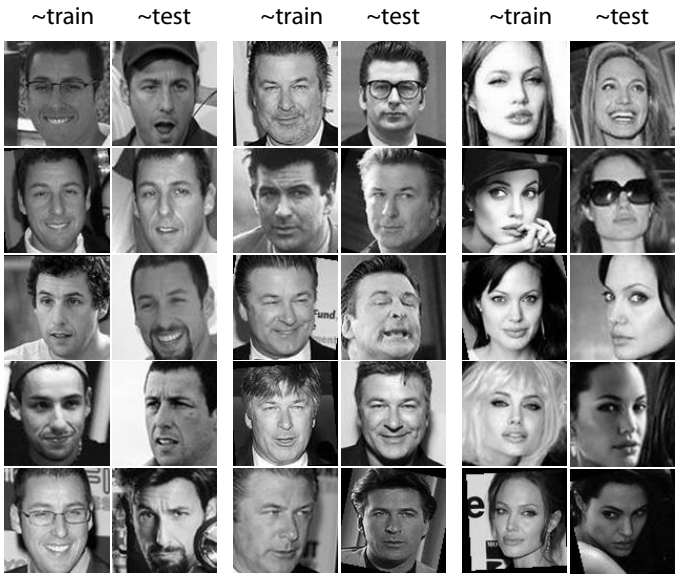

Fig. 2. Images of three individuals in a given split of PubFig83. Here we show five (out of 90) training images and five (out of 10) test images of each individual. The dataset presents many factors of variation in face appearance.

3 Person-Specific Subspace Analysis

The creation of subspaces tailored for faces is a classic technique in the face recognition literature; a variety of matrix-factorization techniques have been applied to faces (*e.g.*, Eigenface [1], Fisherface [4], Tensorface [5], *etc.*), which seek to model structure across a set of training faces, such that new face examples can be projected onto these spaces and can be compared. A principle advantage of projecting onto such subspaces is in the reduction of noise by limiting comparison to few relevant dimensions of variability in faces, as measured across a large number of images. However, while these methods naturally capture general structure across a set of faces, they typically discover either just structure that is common to reconstruct all faces (as in the case of Eigenface), or just structure that is common to discriminate all faces at the same time (as in the case of Fisherface).

In this section, we propose the use of a technique to build person-specific models on any kind of visual representation in $\mathbb{R}^d$. In particular, we build person-specific face subspaces from orthonormal projection vectors obtained by using a discriminative per-individual configuration of partial least squares [27], which we refer to as person-specific PLS or PS-PLS models. While partial least squares methods have been used in other contexts in face recognition before [28], [29], in the absence of a dataset that contains many examples per individual such as PubFig83, it is not possible for PLS methods to model natural variability in face appearance found in unconstrained images.

3.1 Partial Least Squares (PLS)

Partial least squares is a class of methods primarily designed to model relations between sets of observed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

variables by means of latent vectors [27], [30]. It can also be applied as a discriminant tool for the estimation of a low dimensional space that maximizes the separation between samples of different classes. PLS has been used in different areas [31], [32] and, recently, it is also being successfully applied to computer vision problems for dimensionality reduction, regression, and classification purposes [28], [29], [33], [34].

Given two matrices $\mathbf{X}$ and $\mathbf{Y}$ respectively with d and k mean-centered variables and both with n samples, PLS decomposes $\mathbf{X}$ and $\mathbf{Y}$ into

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad \text{and} \quad \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}, \tag{1}$$

where $\mathbf{T}_{n\times p}$ and $\mathbf{U}_{n\times p}$ are matrices containing the desired number p of latent vectors, matrices $\mathbf{P}_{d\times p}$ and $\mathbf{Q}_{k\times p}$ represent the loadings, and matrices $\mathbf{E}_{n\times d}$ and $\mathbf{F}_{n\times k}$ are the residuals.

One approach to perform the PLS decomposition employs the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [27], in which projection vectors $\mathbf{w}$ and $\mathbf{c}$ are determined iteratively such that

$$[cov(\mathbf{t},\mathbf{u})]^2 = \max_{||\mathbf{w}||=||\mathbf{c}||=1} [cov(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2, \tag{2}$$

where $cov(\mathbf{t},\mathbf{u})$ is the sample covariance between the latent vectors $\mathbf{t}$ and $\mathbf{u}$. In order to compute $\mathbf{w}$ and $\mathbf{c}$, given a random initialization of $\mathbf{u}$, the following steps are repeatedly executed [30]:

1) $\mathbf{u}_{old} = \mathbf{u}$
2) $\mathbf{w} = \mathbf{X}^T\mathbf{u}$
3) $||\mathbf{w}|| \rightarrow 1$
4) $\mathbf{t} = \mathbf{X}\mathbf{w}$
5) $\mathbf{c} = \mathbf{Y}^T\mathbf{t}$
6) $||\mathbf{c}|| \rightarrow 1$
7) $\mathbf{u} = \mathbf{Y}\mathbf{c}$
8) if $||\mathbf{u} - \mathbf{u}_{old}|| > \epsilon$, go to Step 1

When there is only one variable in $\mathbf{Y}$, *i.e.*, if $k = 1$, then $\mathbf{u}$ can be initialized as $\mathbf{u} = \mathbf{Y} = \mathbf{y}$. In this case, the steps above are executed only once per latent vector to be extracted [30]. The loadings are then computed by regressing $\mathbf{X}$ on $\mathbf{t}$ and $\mathbf{Y}$ on $\mathbf{u}$, *i.e.*,

$$\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^\mathbf{T}\mathbf{t}) \quad \text{and} \quad \mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^\mathbf{T}\mathbf{u}). \tag{3}$$

In this work, we use PLS to model the relations between face samples and their identities. The relationship between $\mathbf{X}$ and $\mathbf{Y}$ is then *asymmetric* and the *predicted* variables in $\mathbf{Y}$ are modeled as indicators. In the asymmetric case, after computing the latent vectors, matrices $\mathbf{X}$ and $\mathbf{Y}$ are deflated by subtracting their rank-one approximations based on $\mathbf{t}$, that is,

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad \text{and} \quad \mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{t}^T\mathbf{Y}/(\mathbf{t}^\mathbf{T}\mathbf{t}). \tag{4}$$

Such deflation rule ensures orthogonality among the latent vectors $\{\mathbf{t}_i\}_{i=1}^p$ extracted over the iterations. For details about the different types of PLS, their applicability to other problems, and how they compare with other techniques, we refer the reader to [30], [35].

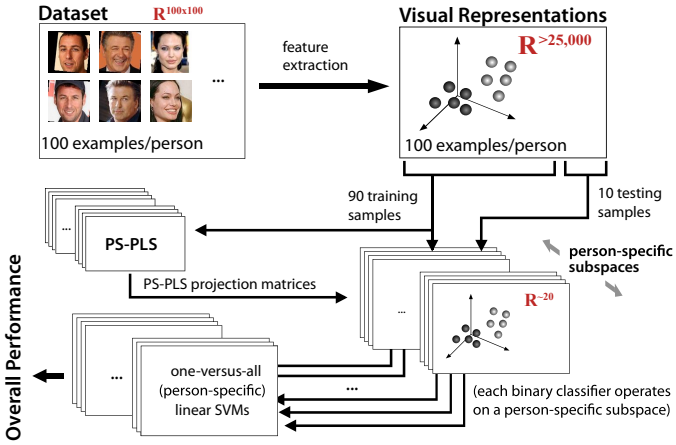

Fig. 3. From the training samples, PS-PLS creates a different face subspace for each individual. A different classifier is then trained in each subspace.

3.2 Person-Specific PLS

We learn face models with PLS for each person c at a time by setting $k = 1$, $\mathbf{Y}_{n\times k} = \mathbf{y}_c$, and $y_{c,s} = 1$ if sample s (out of n) belongs to class c or $y_{c,s} = 0$ otherwise. As $\mathbf{Y}$ has a single variable, this variant of PLS is also known as PLS1 [30]. It is worth recalling from the previous section that when $k = 1$, we can initialize $\mathbf{u} = \mathbf{y}_c$ and that, in this case, obtaining the projection vectors $\{\mathbf{w}\}_{i=1}^p$ is straightforward. In other words, at each iteration i,

$$\mathbf{w}_i = \mathbf{X}i^T\mathbf{y}_c, \tag{5}$$

where $\mathbf{X}i$ is the matrix $\mathbf{X}$ deflated up to iteration i according to Eq. 4.

The person-specific face model that we consider in this case is the subspace spanned by the set of orthonormal vectors $\{\mathbf{w}_i\}_{i=1}^p$ produced by NIPALS for a person c. Given that the variables in $\mathbf{X}$ are also normalized to unit variance, $\mathbf{w}_i$ expresses the relative importance of the face features (*i.e.*, the variables) to discriminate person c from the others. As $\{\mathbf{w}_i\}_{i=1}^p$ are orthogonal, this model accounts for within-person variance in the face appearance, a property also suggested to be relevant in mental representations of familiar faces [8].

In Fig. 3 we illustrate the approach. From the visual representation of the training samples, PS-PLS creates a different face subspace for each individual. All training samples are then projected onto each person-specific subspace, so that a classifier can be trained in each subspace. The classification engine that we use in our experiments is made by linear SVMs in a one-versus-all configuration, but it could be of any type provided it can operate in multiple representation spaces. Given a test sample, an overall decision is made according to decisions made in each person-specific subspace. In this work, we predict the face identity by choosing the person whose corresponding SVM scored highest.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

3.3 Experiments

As already mentioned, PS-PLS models can be learned from arbitrary $\mathbb{R}^d$ input spaces. Hence, we consider four different visual representations in order to evaluate them. The first visual representation that we take into account is based on non-overlapping histograms of LBP patches [3], a well-known baseline method for face description. The second and third representations are called V1-like+ and HT-L2-1st. They are taken from [15] and can be thought of as biologically-inspired visual models of increasing complexity. Finally, the fourth representation is similar in spirit to HT-L2-1st and consists of a three-layer visual hierarchy of the type found in [14]. We refer to this representation as L3+.

The main baseline for PS-PLS models consists of training linear SVMs straight from these visual representations, in which case we call the method RAW. In addition to comparing RAW and PS-PLS, we also consider subspace models obtained via principal component analysis (PCA), linear discriminant analysis (LDA), and Random Projection (RP) [36]. PCA is intuitively appealing in the context of face recognition and decomposes the training set in a way that most of the variance among the samples can be explained by a much smaller and ordered vector basis. LDA is another well-known technique that attempts to separate samples from different classes by means of projection vectors pointing to directions that decrease within-class variance while increasing the between-classes variance. As our PS-PLS setup seeks to maximize the separation only between-class, we argue that this offers a good compromise between LDA and PCA.

We further evaluate *person-specific* PCA models (PS-PCA) and multiclass PLS models with the idea that they would provide insight regarding the value of person-specific spaces. PS-PCA models are built only with the training samples of the person. For the multiclass PLS models, we assume k as the number of classes and make $\mathbf{Y}_{n\times k} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k\}$, with $y_{c,s} = 1$ if sample s belongs to class c or $y_{c,s} = 0$ otherwise. Still, in the inner loop of the NIPALS algorithm, each projection vector is considered after satisfying a convergence tolerance $\epsilon = 10^{-6}$ or after 30 iterations, whichever comes first. In this case, as $\mathbf{Y}$ has multiple variables, this form of PLS is also known as PLS2 [30]. While there remains substantial room to evaluate other subspace and manifold learning methods — including kernelized versions of PCA, LDA, and PLS — we chose here to focus on some of the most popular and straightforward methods available, with the goal of cleanly assessing the benefit of building person-specific subspaces.

The evaluation framework has two parameters: the regularization constant C of the linear SVMs, and the number of projection vectors to be considered, which is relevant in the cases where the projection vectors are ordered by their variance or discriminative power (PCA, PS-PCA, PLS, and PS-PLS). We use a separate *grid search* to estimate these parameters for each split. For this purpose, we re-split the training set so that we obtain 80 samples per class to generate *intermediate* models and 10 samples per class to *validate* them. We consider $\{10^{-3}, 10^{-2}, \ldots, 10^{5}\}$ as possible values to search for C. For the RAW and LDA models, this is the only parameter that we have to search, because, in the RAW case, no projection is made in practice and, in LDA, the number of projection vectors is fixed to the number of classes minus 1.

The possible number of projection vectors that we consider in the search can be represented as $\{1m, 2m, \ldots, 8m\}$. For person-specific subspace models, $m = 10$, *i.e.*, starting from 10, the number of projection vectors is increased by 10 up to the total number of data points per person in the *intermediate* training set. Correspondingly, for the multiclass models, $m = 10n$, where n is the number of persons in the dataset. The only exception is PLS, where $m = n$. Although PLS is a multiclass model, we observed that the ideal number of projection vectors is concentrated in the first few, and so we decided to refine the search accordingly, while keeping the same number of trials as for the other models. For all methods, the Scikit-learn package [37] was used to compute the subspace models and LIBSVM [38] was used to train the linear SVMs. In all cases, the data was scaled to zero mean and unit variance.

TABLE 1
Mean identification rates obtained with different subspace analysis techniques on the PubFig83 dataset. In all cases, the final identities are estimated by linear SVMs. In the last column, we present the most frequent number of projection vectors found by grid search.

model	LBP	V1-like+	HT-L2-1st	L3+	d ($\mathbb{R}^d$)
RAW	65.28±.52	74.81±.35	83.66±.55	88.18±.24	–
RP	61.77±.57	69.04±.44	79.92±.50	85.77±.26	6,640
multiclass unsupervised					
PCA	65.14±.48	74.59±.36	83.36±.47	87.86±.31	6,640
multiclass supervised					
LDA	59.01±.54	76.16±.50	81.14±.30	87.83±.39	–
PLS	63.88±.54	74.90±.45	83.07±.47	87.20±.31	332
person-specific					
PS-PCA	21.70±.58	29.95±.31	44.76±.45	54.58±.36	80
PS-PLS	**67.90±.58**	**77.59±.53**	**84.32±.38**	**89.06±.32**	20

3.4 Results

The results are shown in Table 1. In general, comparisons are done with the first row, where performance is assessed with the RAW visual representations. The remaining rows are divided according to the type of subspace analysis technique. It is possible to observe that the only face subspace in which we could consistently get better results than RAW across the different representations is PS-PLS.

For the RP and PCA subspace analysis techniques, we see no boost in performance above RAW. Since unconstrained face images have a considerable amount

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

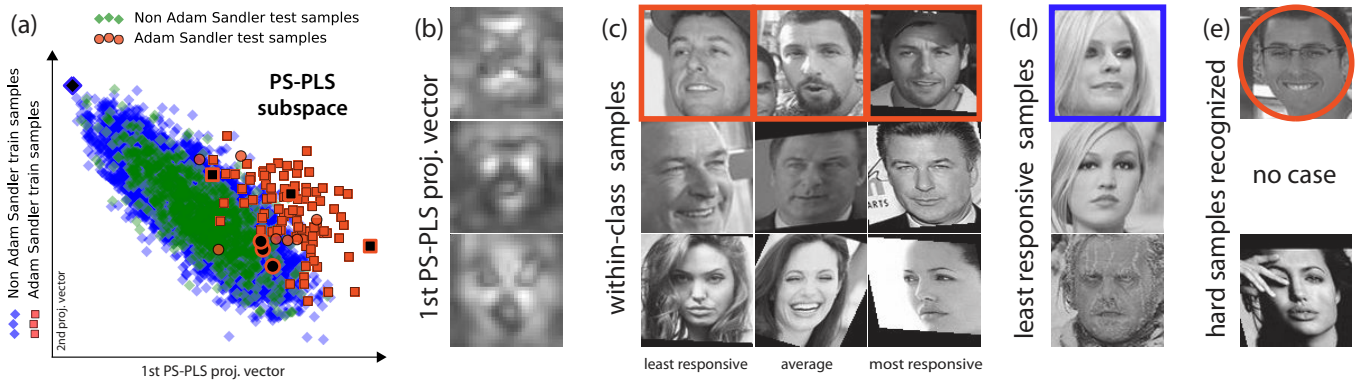

Fig. 4. (a) Scatter plot of training and test samples projected onto the first two PS-PLS projection vectors of Adam Sandler's subspace learned using V1-like+. (b) First PS-PLS projection vector for three individuals in the dataset. (c) Within-class most, on average, and least responsive face samples with respect to the projection onto (b). (d) Overall least responsive training sample w.r.t. (b). (e) Test samples correctly recognized when considering person-specific representations, but mistaken when using the RAW description. Samples in the first row of (b-e) are highlighted in (a).

of background clutter and these techniques do not regard its removal while estimating the models, this is perfectly reasonable. We observe that the visual representation on which the performance of RP dropped most is V1-like+, the largest in terms of input space dimensionality. Both for RP and PCA, the most frequent number of projection vectors found by grid search was 6,640, *i.e.*, the maximum allowed. This gives us the intuition that, operating with these unconstrained face images, the best that RP and PCA can do is to retain as much information in the input space as possible.

For the multiclass *supervised* techniques LDA and PLS, we observe performance increases only with LDA on the V1-like+ representation. While for HT-L2-1st and L3+ this may be simply the case of there being less room for improvement, we think that person-specific manifolds in the multiclass subspace are impaired by a more complex relation among the projection vectors. Since both PLS and PS-PLS follow the same rule for the estimation of the projection vectors, the results corroborate the idea that representing each individual in its own subspace results in better performance.

In the person-specific category, we see that PS-PCA considerably diminishes the predictive power of the features in the input space. In all cases, the best number of projection vectors found by grid search was 80, *i.e.*, the maximum allowed. When compared with PS-PLS, we can see here the importance of person-specific models being also discriminative, besides generative, for this task.

In Fig. 4(a), we present a scatter plot of training and test samples projected onto the first two PS-PLS projection vectors of Adam Sandler's subspace learned from V1-like+ representations.[1] Considering that the samples of Adam Sandler are in red, this plot illustrates one point that we observed throughout the experiments: that the predictive power of the first PS-PLS projection vectors is higher than that of the second one. Indeed, in PS-PLS, we found that the only projection vector that leads to mean projection responses significantly different between positive and negative samples is the first one. Although all subsequent projection vectors considerably increase performance, we believe that, from the second vector on, they progressively account more for person-specific variance than identity information. In our experiments, performance began to saturate around 20 projection vectors.

Fig. 4(b) is the result of mapping the importance of each V1-like+ feature back to the spatial domain, regarding their relative importance found by the first PS-PLS projection vector. Based on these illustrations, we can clearly see that different attributes are being weighted differently for each person, which goes against a rule for specifying attribute weights for faces in general.

Columns in Fig. 4(c) show the person-specific most, average, and least responsive face samples with respect to the projection onto the first PS-PLS projection vector. For Adam Sandler, these samples are highlighted in the plot in Fig. 4(a). It is difficult to infer anything concrete from these images, but we can see that the least responsive samples represent large variations in pose alignment and occlusion.

Still in Fig. 4, column (d) represents the overall least responsive training sample with respect to (b). These samples tend to be of the opposite gender, and hair seems to play a role for the first two individuals. Finally, in column (e) we present one test sample of each person that was not recognized when considering the RAW description of the faces, but that was recognized with the aid of PS-PLS models. Despite showing just one sample for Adam Sandler, there were three such cases, which are all highlighted in Fig. 4(a).

1. As PubFig83 is a dataset with celebrities, we use their names in this discussion. Also, we chose to use V1-like+ in this illustration because the relation of image pixels to the elements of its feature vector is more intuitive.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

In general, we argue that these subspaces are useful both for noise removal and for accentuating discriminative person-specific face aspects. Considering the results obtained with the RAW visual representations, we see that linear SVMs achieve reasonably high level of performance; however, when these same classifiers are trained and operate in PS-PLS subspaces, they perform better, suggesting that these 20-dimensional person-specific subspaces not only embed comparable levels of the available face identity information, but also amplify it.

4 Deep Person-Specific Filters

While person-specific subspace analysis is a promising general approach to learning person-specific representations from arbitrary underlying feature representations, the superior baseline performance of the L3+ visual representation in the previous section led us to explore whether the key theme of person-specific representation could be incorporated more integrally into that representation.

The L3+ representation [14] is based on the use of deep architectures for processing visual information. Such approach has a long tradition in machine learning literature [39], [40], [41], [42], and has been gaining attention due to recent breakthrough results in a number of important vision problems [14], [43], [44]. These techniques seek to mimic the neural computation of the brain in the hope of eventually reproducing its abilities in specific tasks. The basic architecture employs a hierarchical cascade of linear and nonlinear operations, applied in the framework of a generalized convolution.

Since the work of Hinton *et al.* [42], the strategy of greedily learning intermediate levels of representation as a *building block* to construct deep networks has been much discussed. While the focus has been put on unsupervised methods aimed at minimizing some kind of reconstruction error [45], [46], [47], little attention has been devoted to supervised *layer-wise* representation learning. This is possibly because discriminative learning strategies employed at early layers may prematurely discard information that would be critical to learn higher-level features about the target [46].

The work of Pinto *et al.* [14] is of considerable importance to unconstrained face recognition in general and to this work in particular. On the one hand, it achieves state-of-the-art performance in the challenging Labeled Faces in the Wild (LFW) benchmark [48]. On the other hand, it is the basis of our L3+ visual representation. In fact, this representation can be understood as the read-out of a three-layer convolutional neural network whose architecture was determined by performing a brute-force optimization of model hyper-parameters, while using random weights for the network's convolution filters.

Here, we ask if these underlying L3+ representations can be augmented by incorporating a person-specific learning process for setting their linear filter weights, resulting in an architecture that is both "deep" and person-specific. In order to construct these deep person-specific face models, we build on the idea of learning increasingly complex representations (*i.e.*, filter weights), one layer after the other. To be more precise, we are interested in learning person-specific models at the *top* layer of the L3+ network. We focus on the top layer not only because of the potentially disadvantages of discriminative filter learning at early layers but also for other two reasons: (i) the neuroscientific conjecture that class-specific neurons should exist in high levels of the human ventral visual stream hierarchy [23] and (ii) the experimental evidence suggesting that neurons responding to faces of specific individuals should exist in the brain at even deeper stages [24].[2]

4.1 L3+ Top Layer

Given that the top layer of the L3+ network is the object of our interest in the attempt to learn deep person-specific representations, in this section we briefly describe its architecture and operations according to [14]. As we can observe in the left panel of Fig. 5, the third and topmost layer of the L3+ network sequentially performs linear *filtering*, filter response *activation*, and local *pooling*.

The filtering operation takes a $34 \times 34 \times 128$ input from the previous layer corresponding to 128 feature maps and convolves it with filters Φ_i of size $5 \times 5 \times 128$ in order to create k higher level new feature maps

$$\mathbf{f}_i = \mathbf{x} \otimes \Phi_i \quad \forall i \in \{1, 2, \ldots, k\}, \tag{6}$$

where $\mathbf{x}$ is the input, $\otimes$ denotes the convolution operation, and $k = 256$ is the number of filters. The output of the filtering operation is then subjected to an activation function of the form

$$\mathbf{a}_i = \max(0, \mathbf{f}_i), \tag{7}$$

and these activations are, in turn, pooled together and spatially downsampled with a stride of 2 (downsampling factor of 4). In particular, the pooling and downsampling operation can be defined as

$$\mathbf{p}_i = \text{downsample}_2(\sqrt[10]{(\mathbf{a}_i)^{10} \otimes \mathbf{1}_{7\times7}}), \tag{8}$$

where $\mathbf{1}_{7\times7}$ is a 7×7 matrix of ones representing the pooling neighborhood. Note that the pooling operation is simply the L^{10}-norm of the activations in the pooling region, and can be regarded as an approximation of the soft-max pooling of [41]. Finally, after these three operations, the network outputs a visual representation of size $12 \times 12 \times 256$.

2. Another practical reason not to learn discriminative filters at early layers is spatial variance. Face misalignment is a serious problem in unconstrained face recognition that is significantly alleviated at higher levels of the network. For example, in L3+, each input *cell* in the third layer has a *receptive field* corresponding to a region of 65×65 pixels in the input image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

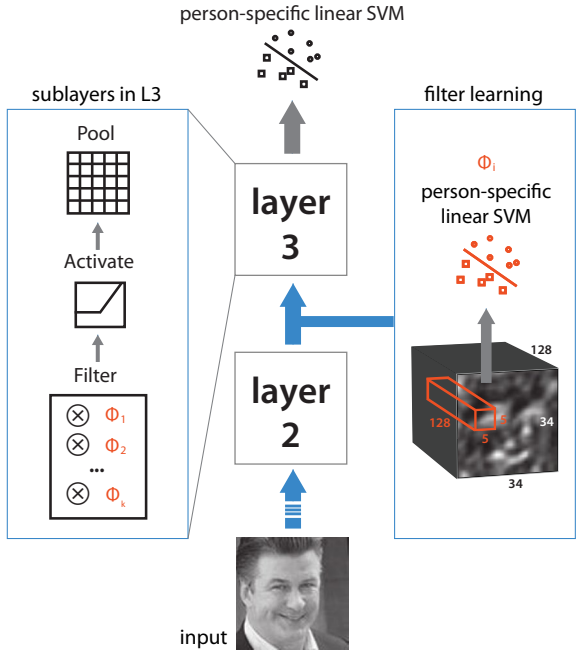

Fig. 5. Schematic diagram of the L3+ convolutional neural network, detailing the operations (sublayers) of its topmost layer (left panel) and illustrating how data from an input image is sampled in order to learn person-specific filters Φ_i at a given neighborhood (right panel). Early steps of the network [14] are omitted to emphasize the processing steps of our interest.

4.2 Proposed Approach

We propose an approach based on linear support vector machines (SVMs) to learn filters on the third layer of the L3+ representation. As we can see in the right panel of Fig. 5, an input image when transformed up to layer 2 is a feature vector $\mathbf{x}$ of size $34 \times 34 \times 128$. From a training set $\mathbf{X}$ with n samples, we are interested in learning $5 \times 5 \times 128$ filters Φ_i that later will be convolved with representations at the same input level. Given that these filters are meant to be person-specific, the type of SVM training that we carry out is one-versus-all, and assumes that filters are going to be learned by taking as input the same neighborhood $\mathcal{N}_i$ of 5×5 elements in space from all samples in $\mathbf{X}$. In Fig. 5, this means to consider features in the same red volume from all images, training an SVM with Alec Baldwin, for example, as the positive class and the other persons as the negative class. By doing so, a person-specific filter expected to be highly responsive to Alec Baldwin's face aspects in $\mathcal{N}_i$ is learned.

Let $\mathbf{X}_{\mathcal{N}_i}$ be the training set at neighborhood $\mathcal{N}_i$ and $\mathbf{y}_c$ be the labels for person c such that $y_{c,s} = +1$ if sample s (out of n) belongs to class c or $y_{c,s} = -1$ otherwise. A filter for c in $\mathcal{N}_i$ is simply the hyperplane Φ_i obtained with the solution of the linear support vector classification problem

$$\min_{\Phi_i, b_i} \frac{1}{2}||\Phi_i||_2 + C \sum_{s}^{n} \max\{0, 1 - y_{c,s}(\Phi_i \cdot \mathbf{x}_{s\mathcal{N}_i} + b_i)\}, \quad (9)$$

where C is the regularization constant that we set to 10^5 in order to obtain a parameter-free hard-margin method. In fact, the filter itself is the pair (Φ_i, b_i) with the intercept b_i ensuring that responses from different filters will be in the same range. For the sake of notation clarity we use only Φ_i to denote this pair.

It is possible to observe that a correspondence between filters $\mathbf{\Phi}$ and neighborhoods $\mathcal{N}$ exists, that is, both Φ_i and $\mathcal{N}_i$ have the same index i specifying from which region the filter is going to be learned. Indeed, there is an important fact in determining i that allows us to train *independent* filters. Recalling that the spatial resolution of the input samples at layer 2 is 34×34 and that filters are 5×5 in space, we can train $(34-5+1)^2 = 900$ such filters. However, we observe that the correlation between filters trained from adjacent regions is undesirably high, and so there is no benefit in considering them all. This is not the case though if we subsample possible neighborhoods with a stride of 3, in which situation the mean correlation among the filters is close to zero. Therefore, the proposed procedure to learn a third person-specific layer in the L3+ hierarchy considers $(\lfloor \frac{34-5}{3} \rfloor + 1)^2 = 100$ filters $\mathbf{\Phi}$.[3]

The final component that we add to our filter learning approach is inspired by an observation about the information flow in the network when operating with random filters.[4] Provided that after drawing the weights from a uniform distribution the filters are mean centered, and given the activation function in Eq. 7, we observed that, on average, half of the random filter responses are set to zero after activation. The enforcement of such "calibrated" sparsity in the *random* case showed to be quite relevant to the network performance in our tests. Therefore, we replicate this behavior in our method by assuming α as the mean response of all person-specific filters on the training set and using an activation function of the form

$$\mathbf{a}_i = \max(0, \mathbf{f}_i - \alpha) \quad (10)$$

instead. Without this shift on activation we found that SVM filters are too selective, *i.e.*, almost all filter responses are set to zero if we rather use Eq. 7.

The observance of the two aforementioned properties of (i) independence and (ii) calibrated sparsity in our learning framework allows the network to represent well face images even of other individuals. No matter which stimuli these person-specific filters are trained to respond best, these properties naturally enable them to be as informative as random filters are. However,

3. Although the number of filters was empirically determined in this study, this number can be seen as a hyperparameter to be adjusted on other problem-domains.

4. As random filters are known to perform surprisingly well in the general class of convolutional neural networks [49], [50], we found valuable to investigate some of their characteristics.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

we expect that when these filters operate in images of the persons whose face aspects they were trained to discriminate, they might significantly increase the ability of the system at recognizing these persons.

Even though the proposed approach is tailored to the deep architecture of our interest and designed to strengthen our hypothesis in the context of person-specific face representation learning, the method seems to extend naturally to other object recognition problems. To our knowledge, this is the first attempt to learn "stackable" layer-wise representations with maximum-margin classifiers. Given the large amount of variation that unconstrained images have (*e.g.*, Fig. 4), even large-scale datasets such as PubFig83 — with thousands of training images — require methods with strong generalization abilities. The idea of piecing together maximum-margin filters in convolutional networks is potentially relevant in this concern.

4.3 Experiments and Results

The experiments that we carry out in order to evaluate our approach consist of clamping both the architecture and filter weights of L3+ up to layer 2 and varying two aspects of its third layer while we measure performance in the PubFig83 dataset. The first aspect is the *filter type*, *i.e.*, how filters are determined, and the second aspect is the *number of filters*.

The obvious baseline with respect to the filter type is the use of random filters, which are used in the standard L3+ visual representation from Sec. 3. We also consider filters of the type proposed by Coates and Ng in [51], whose use in large quantities corroborates the notion that good performance can be achieved with inexpensive filter quantization and encoding techniques [51]. We evaluate their K-means-like method that takes normalized and ZCA whitened patches as input and computes filters using dot-products as the similarity metric rather than the Euclidean distance [51].

In order to compare our approach with competitive configurations of these methods, we scale the number of filters in the third layer up to as many as 2,048, and we vary this number as an experimental parameter. Both for random as well as for K-means-like filters, we assess performance with $k = \{100, 256, 512, 1024, 2048\}$ filters. In the person-specific case, we measure performance with pure $k = 100$ person-specific filters, but we also concatenate them with filters of the two other types, so that the overall number of filters matches the other cases. This gives rise to methods that we call person-specific (PS)+random and PS+K-means-like, that are evaluated with $k = 100 + \{156, 412, 924, 1948\}$ filters.

In addition to random and K-means-like filters, we made a substantial effort to compare our approach with filters trained via backpropagation. However, we found that in this case, even considering a small number of filters, the network rapidly overfits to the training samples, resulting in poor performance on the test set. Considering both the third (convolutional) and the fourth (fully-connected) layers, such network has almost four million parameters when trained with $k = 256$ filters. We believe that the availability of only $n = 7,470$ training samples in PubFig83 did not allow us to obtain good levels of performance in this attempt.

TABLE 2
Comparisons in identification mode. A clear boost in performance can be observed with the use of person-specific filters. It can also be seen that person-specific filters combine well with the other two types. In particular, when combined with 1948 K-means-like filters, the method achieves the best result on PubFig83 to our knowledge.

number of filters	filter type				
	random	K-means like	person-specific-	person-specific+ random	person-specific+ K-means
100	85.38±.26	83.99±.42	**90.62±.27**	–	–
256	88.18±.24	87.69±.29	–	91.43±.24	91.37±.32
512	88.76±.32	89.43±.28	–	91.60±.29	91.87±.23
1024	89.26±.26	90.46±.37	–	91.67±.25	92.17±.27
2048	89.40±.31	91.29±.34	–	91.07±.27	**92.28±.26**

Regardless the filter type and the number of filters, all other operations and architectural parameters in the third layer are preserved. The only exception is the activation function, where Eq. 7 is replaced by Eq. 10 in cases where the filters are learned, *i.e.*, when using person-specific and K-means-like filters.[5] In these cases, α is determined as explained in the previous section. Still concerning filter learning issues, we sample exactly the same patches in both cases; each person-specific filter (out of 100) is learned from a set of n patches, and all K-means-like filters are learned from a training set with the same $100n$ patches. Finally, as in Sec. 3, person-specific linear SVMs are trained on top of all the resulting visual representations.

The experimental results are presented in Table 2 and Figs. 6 and 7 for the methods in identification mode and in Fig. 8 in verification mode. In accordance to all results presented in this paper, we report the mean performance and standard error over ten dataset splits for both modes (see Sec. 2).

From Table 2 and Fig. 6, we clearly see a dramatic boost in performance when considering person-specific filters, especially if we take into account correspondence in the number of filters. Comparable levels of performance with 100 person-specific filters is not achieved even with 2,048 random filters and is only achieved with more than 1,024 K-means-like filters. In addition, we see that person-specific filters combine well with the other two types. In particular, when combined with 1,948 K-means-like filters, the method achieves a mean accuracy of 92.28%, the best result on PubFig83 to our knowledge.

As expected, there is a clear relationship between the number of filters and performance in both random and

5. As advocated in [51], this is in fact a very good encoding scheme to use with large quantities of K-means-like filters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

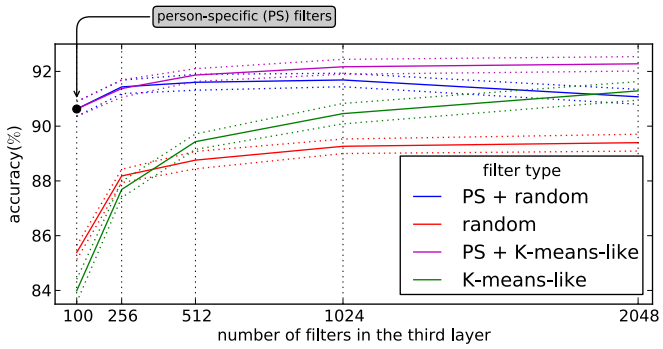

Fig. 6. Plot of the results obtained with Deep Person-Specific Filters in identification mode. Intervals correspond to standard errors.

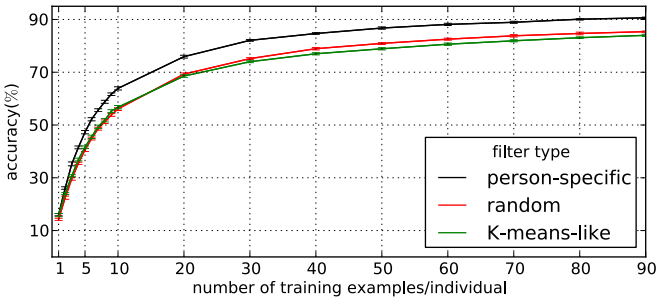

Fig. 7. Performance of each filter type as a function of the number of training images per individual. As the number of images increases, person-specific filters can progressively leverage this information to the point where its advantage becomes clear. The consistent difference in performance observed as the accuracy approaches 90% suggests that the method can steadily aid recognition even at high performance levels.

K-means-like cases. Interestingly, K-means-like performs worse than random with small numbers of filters but achieves much better performance when this number increases. While this may contradict the argument that filter learning becomes crucial with the decrease in filter quantity [51], it appears to corroborate the observation that large numbers of K-means-like filters might span the space of inputs more equitably, which increases the chances that a few filters will be near the input and leads to a few but high activations [51].

In Fig. 7, we present the performance of each filter type as a function of the number of training images per individual. It is possible to observe that person-specific filters do not benefit from learning representations from a single image, and this may be associated with the lack of intrapersonal information from where to extract discriminative face aspects. However, as the number of training images increases, the proposed method can progressively leverage this information to the point where the advantage of person-specific representations becomes clear. The consistent difference in performance observed as the accuracy approaches 90% suggests that the method can steadily aid recognition even at high performance levels.

For comparison purposes, in Table 3 we present other face identification results on PubFig83 available in the literature. While we make a distinction between results using *aligned* images and results using *unaligned* images, the work of Pinto *et al.* [15] gives us an idea about how these setups compare.

Unfortunately, natural comparisons to other top performing methods "in the wild" — especially those evaluated on LFW dataset [48] (*e.g.*, [16], [52]) — are difficult, due to the fundamentally different nature of LFW and PubFig83 protocols. Specifically, LFW is configured for pair-matching, while PubFig83 allows training with a significant number of per-individual examples. In many cases, top performing algorithms on the LFW set are tightly coupled to the pair-matching setting, and it is unclear how they should be adapted to PubFig83.

In Fig. 8, we present the performance of the methods

TABLE 3
Identification results on PubFig83 available in the literature.

images	Pinto *et al.* [15] CVPRW'11	Chiachia *et al.* [22] BMVC'12	Bergstra *et al.* [53] ICML'13	Carlos *et al.* [54] FG'13	This paper
unaligned	85.22±.45	–	86.50±.70	–	–
aligned	87.11±.56	88.75±.26	–	73.47±.41	**92.28±.28**

in verification mode, where the task is to decide whether or not a given test face belongs to a claimed identity. Given that such pair matching is done with the use of only one person-specific model with 100 filters, we considered reasonable to compare methods only with this number of filters. In addition to measuring performance in the closed-set scenario, here we also consider the open-set scenario where some test individuals are not contained in the training set. Precisely, for each split of the dataset, we randomly selected 50 individuals (out of 83) of which person-specific representations are learned from the training images. While the training images of the other 33 individuals are ignored, test images of all the 83 individuals are considered at testing time. Therefore, in each split, the system operates with a distinct set of 50 known and 33 unknown identities.

This open-set evaluation is important not only for practical reasons, but also because humans are constantly exposed to unfamiliar faces and do not mistake them with familiar faces. Therefore, it is desirable that the learned representations also perform well in open-set case.

It is possible to observe in Fig. 8 that the use of person-specific filters causes a significant gain in performance as compared to the use of random or K-means-like filters. Even though there exists a natural drop in performance between the closed- (83) and the open-set (50+33) scenarios, the advantages of the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2014.2359543, IEEE Transactions on Information Forensics and Security

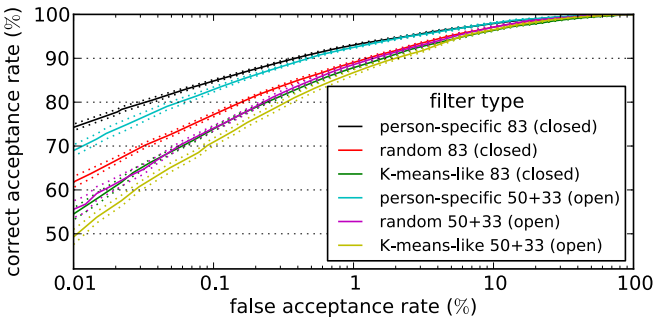

Fig. 8. Comparisons with Deep Person-Specific Filters in closed-set (83) and open-set (50+33) verification regimes. When the system is set to wrongly accept only 0.01% of the test cases, we can observe a dramatic improvement in performance, which is especially relevant to high security applications.

approach can be perceived in both settings. Moreover, when the system is set to wrongly accept only 0.01% of the test cases, the use of person-specific filters results in a great improvement in correct acceptance. In high security applications, this difference is of extreme relevance, suggesting that the approach of learning person-specific representation is not only conceptually relevant — as we observed throughout the paper — but also readily applicable in the verification scenario. Moreover, given that these results could possibly be improved with the use custom-tailored open-set classifiers [55], [56], open-set person-specific face verification is certainly a topic of interest for future work.

5 Conclusions

In this paper, we presented two techniques, based on different learning principles, to explicitly and progressively build on the idea that generating person-specific representations can boost face recognition performance.

We motivated the idea as an attempt to model two different attributes of human face perception, and conducted interrelated experiments on a dataset of faces in the wild, achieving not only insight into the value of person-specific representation, but also state-of-the-art results.

We first proposed the use of person-specific subspaces to leverage any kind of input visual representation in $\mathbb{R}^d$. We believe that this approach represents a first step towards the incorporation of the notion of face "familiarity" into face recognition systems — a notion that is known to be of key importance in biological vision. In addition, we introduced an original framework that uses SVMs to learn "deep" person-specific filters in a convolutional neural network, again achieving superior recognition performance.

With the consistent improvements that we observed throughout the experiments in both face identification and face verification tasks, we showed that the use of intermediate, person-specific representation has the power to boost recognition performance beyond what either generic face representation learning, or traditional supervised learning can achieve alone.

While any sort of supervised learning might arguably be considered a form of "person-specific" representation, here we have found that the inclusion of intermediate, problem-driven person-specific representation learning steps lead to significant boosts in performance. One possible explanation for this phenomenon is that such representations introduce an intermediate form of regularization to the face recognition problem, allowing the classifiers to generalize better by enforcing them to use less but more relevant features. Exploring this hypothesis, and continuing to explore the wide range of possible techniques for learning person-specific representations from faces in the wild will be a promising area for future research.

Acknowledgments

The authors would like to thank FAPESP (2010/00994-8, 2010/05647-4, 2013/11359-0), CNPq (303673/2010-9, 304352/2012-8, 477662/2013-7, 479070/2013-0), Microsoft Research, and The Rowland Institute at Harvard. They also thank the reviewers for insightful comments and suggestions to improve an earlier draft of this paper.

References

[1] M. Turk and A. Pentland, "Face Recognition using Eigenfaces," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1991.
[2] L. Wiskott, J.-M. Fellous, N. Krger, and C. V. D. Malsburg, "Face Recognition By Elastic Bunch Graph Matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, 1997.
[3] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.
[4] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
[5] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear Image Analysis for Facial Recognition," in *IEEE Intl. Conf. on Pattern Recognition*, 2002.
[6] L. Wolf, T. Hassner, and Y. Taigman, "The One-Shot Similarity Kernel," in *IEEE Intl. Conf. on Computer Vision*, 2009.
[7] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
[8] A. M. Burton, R. Jenkins, and S. R. Schweinberger, "Mental Representations of Familiar Faces," *British Journal of Psychology*, vol. 102, no. 4, pp. 943–58, 2011.
[9] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.
[10] A. M. Burton, S. Wilson, M. Cowan, and V. Bruce, "Face Recognition in Poor-Quality Video: Evidence From Security Surveillance," *Psych. Science*, vol. 10, no. 3, pp. 243–248, 1999.
[11] R. Chellappa, P. Sinha, and P. Phillips, "Face Recognition by Computers and Humans," *IEEE Computer*, vol. 43, no. 2, pp. 46–55, 2010.
[12] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," in *IEEE Intl. Conf. on Computer Vision*, 1999.

[13] B. Heisele, P. Ho, and T. Poggio, "Face Recognition with Support Vector Machines: Global versus Component-based Approach," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[14] N. Pinto and D. D. Cox, "Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition," in *IEEE Conf. on Automatic Face and Gesture Recognition*, 2011.

[15] N. Pinto, Z. Stone, T. Zickler, and D. D. Cox, "Scaling-up Biologically-Inspired Computer Vision: A Case Study in Unconstrained Face Recognition on Facebook," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2011.

[16] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[17] V. Bruce and A. Young, "Understanding Face Recognition." *British Journal of Psychology*, vol. 77, pp. 305–327, 1986.

[18] S. Krishna, J. Black, and S. Panchanathan, "Using Genetic Algorithms to Find Person-Specific Gabor Feature Detectors for Face Indexing and Recognition," in *IAPR Intl. Conf. on Biometrics*. Springer, 2005, pp. 182–191.

[19] S. Zafeiriou, A. Tefas, and I. Pitas, "Learning Discriminant Person-Specific Facial Models using Expandable Graphs," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 1, pp. 55–68, 2007.

[20] J. Sivic, M. Everingham, and A. Zisserman, ""Who are you?": Learning Person Specific Classifiers from Video"," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[21] A. Burton, R. Jenkins, P. Hancock, and D. White, "Robust Representations for Face Recognition: The Power of Averages," *Cognitive Psychology*, 2005.

[22] G. Chiachia, N. Pinto, W. R. Schwartz, A. Rocha, A. X. Falcão, and D. Cox, "Person-Specific Subspace Analysis for Unconstrained Familiar Face Identification," in *British Machine Vision Conference*, 2012.

[23] T. Poggio, "The Computational Magic of the Ventral Stream," *Nature Precedings*, 2011, doi:10.1038/npre.2012.6117.3.

[24] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant Visual Representation by Single Neurons in the Human Brain," *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

[25] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *IEEE Intl. Conf. on Computer Vision*, 2009.

[26] J.-K. Kamarainen, A. Hadid, , and M. Pietikainen, "Local Representation of Facial Features," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds. Springer, 2011, ch. 4, pp. 79–108.

[27] H. Wold, "Partial Least Squares," *Wiley Encyclopedia of Statistical Sciences*, vol. 6, pp. 581–591, 1985.

[28] W. R. Schwartz, H. Guo, J. Choi, and L. S. Davis, "Face Identification Using Large Feature Sets," *IEEE Trans. on Image Processing*, vol. 21, no. 4, pp. 2245–2255, 2011.

[29] H. Guo, W. R. Schwartz, and L. S. Davis, "Face Verification using Large Feature Sets and One Shot Similarity," in *IEEE Intl. Joint Conf. on Biometrics*, 2011.

[30] R. Rosipal and N. Kramer, "Overview and Recent Advances in Partial Least Squares," *Springer LNCS: Subspace, Latent Structure and Feature Selection Techniques*, pp. 34–51, 2006.

[31] F. Lindgren, P. Geladi, A. Berglund, M. Sjostrom, and S. Wold, "Interactive Variable Selection (IVS) for PLS. Part II: Chemical Applications," *Wiley Journal of Chemometrics*, vol. 9, no. 5, pp. 331–342, 1995.

[32] D. V. Nguyen and D. M. Rocke, "Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data," *Oxford Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[33] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human Detection Using Partial Least Squares Analysis," in *IEEE Intl. Conf. on Computer Vision*, 2009.

[34] A. Kembhavi, D. Harwood, and L. Davis, "Vehicle Detection Using Partial Least Squares," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1250–1265, 2011.

[35] H. Abdi, "Partial Least Squares Regression and Projection on Latent Structure Regression," *Wiley Int. Reviews: Computational Statistics*, vol. 2, no. 4, 2010.

[36] E. Bingham and H. Mannila, "Random Projection in Dimensionality Reduction: Applications to Image and Text Data," in *ACM Conf. on Knowledge Discovery and Data Mining*, 2001.

[37] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011.

[38] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[39] K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition unaffected by Shift in Position," *Springer Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[40] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *MIT Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[41] M. Riesenhuber and T. Poggio, "Hierarchical Models of Object Recognition in Cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999.

[42] G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *MIT Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[44] Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, "Building High-level Features Using Large Scale Unsupervised Learning," in *Intl. Conf. on Machine Learning*, 2012.

[45] M. A. Ranzato, F. J. Huang, Y. lan Boureau, and Y. Lecun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[46] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-wise Training of Deep Networks," in *Advances in Neural Information Processing Systems*, 2007.

[47] Q. V. Le, J. Ngiam, Z. Chen, D. Chia, P. W. Koh, and A. Y. Ng, "Tiled Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2010.

[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild," Univ. of Massachusetts, Amherst, Tech. Rep., 2007.

[49] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the Best Multi-Stage Architecture for Object Recognition?" in *IEEE Intl. Conf. on Computer Vision*, 2009.

[50] A. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Ng, "On Random Weights and Unsupervised Feature Learning," in *Intl. Conf. on Machine Learning*, 2011.

[51] A. Coates and A. Ng, "The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization," in *Intl. Conf. on Machine Learning*, 2011.

[52] 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*, 2013.

[53] J. Bergstra, D. Yamins, and D. D. Cox, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," in *Intl. Conf. on Machine Learning*, 2013.

[54] G. P. Carlos, H. Pedrini, and W. R. Schwartz, "Fast and Scalable Enrollment for Face Identification based on Partial Least Squares," in *IEEE Conf. on Automatic Face and Gesture Recognition*, 2013.

[55] F. de O. Costa, E. Silva, M. Eckmann, W. Scheirer, and A. Rocha, "Open set source camera attribution and device linking," *Elsevier Pattern Recognition Letters*, vol. 39, pp. 91–101, 2014.

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