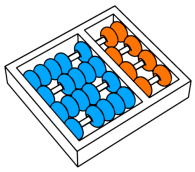
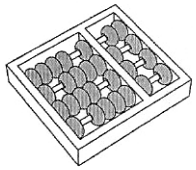


Giovani Chiachia

"Learning Person-Specific Face Representations"

"*Aprendendo Representações Específicas para a Face de cada Pessoa*"

CAMPINAS
2013



University of Campinas
Institute of Computing

Universidade Estadual de Campinas
Instituto de Computação

Giovani Chiachia

"Learning Person-Specific Face Representations"

Supervisor:
Orientador(a): Prof. Dr. Alexandre Xavier Falcão

Co-Supervisor:
Co-orientador(a): Prof. Dr. Anderson Rocha

"Aprendendo Representações Específicas para a Face de cada Pessoa"

Ph.D. Thesis presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a Doctor degree in Computer Science.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Doutor em Ciência da Computação.

This volume corresponds to the final version of the Thesis defended by Giovani Chiachia, under the supervision of Prof. Dr. Alexandre Xavier Falcão.

Este exemplar corresponde à versão final da Tese defendida por Giovani Chiachia, sob orientação de Prof. Dr. Alexandre Xavier Falcão.

Supervisor's signature / *Assinatura do Orientador(a)*

CAMPINAS
2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

C43L

Chiachia, Giovani, 1981-
Learning person-specific face representations / Giovani Chiachia. – Campinas, SP : [s.n.], 2013.

Orientador: Alexandre Xavier Falcão.
Coorientador: Anderson de Rezende Rocha.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Identificação biométrica. 2. Reconhecimento facial (Computação). 3. Visão por computador. 4. Aprendizado de máquina. I. Falcão, Alexandre Xavier,1966-. II. Rocha, Anderson de Rezende,1980-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Aprendendo representações específicas para a face de cada pessoa
Palavras-chave em inglês:
Biometric identification
Human face recognition (Computer science)
Computer vision
Machine learning
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Alexandre Xavier Falcão [Orientador]
Hélio Pedrini
Eduardo Alves do Valle Junior
Zhao Liang
Walter Jerome Scheirer
Data de defesa: 27-08-2013
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Tese Defendida e Aprovada em 27 de Agosto de 2013, pela Banca examinadora composta pelos Professores Doutores:

Prof. Dr. Walter Jerome Scheirer
DMCB / HARVARD UNIVERSITY

Prof. Dr. Eduardo Alves do Valle Junior
FEEC / UNICAMP

Prof. Dr. Zhao Liang
ICMC / USP

Prof. Dr. Hélio Pedrini
IC / UNICAMP

Prof. Dr. Alexandre Xavier Falcão
IC / UNICAMP

Institute of Computing /*Instituto de Computação*
University of Campinas /*Universidade Estadual de Campinas*

Learning Person-Specific Face Representations

Giovani Chiachia[1]

August 27, 2013

Examiner Board/*Banca Examinadora*:

- Prof. Dr. Alexandre Xavier Falcão (Supervisor)
- Prof. Dr. Hélio Pedrini
  Institute of Computing - UNICAMP
- Prof. Dr. Eduardo Alves do Valle Junior
  School of Electrical and Computer Engineering - UNICAMP
- Prof. Dr. Zhao Liang
  Institute of Mathematics and Computer Science - USP
- Prof. Dr. Walter Jerome Scheirer
  School of Engineering and Applied Sciences - Harvard University

[1]Financial support: FAPESP scholarship (2010/00994-8) 2009–2013

Abstract

Humans are natural face recognition experts, far outperforming 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. Researchers in psychology have even suggested that face representations might be partially tailored or optimized for familiar faces. 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 a collection of methods enabling this approach and progressively justifying our claim. By learning and operating within person-specific representations of faces, we are able to consistently improve performance on both the constrained and the unconstrained face recognition scenarios. In particular, we achieve state-of-the-art performance on the challenging PubFig83 familiar face recognition benchmark. We suggest that such person-specific representations introduce an intermediate form of regularization to the problem, allowing the classifiers to generalize better through the use of fewer — but more relevant — face features.

Resumo

Os seres humanos são especialistas natos em reconhecimento de faces, com habilidades que excedem em muito as dos métodos automatizados vigentes, especialmente em cenários não controlados, onde não há a necessidade de colaboração por parte do indivíduo sendo reconhecido. No entanto, uma característica marcante do reconhecimento de face humano é que nós somos substancialmente melhores no reconhecimento de faces familiares, provavelmente porque somos capazes de consolidar uma grande quantidade de experiência prévia com a aparência de um certo indivíduo e de fazer uso efetivo dessa experiência para nos ajudar no reconhecimento futuro. De fato, pesquisadores em psicologia têm até mesmo sugerido que a representação interna que fazemos das faces pode ser parcialmente adaptada ou otimizada para rostos familiares. Enquanto isso, a situação análoga no reconhecimento facial automatizado — onde um grande número de exemplos de treinamento de um indivíduo estão disponíveis — tem sido muito pouco explorada, apesar da crescente relevância dessa abordagem na era das mídias sociais. Inspirados nessas observações, nesta tese propomos uma abordagem em que a representação da face de cada pessoa é explicitamente adaptada e realçada com o intuito de reconhecê-la melhor. Apresentamos uma coleção de métodos de aprendizado que endereça e progressivamente justifica tal abordagem. Ao aprender e operar com representações específicas para face de cada pessoa, nós somos capazes de consistentemente melhorar o poder de reconhecimento dos nossos algoritmos. Em particular, nós obtemos resultados no estado da arte na base de dados PubFig83, uma desafiadora coleção de imagens instituída e tornada pública com o objetivo de promover o estudo do reconhecimento de faces familiares. Nós sugerimos que o aprendizado de representações específicas para face de cada pessoa introduz uma forma intermediária de regularização ao problema de aprendizado, permitindo que os classificadores generalizem melhor através do uso de menos — porém mais relevantes — características faciais.

À minha esposa, Thais Regina de Souza Chiachia.

Agradecimentos

Após anos de trabalho, muitas são as pessoas às quais eu gostaria de expressar meus sinceros agradecimentos.

Primeiramente, eu gostaria de agradecer aos professores doutores Alexandre Xavier Falcão e Anderson Rocha pela orientação que me deram durante o desenvolvimento deste trabalho. Não foram poucas as ocasiões que demandaram suporte, reflexões e mudanças de rumo, e eles administraram isso com sabedoria, dando-me autonomia na medida certa para que hoje, de fato, eu sinta-me pesquisador. Aprendi muito com eles.

Sem o acolhimento da Universidade Estadual de Campinas (UNICAMP) e o fomento da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2010/00994-8), este trabalho tampouco seria possível. Gostaria de agradecê-las profundamente pelo apoio. Nos anos vindouros, comprometo-me a retribuí-lo através da aplicação dos conhecimentos que obtive em prol de uma sociedade mais próspera.

Também gostaria de agradecer aos doutores David Cox e Nicolas Pinto por terem me recebido tão gentilmente nos EUA durante os seis meses em que lá estive na Universidade de Harvard. Essa passagem foi fundamental para o sucesso deste projeto. Por lá, também aprendi muito.

De forma geral, gostaria de agradecer a todos que colaboraram, direta ou indiretamente, com este trabalho. Ressalto aqui a colaboração com o doutor William R. Schwartz, tão oportuna, e as conversas com os doutores Nicolas Poilvert e James Bergstra, tão produtivas e inspiradoras. A todos os colegas de trabalho, discentes, docentes e administrativos, meu muito obrigado.

É preciso muita perseverança para seguir o caminho do doutoramento e minha família é, sem dúvida, uma das grandes responsáveis por eu tê-la mantido durante esses anos. Pelo amor incondicional que me dão — e de amor deriva-se suporte, motivação, paciência, empatia, *etc.* — por fim eu gostaria de agradecer à minha esposa e aos meus pais. Pelos mesmos motivos, gostaria de agradecer à minha avó, aos meus irmãos e a todos "lá em casa". São pessoas que genuinamente me querem bem, assim como meus grandes amigos, que aqui também agradeço.

"Programming, like all engineering, is a lot of work: we have to build everything from scratch. Learning is more like farming, which lets nature do most of the work. Farmers combine seeds with nutrients to grow crops. Learners combine knowledge with data to grow programs."

Pedro Domingos

Related Publications

I. Giovani Chiachia, Alexandre X. Falcão, and Anderson Rocha. Person-specific Face Representation for Recognition. In *IEEE/IAPR Intl. Joint Conference on Biometrics*, Washington DC, 2011.

II. Giovani Chiachia, Nicolas Pinto, William R. Schwartz, Anderson Rocha, Alexandre X. Falcão, and David Cox. Person-Specific Subspace Analysis for Unconstrained Familiar Face Identification. In *British Machine Vision Conference*, Surrey, 2012.

III. Manuel Günther, Artur Costa-Pazo, Changxing Ding, Elhocine Boutellaa, and Giovani Chiachia *et al.* The 2013 Face Recognition Evaluation in Mobile Environment. In *IEEE/IAPR Intl. Conference on Biometrics*, Madrid, 2013.

IV. Giovani Chiachia, Alexandre X. Falcão, Nicolas Pinto, Anderson Rocha, and David Cox. Learning Person-Specific Face Representations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (submitted), 2013.

Contents

List of Tables

List of Figures

Chapter 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 environments, these features might undergo significant variation due to changes in lighting, viewing angle, occlusion, *etc.* Exploring feature extraction that is tailored to specific individuals of interest is a potentially promising approach to tackling this problem.

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] and that the advantage of humans over machines in this scenario is still substantial [11], face familiarity is a specially relevant notion to pursue in the design of robust face recognition systems [12].

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

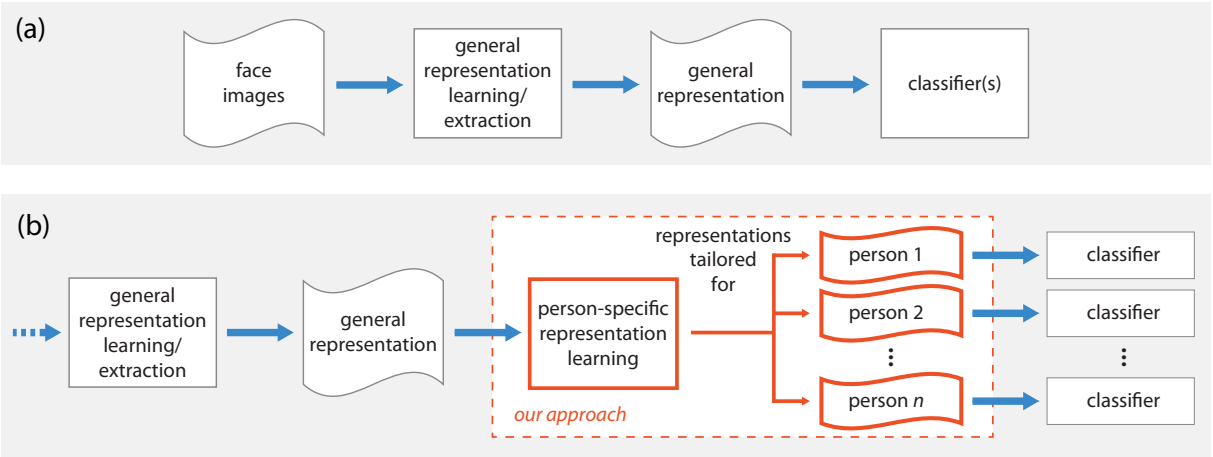


Figure 1.1: Pipelines illustrating how methods can be regarded with respect to the face representation approach they employ. Both pipelines (a) and (b) transform the input images into a feature set where the faces are described by the same, general attributes. Common techniques to derive this representation 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.

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. 1.1, 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 representation are Eigenface [1], Gabor wavelets [2], Local Binary Patterns [3], Fisherface [4], Scale-Invariant Feature Transform [13], 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 [14, 15, 16], learning similarity measures [6, 17], 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 an *intermediate*

learning task that is 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.

While few previous works have already considered the use of person-specific representations in face recognition [18, 19, 20], the advantages of the underlying concept has never been attested before. Here we validate the concept of person-specific face representations, and describe approaches to building them ranging from a patch-based method, to subspace learning, to deep convolutional network features. Taken together, we argue that these techniques show that the person-specific representation learning approach holds great promise in advancing face recognition research.

1.1 Thesis Organization and Contributions

As a consequence of being one of the most active pursuits in computer vision [12], the face recognition problem has been addressed from many different perspectives. In spite of this fact, it is still possible to devise seminal works in the area. Likewise, it is also possible to draw a connection between the progress made in the development of the algorithms and the recognition scenario that they are targeted to. Therefore, in order to better contextualize this thesis, in Chapter 2 we present a summary of face representation techniques and recognition scenarios as they evolved over time.

Our experiments consider both the *constrained* and the *unconstrained* face recognition scenarios respectively represented by the UND [21] and the PubFig83 [16] datasets introduced in Chapter 3. After describing these datasets, we then present and evaluate three distinct methods for person-specific representation learning, with the goal of progressively validating the overarching approach.

The first method, presented in Chapter 4, is designed to be as simple as possible and is based on an algorithm that we call "discriminant patch selection" (DPS) [22]. This algorithm enables us to carry out an evaluation of the idea of person-specific representations in a constrained face recognition scenario where an intuitive understanding is more tenable.

Second, in Chapter 5, we explore a more powerful set of techniques based on subspace projection [23]. In particular, we introduce a person-specific application of partial least squares (PS-PLS) to generate per-individual subspaces, and show that operating in these subspaces yields state-of-the-art performance on the PubFig83 benchmark dataset. 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 generating person-specific subspaces yields significant improvements in face recognition performance as compared to either "general" representation learning approaches or classic supervised learning alone. Further, we show that such subspace methods, when applied atop a deep convolution neural network representation can achieve recognition performance that exceeds previous state-of-the-art performance.

Therefore, in our third and last method, we incorporate person-specific learning directly into a deep convolutional neural network. We demonstrate in Chapter 6 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 a simple approach based on SVMs. The inspiration to this approach comes from the assumption that class-specific transformations might be learned at the top of the human ventral visual stream hierarchy [24], and that neurons responding to specific faces might exist in the brain at even deeper stages [25]. 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.

Finally, a compilation of our contributions and experimental findings, along with new directions to this line of research, are presented in Chapter 7.

Chapter 2

Background

There is a sensible relationship between the progress made in the development of face representation algorithms and the recognition scenario that they are targeted to. In this chapter, we present a summary of these techniques and scenarios as they evolved over time.

2.1 Face Representation

Since the seminal work of Kanade [26] in automated face recognition, the task of transforming pixel values into features conveying more important information is a paramount step in any face recognition pipeline. Intuitively, pixel values are highly correlated and uninformative by their own. So, back in 1973, Kanade proposed to represent faces based on distances and angles between fiducial points such as eye corners, mouth extrema, nostrils, among others, with procedures to automatically detect them [26]. This work is the first milestone that we consider in the timeline presented in Fig. 2.1 about groundbreaking contributions to the topic of face representation.

Methods solely based on geometric attributes, as proposed by Kanade, are today known to discard rich information of facial appearance. After a dormant period [27], face recognition revived in 1991 with the advent of Eigenface, a technique based on principal component analysis (PCA) for learning and extracting low dimensional face representations via subspace projection [1]. Indeed, Eigenface gave rise to a class of face representation methods known as holistic [28], with projection vectors operating in the full image domain. While the Eigenface method learns projection vectors according to the principle of overall maximal variance, Fisherface, based on linear discriminant analysis (LDA), learns basis vectors with the objective of maximizing the ratio of between-class and within-class variance [4]. The incorporation of class label information in the framework of holistic methods was an important step towards better face representations. Hence,

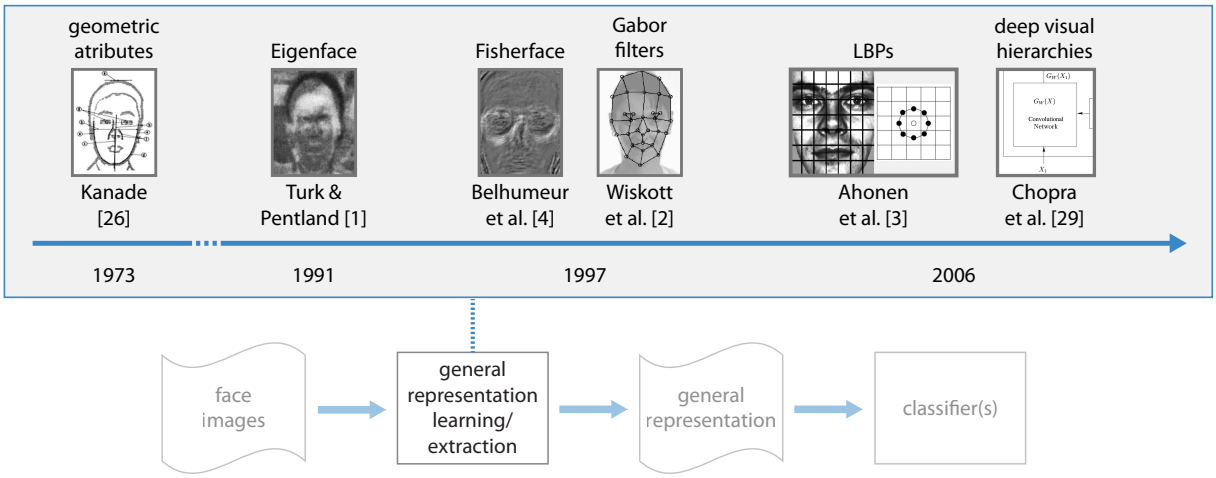


Figure 2.1: Milestones in the history of face representation, from Kanade's seminal work [26] to the use of deep visual hierarchies [29]. In order to better contextualize the techniques, we replicate the traditional recognition pipeline of Fig 1.1(a) at the bottom.

Fisherface is the third, key face representation technique that we highlight in Fig. 2.1.

Contemporary to Fisherface, another milestone in face representation was the use of Gabor filter responses to represent the appearance of regions around facial fiducial points [2]. This approach can be seen as an extension of Kanade's method, in that the representation relies on fiducial points. However, while Kanade only relied on geometric measures computed from these points, appearance information extracted from their neighborhood provide much richer information for the task at hand. This original work introduced the broad idea of locally representing facial features, and inspired a fruitful vein of representation methods. Within this vein, we point out in Fig. 2.1 the widely used local binary patterns (LBPs) [3]. It consists of a simple and fast image transformation based on local pixel value comparisons that leads to a compact texture description. In fact, the use of LBPs for face representation is coupled with the extraction of local histograms, resulting in a representation with a certain degree of translation invariance [3].

Finally, the last approach for face representation considered in this overview refers to a class of representation methods based on deep visual hierarchies, whose first application on raw face images [29] was at about the same time as LBPs. These hierarchies can be seen as a form of face representation that departs from the idea of "engineered" features by instead using a cascade of linear and nonlinear local operations that are — to some extent — learned directly from the face images, as in the original work of Chopra *et al.* [29] with convolutional neural networks.

Today, most of the principles underlying these representation techniques are present in state-of-the-art approaches. For example, among the best performing methods in un-

constrained face verification, there are systems that rely on fiducial points to extract face features from their neighborhood [30, 31], something that borrows ideas from the first (fiducial points), the fourth (local features), and the fifth (LBPs and related) milestones presented in Fig. 2.1. PCA and LDA are vastly used as an intermediate processing step of many current top performers [31, 32]. Deep visual hierarchies have definitely demonstrated their potential for unconstrained face recognition [16]. In addition, each of these methods were unfolded and combined in a profusion of ways that are beyond the scope of this overview. As we shall see throughout the thesis, there is a good overlap between the general representation techniques highlighted in Fig. 2.1 and the techniques that serve us as basis to learn person-specific face representations.

2.2 Recognition Scenarios

Research on automatic face recognition in the 1990s and the early 2000s was mostly based on mugshot-like images with controlled levels of variation. Indeed, it all started with the Facial Recognition Technology (FERET) program in 1994 [33], that can be regarded as the first attempt to organize the area around a well-defined problem. After 1994, FERET evaluations were carried out for more two years and images from the last edition, in 1996, are still available for research purposes. They are similar to the images of the FRGC (experiment 1) [34] and the UND (collection X1) [21] datasets, shown in the left part of Fig. 2.2. Since the users were asked to meet specific poses and expressions, and illumination conditions were carefully taken into account, this image acquisition scenario is referred to as *constrained*.

From constrained images, many lessons have been learned. Among them, for example, the fact that females are harder to recognize than males [35]. These findings and, more importantly, research directions — such as the need to make systems more robust to changes in illumination — were only possible with the concerted effort of institutions like the National Institute of Standards and Technology (NIST), which was in charge of the FERET and FRGC programs, and currently promotes advances in the area by means of challenges such as FRVT [36] and GBU [37], among others.[1] Nowadays, many benchmarks for automatic face recognition consider more realistic, uncontrolled face images in their protocol. For example, the GBU challenge considers face pictures taken outdoors and in hallways [37]. Likewise, a recent competition on mobile face recognition [32] — based on the MOBIO dataset [38] — was carried out on images captured with little to no control,[2] under conditions approaching the *unconstrained* setting (Fig. 2.2).

A new perspective to face recognition research was introduced with the release of

[1] `http://www.nist.gov/itl/iad/ig/face.cfm`

[2] In fact, users were asked to be seated and pictures were taken indoors.

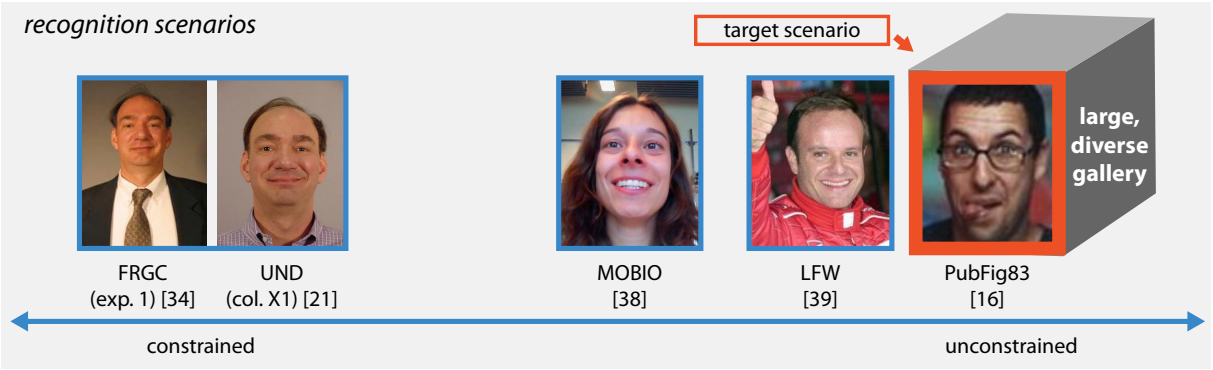


Figure 2.2: Face recognition from the constrained to the unconstrained scenario. The scenario around which the area was first organized was constrained, in that individuals were asked to meet specific poses and expressions, and illumination conditions were carefully taken into account. With advances on the technology and the advent of the Internet, nowadays many research groups target their algorithms to the unconstrained scenario, where requirements on individuals are minimal.

Labeled Faces in the Wild (LFW) [39], a dataset based on the original idea of collecting images of celebrities from the Internet with the only requirement that their faces were detectable by the Viola-Jones algorithm [40]. Even though the resulting dataset was biased towards face pictures typically found in news media, it embodied a new factor of variation in the recognition scenario: diversity in appearance. Due to its interesting properties and ease of use, and possibly also because its curators constantly update and report progress made on it,[3] LFW is currently largely adopted. Indeed, LFW has motivated the creation of many other datasets, among them PubFig [11] and its refined version PubFig83 [16], which have similar recording conditions, but serves to other purposes (Fig 2.2).

While LFW contains over 5,000 people, only five individuals have more than 100 images. In contrast, PubFig83 has 83 people, but each individual has at least 100 images. While LFW — like most NIST challenges, including GBU [37] — is designed for pair matching tests, and has a protocol that does not allow learning any parameter from gallery images,[4] PubFig83 is designed to approach familiar face recognition, and has a protocol that actually fosters learning algorithms to take most out of gallery images.

Notwithstanding the fact that many other interesting datasets remain to be cited,[5] we believe that the ones that we mentioned here illustrate well the continuum from the constrained to the unconstrained recognition scenarios. For example, Multi-PIE [41] is

[3] `http://vis-www.cs.umass.edu/lfw/results.html`

[4] In fact, LFW is conceived in terms of "pairs", not individual images. The notion of gallery and probe images does not even exists in this dataset.

[5] A non-exhaustive list can be found at `http://www.face-rec.org/databases`.

an interesting dataset to study face recognition under severe pose variations. To this purpose, a laborious setup was used to acquire images from precisely different viewpoints. Its highly controlled nature enables researchers to factor out other sources of variation and carefully address the problem. However, exactly because of its motivation, the dataset reflects a constrained recognition scenario.

Overall, we consider PubFig83 as our target scenario in this work because its has a large pool of heterogeneous face images for each individual and its evaluation protocol allows us to learn from these images. In fact, there is a perfect match between the recognition scenario that this dataset reproduces and the motivation of this thesis.

Chapter 3

Datasets and Evaluation Protocol

We follow the idea of gaining insight into the constrained scenario, where factors interfering in the results are alleviated, to later extending our representation learning methods to a scenario that best suits the approach. In the following sections, we present the controlled and the uncontrolled datasets of our choice, with their respective evaluation protocol, to accomplish this goal.

3.1 Constrained: UND

Our experiments in the controlled scenario are based on the X1 collection of the UND face dataset [21]. This dataset is arranged in weekly acquisition sessions in which four face images were obtained by the combination of a small variation in illumination and two slightly different facial expressions.

We designed an evaluation protocol that allows us to learn person-specific representations from gallery images as well as to account for variability in our tests. In particular, we considered the 54 subjects whose attendance to the acquisition sessions were highest, so that each person was recorded at least in seven and at most in ten sessions. This procedure resulted in a dataset with 1,864 images — with at least 28 images per individual — which enabled us to split the dataset into ten pairs of training and test sets. Considering the images in chronological order, for each split, we selected two images of each individual for the training set and used the remaining images as test samples. In addition, all images were registered by the position of the eyes, cropped with an elliptical mask, and were made 260×300 pixels in size.

Fig. 3.1 presents training and test images of four individuals in UND. We can see that test images differ from training images only by a small amount, specially due to facial expression. UND represents the typical dataset used in automated face recognition research until the late 1990s and the early 2000s. While our target scenario is unconstrained face

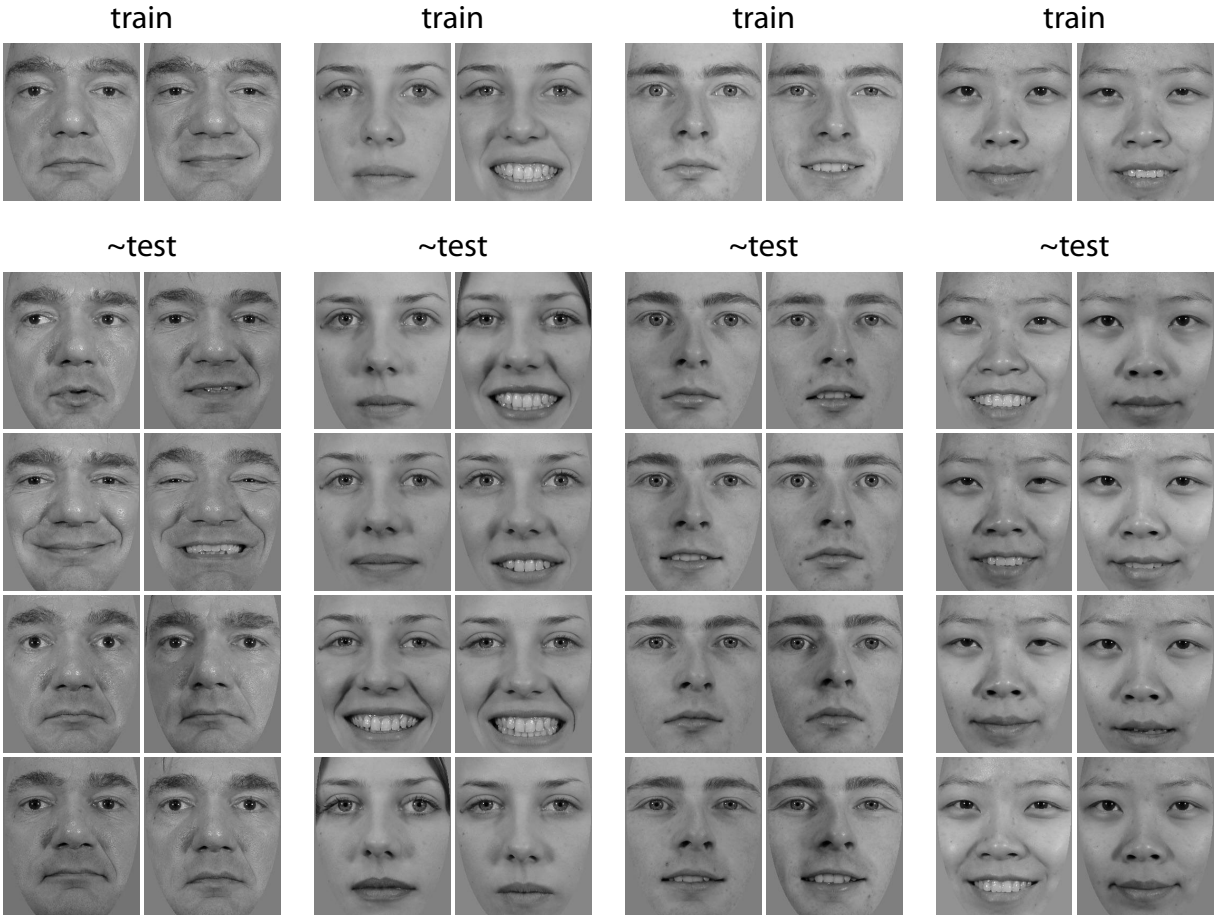


Figure 3.1: Training and test images of four individuals in the UND dataset. As we can see, test images differ from training images only by a small amount. This recognition scenario was typical in automated face recognition research of the early 2000s.

recognition, in this thesis, the controlled images of UND serve to provide insight regarding the value of person-specific representations.

Evaluations in this dataset are performed in identification mode, where the task is to identify which of a set of previously-known faces a new test face belongs to.

3.2 Unconstrained: PubFig83

The PubFig83 dataset [16] is a subset of the PubFig dataset [11], which is, in turn, a large collection of real-world images of celebrities collected from the Internet. This subset was established and released to promote research on familiar face recognition from unconstrained images, and it 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 individuals.

To our knowledge, this is the publicly available face dataset with the largest amount of unconstrained, uncorrelated images per individual. This characteristic is fundamental in validating our claim — which has a perfect fit with the dataset motivation — and that is why this thesis is mostly validated on PubFig83.[1]

We aligned the images by the position of the eyes and followed the original evaluation protocol of [16], where the dataset is split into ten pairs of training and test sets with images selected randomly and without replacement. For each individual, 90 images were considered for training and 10 for test.

In Fig. 3.2, we present images of four individuals in a given split of PubFig83. While here we only have space to show 10 (out of 90) training images of each individual, all their respective test images are presented. We can observe that this dataset is considerably more challenging than UND. Indeed, due to its unconstrained nature, PubFig83 presents at the same time all factors of variation in face appearance: pose, expression, illumination, occlusion, hairstyle, aging, 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 [42]. Another challenging aspect of the dataset is that images are originally 100×100 pixels in size.

On PubFig83, we report results both in identification mode as well as in verification mode. In the later, the task is to decide whether or not a given test face belongs to a claimed identity.

[1] Though, in part, we additionally validate our methods on the private Facebook100 dataset [16], as we shall see in Appendix B.

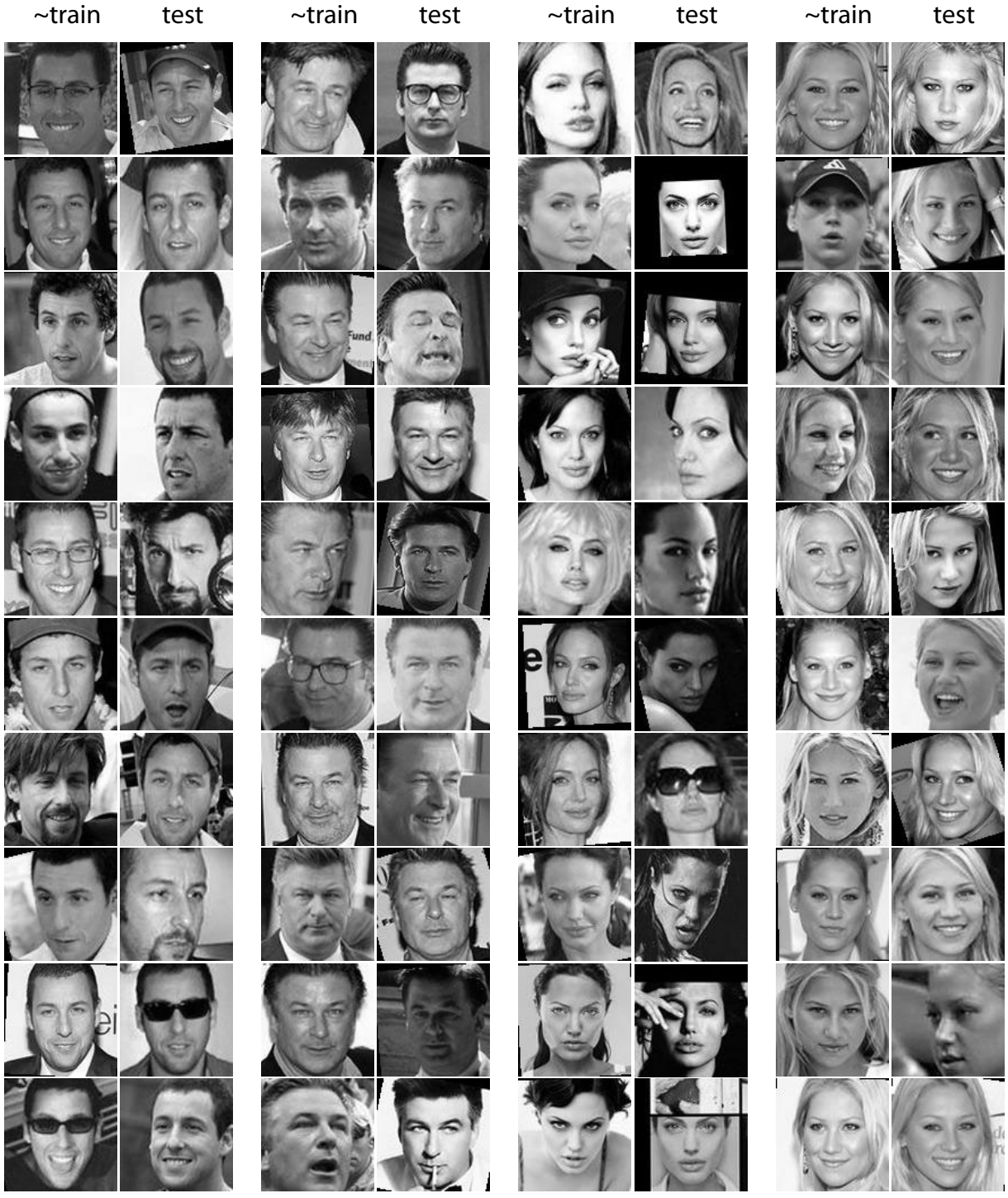


Figure 3.2: Images of four individuals in a given split of PubFig83. While here we only have space to show 10 (out of 90) training images of each individual, all their respective test images are presented. We can observe that this dataset is considerably more challenging than UND. Indeed, due to its unconstrained nature, PubFig83 presents at the same time all factors of variation in face appearance.

Chapter 4

Preliminary Evaluation

This preliminary evaluation is aimed at being as simple and intuitive as possible. Therefore, here we follow the basic idea of matching face images via histograms of Local Binary Patterns (LBPs) extracted from patches on different positions of the face. Indeed, the approach presented in this section is closely related to the methods in [3], but using a different patch selection mechanism that is crucial to our purpose.

Given that we calculate LBPs from an 8-neighborhood, our matching schema considers histograms with 256 bins. Formally, let $\mathcal{P}'$ be the set of patches considered for the matching and H_p be the histogram of the LBPs from patch p. The patch-based dissimilarity between images $\mathbf{I}_1$ and $\mathbf{I}_2$ is

$$D(\mathbf{I}_1, \mathbf{I}_2, \mathcal{P}') = \sum_{\forall p \in \mathcal{P}'} \sum_{b=1}^{256} |H_{p,b}(\mathbf{I}_1) - H_{p,b}(\mathbf{I}_2)|, \tag{4.1}$$

where $H_{p,b}$ represents the value of bin b of patch p. In other words, the dissimilarity corresponds to the summation of the absolute difference over the bins of each patch histogram, *i.e.*, the L_1 distance.

4.1 Discriminant Patch Selection (DPS)

The concept of selecting patches to better describe object classes in images has been studied in many contexts. For example, in [43], the authors present methods for selecting patches that are informative to detect objects, and, in [44], patch selection is proposed in a probabilistic framework for the recognition of vehicle types.

The idea of our DPS procedure is to determine (x, y) coordinates for patch selection according to the *discriminability* they have in a group of aligned training images with at least two images per category. For a given patch in a given image, its discriminability

Algorithm 1 DISCRIMINANT PATCH SELECTION

INPUT:	Set of training images $\mathcal{T}$ and classes $\mathcal{C}$, set of patch positions $\mathcal{P}$, discriminability function $F(p, \mathbf{I}, \mathcal{G})$, and *patch selection criterion*.
OUTPUT:	Class-specific models M_c and patches $\mathcal{P}'$ selected according to the provided criterion.
AUXILIARY:	Function $C(\mathbf{I})$, image $\mathbf{I}$, and variables c and d.

1. **For each** $c \in \mathcal{C}$ *and* $p \in \mathcal{P}$ **do** $M_{c,p} \leftarrow 0$
2. **For each** *patch position* $p \in \mathcal{P}$ **do**
3. | **For each** *image* $\mathbf{I} \in \mathcal{T}$ **do**
4. | | $c \leftarrow C(\mathbf{I})$.
5. | | $d \leftarrow F(p, \mathbf{I}, \mathcal{T} \backslash \{\mathbf{I}\})$.
6. | | $M_{c,p} \leftarrow M_{c,p} + d$.
7. **Select patches** *from models* M_c *into* $\mathcal{P}'$ *according to the criterion related to their discriminability.*

is measured on an individual basis with respect to patches of the other training images. By interchanging such image, the discriminability of patches at the same position is computed for all classes. This is done for the whole set of patches. At the end, each class is associated with one discriminability value per patch position. We refer to these mappings as the *class-specific* models that we use for patch *selection*.

Let $\mathcal{T}$ be a set of labeled training images and $\mathcal{P}$ be a set with all patch positions considered for selection. Assuming that function $F(p, \mathbf{I}, \mathcal{G})$ measures how good a patch at $p \in \mathcal{P}$ in image $\mathbf{I}$ discriminates its class with respect to other patches in the image subset $\mathcal{G} = \mathcal{T} \backslash \{\mathbf{I}\}$, and considering that function $C(\mathbf{I})$ retrieves the correct class $c \in \mathcal{C}$ to which image $\mathbf{I}$ belongs, a pseudocode for the method can be defined as in Alg. 1.

Note that M_c in Alg. 1 is considered a class-specific model in the sense that the discriminability of patches at $p \in \mathcal{P}$ with respect to class c are accumulated in $M_{c,p}$. While the patch selection criterion may take into account the discriminability of the patches by the problem classes (*i.e.*, by M_c), it may also fuse the models in order to consider patch discriminabilities common to the whole training set, in which case we obtain *general* models.

4.2 DPS Setup

For both experiments in the constrained and in the unconstrained scenario, we consider $\mathcal{T}$ as the training set of a particular dataset split. The set $\mathcal{P}$ contains all possible patch

positions regarding patch sizes of 20×20 and 10×10 pixels — empirically chosen for the UND and the PubFig83 datasets, respectively — lying in the image domain.

Concerning the discriminant function $F(p, \mathbf{I}, \mathcal{T} \backslash \{\mathbf{I}\})$, we measure the discriminability of a patch at position p in a given pivot image $\mathbf{I}$ as the identification rank obtained by matching it with all patches at the *same position* in the remaining images. The discriminant criterion is actually the negation of the rank, provided that the lower the rank, the more discriminant the patch. Such measurement of discriminability by the identification rank was only possible because we consider at least two training images per class in the dataset splits (Chapter 3).

Finally, we select patches based on models M_c according to the experiment we want to evaluate. Our main purpose is to build person-specific representations via the selection of the most discriminant patches from each M_c model. In order to avoid overlapping patches, we constrain the selection so that each new selected patch must have its center at a minimum distance from the previously selected ones.

4.3 Experiments in the Controlled Scenario

As shown in Fig. 1.1(b), the learning of person-specific representations results in representation spaces associated to each subject. Therefore, a classification engine is required to operate in each of these spaces. For the sake of simplicity, the experiments in the controlled scenario are based on *nearest neighbor* (1-NN) classifications. In order to recognize a test face, we match it to all faces in the gallery in each representation space according to Eq. 4.1. As a result, we obtain a number of 1-NN predictions. These predictions are then fused by a voting scheme, *i.e.*, the identity with the greater number of votes is given to the test face. See Appendix A for a running example of this identification scheme.

The experiments with controlled images consist of comparing the identification rate obtained in the UND dataset with the selection of patches according to six different criteria. We start with the selection of the person-specific *most* discriminant patches, *i.e.*, the criterion that implements the idea of learning a good face representation specific to each person, and call this selection strategy as experiment A.

In Table 4.1, we present the characteristics of each experiment along with the mean accuracy and the standard error obtained across the ten dataset splits (Sec. 3.1). As we can see, in experiment A we have $n = |\mathcal{C}| = 54$ person-specific representation spaces — corresponding to the number of subjects in the dataset — each one composed by the concatenation of histograms of LBPs computed from the 48 most discriminant patches of that person.[1] An illustration of a given person-specific model is provided in Fig. 4.1(a)

[1]We decided to select 48 patches for each person because such number seemed to us appropriate to describe a large portion of the face.

Table 4.1: Experimental details and performance evaluation in the controlled scenario. As we can see from experiment A, the representation based on the person-specific most discriminant patches resulted in better identification rates. A per-split performance plot is presented in Fig. 4.2.

exp.	patch sel. criteria	person-specific	# rep. spaces	# patches /space	accuracy (%)
A	most disc.	yes	n	48	97.07±.36
B	least disc.	yes	n	48	92.72±.75
C	most disc.	yes	1	$48n$	94.89±.65
D	most disc.	no	1	48	94.87±.39
E	random	yes	n	48	96.21±.45
F	non-overlap	no	1	13×15	96.23±.42

as well as the patches that were selected to represent this individual in experiment A.

The first alternative patch selection criterion that we compare with experiment A is to select the *least* discriminant patches for the person-specific representation spaces. This can be viewed as a sanity check to assure that DPS is behaving as expected. Compared with A, it is possible to observe that experiment B presents a significant drop in performance.

The next comparison is the most interesting outcome of this preliminary evaluation. It consists of contrasting experiment A with experiment C, whose patch selection strategy is the same, but the patches are *assembled* into a single representation space. The interesting point to observe is that the same data are employed by both methods. In experiment A, we consider 54 representation spaces with 48 patches each, while in experiment C, we consider a single feature space with the same 54×48=2,592 patches. The difference in performance observed between experiments A and C suggests that undesirable *cancellations* are occurring when the person-specific representations are tiled in a single space. We consider this fact as a good support to our hypothesis.[2]

Experiment D refers to the selection of the 48 patches that are the most discriminant for all persons simultaneously. In this case, the patch discriminability from the persons are correspondingly merged by summing them up before the selection, leading to a set of *general* discriminant patches (see Sec. 4.1). This strategy is well-known in the literature and reflects the paradigm of creating a representation space that highlights the importance of face aspects that better distinguish among all individuals. Fig. 4.1(b) shows the model obtained in D along with the corresponding most discriminant patches. With respect to accuracy, we can also observe in Table 4.1 a significant difference between experiments A

[2]Note that because we are using 1-NN classifiers, these cancellations only occur due to the voting scheme (Appendix A) employed in experiment A before the final prediction. Otherwise, experiments A and C would perform exactly the same.

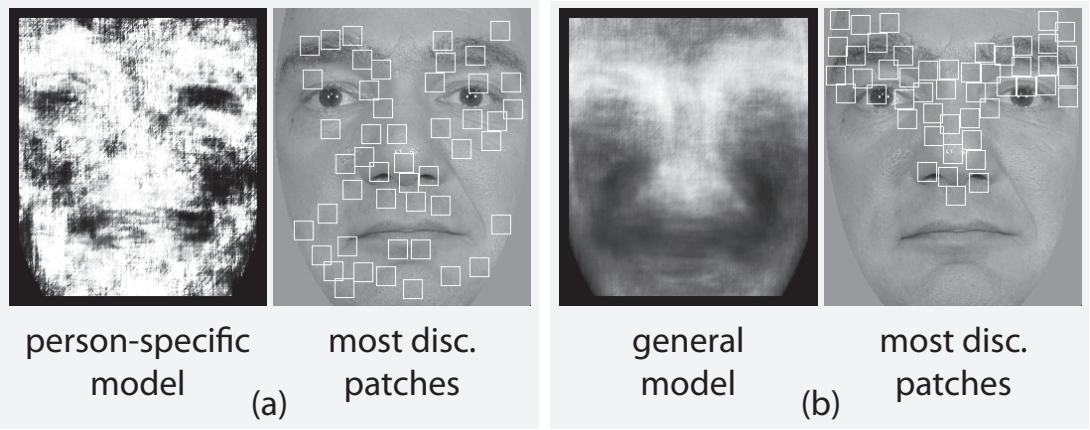


Figure 4.1: A given person-specific model and the most discriminant patches for this individual (a). The general model obtained with the summation of all person-specific models and the corresponding most discriminant patches (b). Models learned from the first dataset split.

and D. Aside from this fact, here it is possible to notice the importance of the eyebrows in face recognition, which are facial features known to contribute in an important way in human face perception [9, 12].

We also evaluate the random selection of 48 patches per individual within the elliptical face domain, following the same matching strategy used in experiments A and B. This experiment is called E and performed worse than A as well. Interestingly, however, experiment E performed better than C and D. We believe that the random criterion, by being uniform and not allowing overlapping patches, enabled a well distributed selection of patches within and among the person-specific representation spaces. This possible representation regularly covering the face image domain may have led to this good performance.

Therefore, the last experiment in the controlled scenario, named F, stands for a regular grid composition of non-overlapping patches covering the entire image. Given that images in the UND dataset are 260×300 pixels in size and patches are 20×20 pixels, this method employs a grid of 13×15=195 patches to describe the faces. We can observe in Table 4.1 that experiment A also prevails over F, and that they are the top performing representation strategies.

Notwithstanding the proximity among accuracies presented in Table 4.1, in Fig. 4.2 we provide a per-split comparison of the experiments. This visualization enables us to see that experiment A achieves a consistently better performance across the splits. Therefore, when the experiments are paired by the splits and a Wilcoxon signed-rank test is carried out, the performance of A is significantly different from all other experiments ($p < 0.01$).

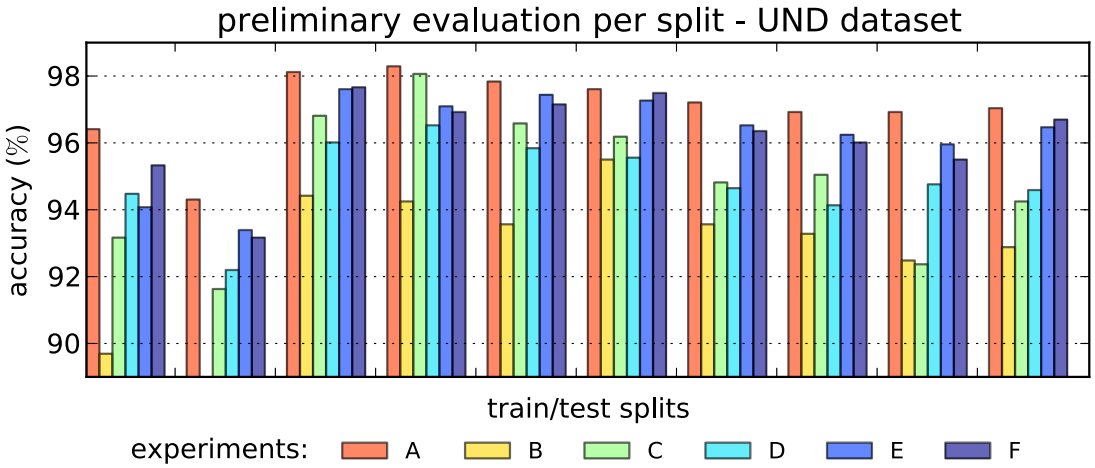


Figure 4.2: Per-split evaluation on the UND dataset. Details of each experiment are available in Table 4.1. When individually compared with each other experiment, the performance of A is significantly different according to the Wilcoxon signed-rank test ($p < 0.01$).

Overall, the results measured in this controlled setting provide early evidence about the potential importance of learning person-specific face representations.

4.4 Experiments in the Unconstrained Scenario

With the use of the PubFig83 dataset, this preliminary evaluation gains in importance not only due to the uncontrolled conditions in which the images were obtained, but also due to the scale of the learning task, which increases from two images of 54 individuals in the UND dataset to 90 images of 83 subjects in PubFig83.

We start this round of evaluation by considering the top two performing methods from the previous section, namely experiments A and F. As mentioned in Sec. 3.2, we report results on PubFig83 following the same protocol of [16], providing mean accuracy and standard error obtained from ten random training/test dataset splits.

In Table 4.2, we can observe that the difference in performance between experiments A and F is considerably greater than the difference between the same methods in the controlled scenario (Table 4.1). Given that predictions in both methods are made with 1-NN classifiers, we understand that the selection of person-specific discriminant patches provides an important aid in the classifier generalization.

Despite the relative superiority of experiment A over F, when compared with state-of-the-art methods [16, 23], which use robust visual representations and powerful classification engines, the performance obtained with experiment A on PubFig83 is substantially

Table 4.2: Preliminary evaluation in the unconstrained scenario. Here the difference in performance between the top two methods in the controlled scenario (exp. A and F) is much greater. However, as exp. G and H suggest, in this scenario we must consider more robust learning techniques.

exp.	patch sel. criteria	person-specific	# rep. spaces	classi-fier	accuracy (%)
A	most disc.	yes	n	1-NN	45.25±.56
F	non-overlap	no	1	1-NN	32.16±.71
G	most disc.	yes	n	SVM	62.94±.28
H	non-overlap	yes	1	SVM	65.28±.52

lower. In order to evaluate the impact of using a better classifier on top of the same visual representations, we replace the 1-NN classifier in experiments A and F with linear SVMs, and call these new experiments as G and H, respectively. We use LIBSVM [45] to train the linear machines and, for each split, we estimate the SVM regularization constant C via *grid search*, considering a re-split of the training set and possible C values of $\{10^{-3}, 10^{-2}, \ldots, 10^{5}\}$.

As expected, in Table 4.2 we can see that the use of SVMs in experiments G and H results in a significant performance boost. We note that experiment H, although does not operate in person-specific representations, is presented as person-specific. This is because we use a one-versus-all learning strategy when training the classifiers. More interesting, however, is the fact that SVM operates better in experiment H, when it is provided with the whole set of LBP histograms, so that its learning principle can make the most out of the training data.

In general, the experiments conducted in this section give us the idea that we need better visual representations in order to obtain satisfactory performance on the PubFig83 dataset. Moreover, we observe that the combination of our discriminant patch selection (DPS) method with the 1-NN classifier, which was fundamental in providing insight into the controlled problem, cannot cope with the challenging problem imposed by PubFig83. Therefore, we conclude that beyond better visual representations, we also need more robust techniques to further pursue the idea of explicitly learning person-specific face representations.

Chapter 5

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 [46], 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 [47, 48], 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. Even though any projection technique that attempts to discriminate between face identities, one at a time, can be considered person-specific in some sense, subspace models can offer more degrees of freedom to accommodate within-class variance in appearance.

5.1 Partial Least Squares (PLS)

Partial least squares is a class of methods primarily designed to model relations between sets of observed variables by means of latent vectors [46, 49]. 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 [50, 51] and, recently, it is also being successfully applied to computer vision problems for dimensionality reduction, regression, and classification purposes [47, 48, 52, 53, 54].

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

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad \text{and} \quad \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}, \tag{5.1}$$

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 [46], in which projection vectors $\mathbf{w}$ and $\mathbf{c}$ are determined iteratively such that

$$[cov(\mathbf{t}, \mathbf{u})]^2 = \max_{||\mathbf{w}||=||\mathbf{c}||=1} [cov(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2, \tag{5.2}$$

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 [49]:

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 [49]. The loadings are then computed by regressing $\mathbf{X}$ on $\mathbf{t}$ and $\mathbf{Y}$ on $\mathbf{u}$, *i.e.*,

$$\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^\mathbf{T}\mathbf{t}) \quad \text{and} \quad \mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^\mathbf{T}\mathbf{u}). \tag{5.3}$$

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,

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad \text{and} \quad \mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{t}^T\mathbf{Y}/(\mathbf{t}^\mathbf{T}\mathbf{t}). \tag{5.4}$$

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 regression and other problems, and how they compare with other techniques, we refer the reader to [49, 55, 56].

5.2 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 [49]. It is worth recalling from Sec. 5.1 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,

$$\mathbf{w}_i = \mathbf{X}i^T\mathbf{y}_c, \tag{5.5}$$

where $\mathbf{X}i$ is the matrix $\mathbf{X}$ deflated up to iteration i according to Eq. 5.4.

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 throughout the samples, a property also suggested to be relevant in mental representations of familiar faces [8].

In Fig. 5.1 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 by considering the different representations of the samples over the subspaces. 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.

5.3 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 the one that performed best in our preliminary evaluation on PubFig83 (Sec. 4.4, experiment H), and it is based on non-overlapping histograms of LBP patches. The second and third representations are called V1-like+ and HT-L2-1st. They are taken from [16] and can be thought of as biologically-inspired visual models of increasing complexity. Finally, the fourth visual representation is similar in spirit to HT-L2-1st and consists of a three-layer hierarchical convolutional

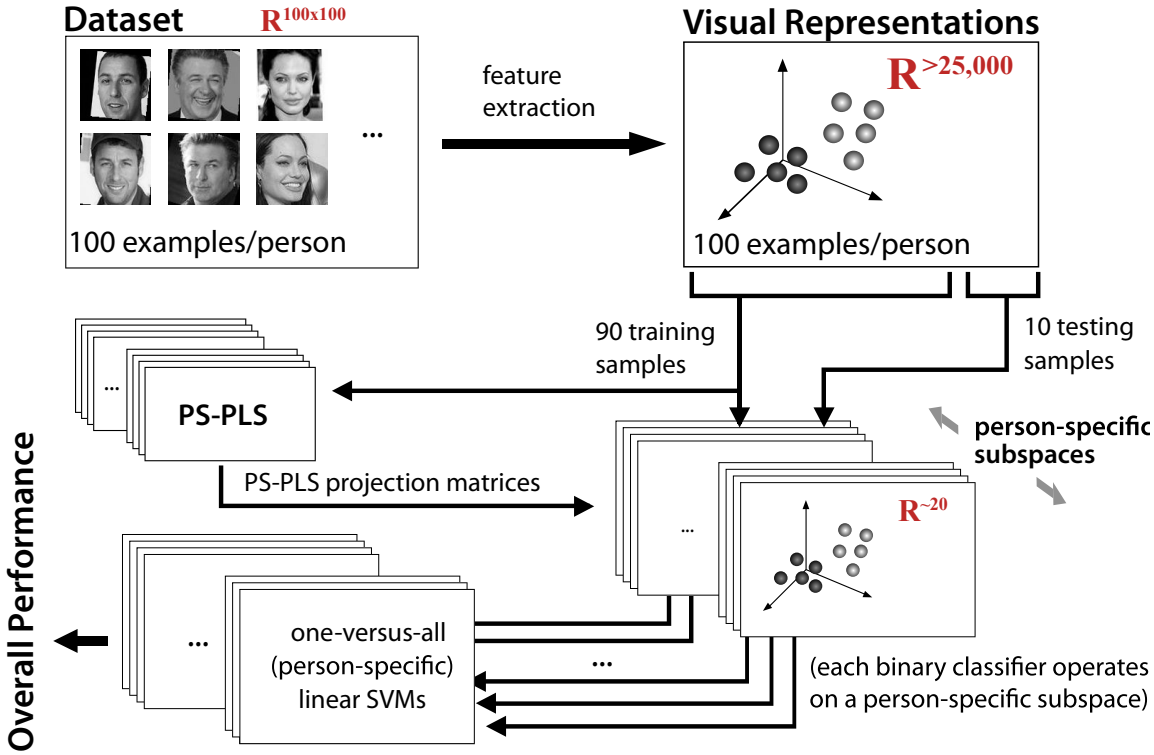


Figure 5.1: From the training samples, PS-PLS creates a different face subspace for each individual. A different classifier is then trained in each subspace.

network. We refer to this visual representation as L3+, as it is a slight modification of the HT-L3-1st network found in [15].[1]

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). 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. Finally, due to its interesting properties [57, 58], we also consider RP vectors sampled from a univariate normal

[1]The only difference between L3+ and HT-L3-1st is that the later performs an additional normalization as a last step.

distribution.

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 (see Sec. 5.1 for details). In this case, as $\mathbf{Y}$ has multiple variables, this form of PLS is also known as PLS2 [49]. While there remains substantial room to evaluate other subspace methods — including kernelized versions of PCA [59], LDA [60], and PLS [61] — 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 *validation* 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 [62] was used to compute the subspace models and LIBSVM [45] was used to train the linear SVMs. In all cases, the data was scaled to zero mean and unit variance.

Table 5.1: Mean identification rates obtained with different face 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 (see Sec. 5.3 for details).

Models	LBP	V1-like+	HT-L2-1st	L3+	
RAW	65.28±.52	74.81±.35	83.66±.55	88.18±.24	d ($\mathbb{R}^d$)
Multiclass Unsupervised					
RP	61.77±.57	69.04±.44	79.92±.50	85.77±.26	6,640
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

5.4 Results

The results are shown in Table 5.1. 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.[2] 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.

With the *multiclass unsupervised* techniques, we see no boost in performance above RAW. Since unconstrained face images have a considerable amount of noise 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* subspace models, 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

[2]Note that the performance obtained with RAW LBP representations is the same of exp. H in Table 4.2, as the methods are the same.

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. We cannot disregard noise in the unconstrained scenario.

In Appendix B, the very same performance pattern is observed with other two visual representations and on an additional private dataset called Facebook100. We omit these numbers here for a better flow in reading and also because, due to privacy concerns, results on Facebook100 are non-replicable. In any case, these extra experiments strengthen the advantage of person-specific subspace analysis via PLS in the familiar face identification setting.

In Fig. 5.2(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.[3] Similar plots for PCA, LDA and multiclass PLS are available in Appendix C. Considering that the samples of Adam Sandler are in red, Fig. 5.2(a) illustrates one point that we observed throughout the experiments, *i.e.*, 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 discriminative information. In our experiments, performance began to saturate around 20 projection vectors.

Fig. 5.2(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 roughly see that higher importance is being given to Adam Sandler's mouth and forehead (first row), to Alec Baldwin's eyes, hairstyle, and chin (second row), and to the configural relationship of Angelina Jolie's face attributes (third row).

Columns in Fig. 5.2(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

[3]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.

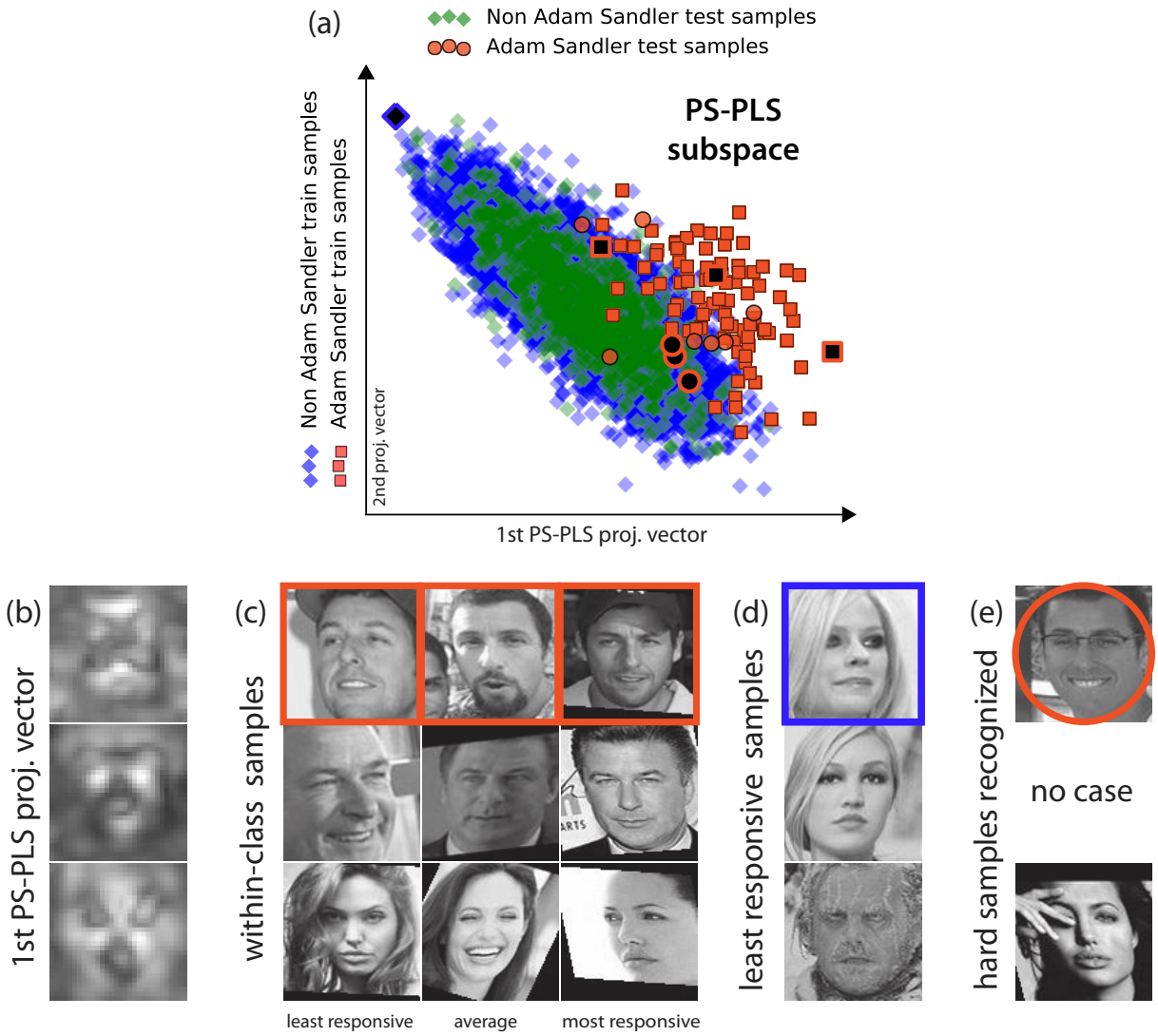


Figure 5.2: (a) Scatter plot of training and test samples projected onto the first two PS-PLS projection vectors of Adam Sandler's subspace learned using V1-like+. (b) First PS-PLS projection vector for three individuals in the dataset. (c) Within-class most, on average, and least responsive face samples 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).

Adam Sandler, these samples are highlighted in the plot in Fig. 5.2(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. 5.2, 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. 5.2(a).

In general, we argue that these subspaces are useful both for noise removal and for accentuating discriminative person-specific face aspects. In unconstrained face recognition settings, both of these issues are of fundamental importance. 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.

Chapter 6

Deep Person-Specific Models

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 Sec. 5.4 led us to explore whether the key theme of person-specific representation could be incorporated more integrally into that feature representation.

The L3+ representation is based on the use of deep architectures for processing visual information [15]. Such approach has a long tradition in the machine learning literature [63, 64, 65, 66], and has been gaining attention due to recent breakthrough results in a number of important vision problems [15, 67, 68, 69]. 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. For an overview on this type of visual representation, see Appendix D.

Since the work of Hinton *et al.* [66], 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 [70, 71, 72], 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 [71].

The work of Pinto *et al.* [15] 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* benchmark [39].[1] On the other hand, it is the basis of our L3+, a likewise best performing face representation in the ICB-2013 competition (Appendix E). In fact, this representation can be understood

[1] `http://vis-www.cs.umass.edu/lfw/results.html`

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 [15].

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 [24] and (ii) the experimental evidence suggesting that neurons responding to faces of specific individuals should exist in the brain at even deeper stages [25].[2]

6.1 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 [15]. As we can observe in the left panel of Fig. 6.1, the third and topmost layer of the L3+ network sequentially performs linear *filtering*, filter response *activation*, and local *pooling*.[3]

The filtering operation takes a $34\times34\times128$ 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

$$\mathbf{f}_i = \mathbf{x} \otimes \Phi_i \quad \forall i \in \{1, 2, \ldots, k\}, \tag{6.1}$$

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

$$\mathbf{a}_i = \max(0, \mathbf{f}_i), \tag{6.2}$$

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

[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.

[3]For an intuitive explanation of these operations, we refer the reader to Appendix D.

operation can be defined as

$$\mathbf{p}_i = \text{downsample}_2(\sqrt[10]{(\mathbf{a}_i)^{10} \otimes \mathbf{1}_{7\times 7}}), \tag{6.3}$$

where $\mathbf{1}_{7\times 7}$ 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 a soft-max pooling in the sense of [65]. Finally, after these three operations, the network outputs a visual representation of size $12 \times 12 \times 256$.

6.2 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. 6.1, 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. 6.1, 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

$$\min_{\Phi_i, b_i} \frac{1}{2}||\Phi_i||_2 + C \sum_{s}^{n} \max\{0, 1 - y_{c,s}(\Phi_i \cdot \mathbf{x}_{s\mathcal{N}_i} + b_i)\}, \tag{6.4}$$

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 $\boldsymbol{\Phi}$ and neighborhoods $\boldsymbol{\mathcal{N}}$ 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

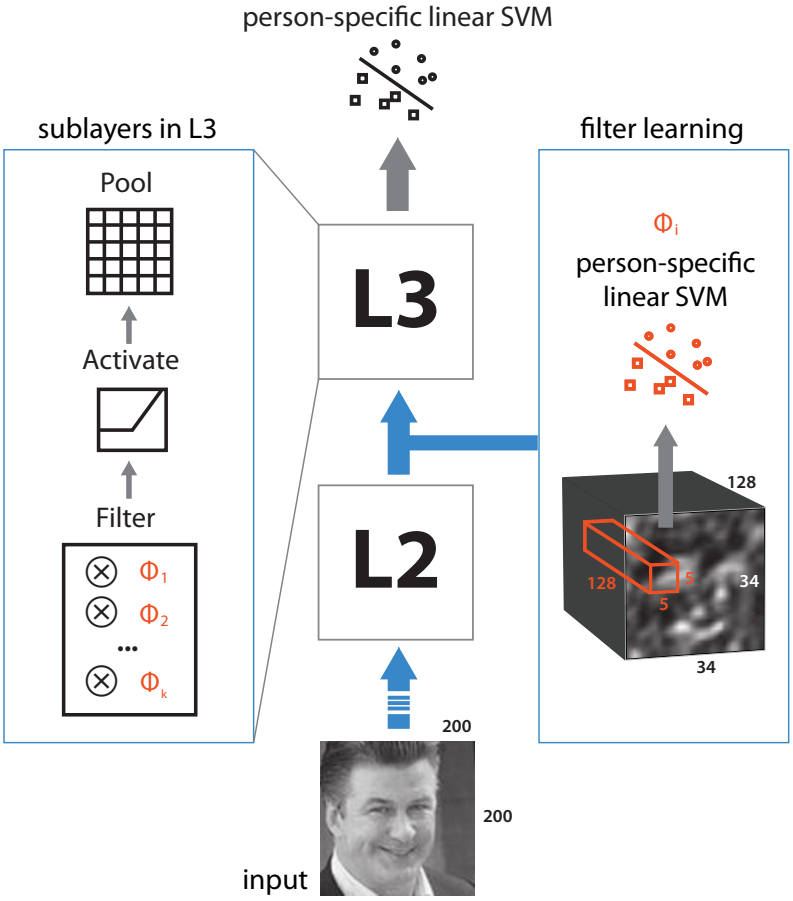


Figure 6.1: Schematic diagram of the L3+ convolutional neural network, detailing the operations (sublayers) of its topmost layer (left panel) and illustrating how data from an input image is sampled in order to learn person-specific filters Φ_i at a given neighborhood (right panel). Early steps of the network [15] are omitted to emphasize the processing steps of our interest.

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 $\mathbf{\Phi}$.[4]

The final component that we add to our filter learning approach is inspired by an ob-

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

servation about the information flow in the network when operating with random filters.[5] Provided that after drawing the weights from a uniform distribution the filters are mean centered, and given the activation function in Eq. 6.2, we observed that, on average, half of the linear filtering responses after activation are set to zero. The enforcement of such "calibrated" sparsity showed to be quite relevant to the network performance in our tests and, therefore, we replicate this behavior by assuming α as the mean response of the person-specific filters on the training set and using an activation function of the form

$$\mathbf{a}_i = \max(0, \mathbf{f}_i - \alpha) \tag{6.5}$$

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. 6.2.

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, 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. 5.2), 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.

6.3 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*.

[5] As random filters are known to perform surprisingly well in the general class of convolutional neural networks [73, 74], we found valuable to investigate some of their characteristics.

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. 5.4. We also consider filters of the type proposed by Coates and Ng in [75], whose use in large quantities corroborates the notion that good performance can be achieved with inexpensive filter quantization and encoding techniques [75]. 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 [75].

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.

Regardless the filter type and the number of filters, all other operations and architectural parameters in the third layer are made as presented in Sec. 6.1. The only exception is the activation function, where Eq. 6.2 is replaced by Eq. 6.5 in cases where the filters are learned, *i.e.*, when using person-specific and K-means-like filters.[6] In these cases, α is determined as explained in Sec. 6.2. 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, on top of all the resulting visual representations, hard-margin person-specific linear SVMs are trained with $C = 10^5$.

The experimental results are presented in Table 6.1 and Fig. 6.2 for the methods in identification mode and in Fig. 6.3 in verification mode. In accordance to all results presented in this thesis, we report the mean accuracy and standard error over ten dataset splits (see Sec. 3.2).

[6] As advocated in [75], this is in fact a very good encoding scheme to use with large quantities of K-means-like filters.

Table 6.1: Results 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

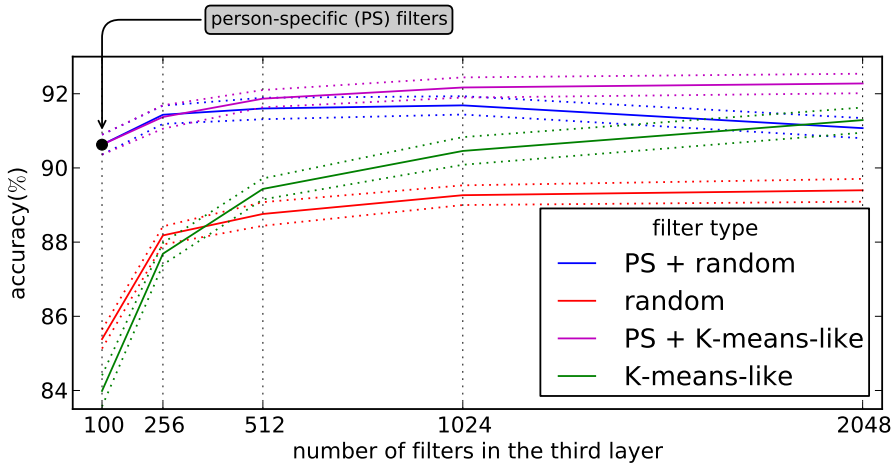


Figure 6.2: Plot of the results obtained in the identification scenario with intervals corresponding to standard errors.

From Table 6.1 and Fig. 6.2, 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.[7] 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

[7]Note that the performance achieved with 256 random filters in Table 6.1 is the same as in Table 5.1 when we use the RAW L3+ representation. In fact, they are the same method.

Table 6.2: Identification results on PubFig83 available in the literature.

images	Pinto *et al.* [16] CVPRW'11	Chiachia *et al.* [23] BMVC'12	Bergstra *et al.* [76] ICML'13	Carlos *et al.* [77] FG'13	This thesis
unaligned	85.22±.45	–	86.50±.70	–	–
aligned	87.11±.56	88.75±.26	–	73.47±.41	92.28±.28

in both random and 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 [75], 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 [75].

For comparison purposes, in Table 6.2 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.* [16] gives us an idea about how these setups compare.

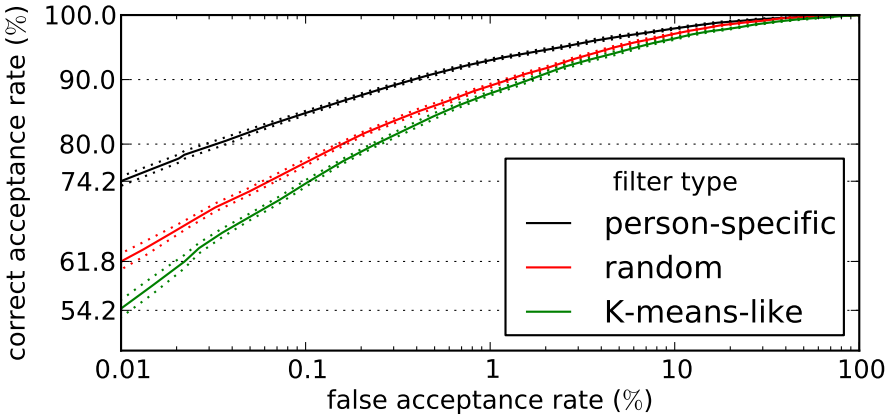


Figure 6.3: Comparison of the methods in the verification scenario. When the system is set to wrongly accept only 0.01% of the test cases, we can observe a dramatic improvement in performance, which is especially relevant to high security applications.

In Fig. 6.3, we present the performance of the methods 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 found reasonable to compare methods only with this number of filters. It is possible

to observe that the use of person-specific filters results in a great improvement in correct acceptance, especially when the system is set to wrongly accept only 0.01% of the test cases. 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 thesis — but also readily applicable in the verification scenario.

Chapter 7

Conclusion and Future Work

In this thesis, we presented three 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 in both constrained and unconstrained settings, achieving not only insight into the value of person-specific representation, but also state-of-the-art results.

To tackle the challenging problem of unconstrained face recognition, we first introduced 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 proposed an original framework that uses SVMs to learn "deep" person-specific models 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.

An important direction to this line of research is to assess the boundaries within

which this hypothesis holds true. For example, in Appendix E, we show that the lack of diversity in learnable face images — being this diversity an assumption in human face familiarity [8] — impairs our approach in a recognition scenario unarguably easier than PubFig83. Determining the extent of applicability of our approach, and continuing to explore the wide range of possible techniques for learning person-specific representations will be a promising area for future research.

In the short term, we also envision the extension of our deep person-specific models to other problem-domains, in which case they will be *class-specific*. Indeed, the notion of learning "stackable" layer-wise representations with maximum-margin classifiers — that usually leads to classifiers with strong generalization abilities — might be interesting to explore in problems where training samples (compared to the problem difficulty) are scarce, *i.e.*, most unconstrained computer vision problems that we currently deal with, such as face or object recognition. In situations where we have more filters than we would like to use, we also plan to use Adaboost [78] or some related method to select them, similar to the approach of Berg and Belhumeur to select a good combination of SVM classifiers [30]. Finally, we can also investigate a potential compromise between unsupervised and supervised layer-wise filter learning. Semi-supervised filter learning in the sense of [79] is also a possibility.

Bibliography

[1] M. Turk and A. Pentland, "Face Recognition using Eigenfaces," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1991. 1, 2, 5, 21

[2] L. Wiskott, J.-M. Fellous, N. Krüger, 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. 1, 2, 6

[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. 1, 2, 6, 14

[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. 1, 2, 5, 21

[5] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear Image Analysis for Facial Recognition," in *IEEE Intl. Conf. on Pattern Recognition*, 2002. 1, 21

[6] L. Wolf, T. Hassner, and Y. Taigman, "The One-Shot Similarity Kernel," in *IEEE Intl. Conf. on Computer Vision*, 2009. 1, 2

[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. 1, 2

[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. 1, 23, 40

[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. 1, 18

[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. 1

[11] 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. 1, 8, 12

[12] R. Chellappa, P. Sinha, and P. Phillips, "Face Recognition by Computers and Humans," *IEEE Computer*, vol. 43, no. 2, pp. 46–55, 2010. 1, 3, 18

[13] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," in *IEEE Intl. Conf. on Computer Vision*, 1999. 2

[14] 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. 2

[15] 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. 2, 24, 30, 31, 33, 55, 56

[16] 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. 2, 3, 7, 8, 12, 19, 23, 37, 51, 58, 59

[17] 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. 2

[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. 3

[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. 3

[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. 3

[21] X. Chen, P. J. Flynn, and K. W. Bowyer, "Visible-light and Infrared Face Recognition," in *ACM Workshop on Multimodal User Authentication*, 2003. 3, 7, 10, 58, 59

[22] G. Chiachia, A. X. Falcão, and A. Rocha, "Person-specific Face Representation for Recognition," in *IEEE Intl. Joint Conf. on Biometrics*, 2011. 3

[23] 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. 3, 19, 37

[24] T. Poggio, "The Computational Magic of the Ventral Stream," *Nature Precedings*, 2011, doi:10.1038/npre.2012.6117.3. 4, 31

[25] 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. 4, 31

[26] T. Kanade, "Picture Processing by Computer Complex and Recognition of Human Faces," Ph.D. dissertation, Kioto University, 1973. 5, 6

[27] S. Z. Li and A. K. Jain, "Introduction," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds. Springer, 2011, ch. 1, pp. 1–15. 5

[28] W. Zhao, R. Chellapa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, December 2003. 5

[29] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. 6

[30] T. Berg and P. N. Belhumeur, "Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification," in *British Machine Vision Conference (BMVC)*, 2012. 7, 40

[31] 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. 7

[32] M. Günther, A. Costa-Pazo, C. Ding, E. Boutellaa, and G. C. et al., "The 2013 Face Recognition Evaluation in Mobile Environment," in *IEEE Intl. Conf. on Biometrics*, 2013. 7, 58, 63

[33] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, 2000. 7

[34] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary Face Recognition Grand Challenge Results," in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 15–24. 7

[35] P. J. Phillips, P. Grother, and R. Micheals, "Evaluation Methods in Face Recognition," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds. Springer, 2011, ch. 21, pp. 550–574. 7

[36] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 Large-Scale Experimental Results," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831–846, 2010. 7

[37] P. J. Phillips, J. R. Beveridge, B. Draper, G. H. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. A. Sahibzada, and S. Weimer, "An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem," in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2011. 7, 8

[38] 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 ICME Workshop on Hot Topics in Mobile Multimedia*, 2012. 7, 58

[39] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild," Univ. of Massachusetts, Amherst, Tech. Rep., 2007. 8, 30

[40] P. Viola and M. Jones, "Robust Real-time Object Detection," *Intl. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. 8

[41] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2008. 8

[42] 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. 12

[43] A. Vashist, Z. Zhao, A. Elgammal, I. Muchnik, and C. Kulikowski, "Discriminative Patch Selection using Combinatorial and Statistical Models for Patch-Based Object

Recognition," *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2006. 14

[44] M. S. Sarfraz and M. Khan, "A Probabilistic Framework for Patch based Vehicle Type Recognition," in *Intl. Conf. on Computer Vision Theory and Applications*, 2011. 14

[45] 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. 20, 25

[46] H. Wold, "Partial Least Squares," *Wiley Encyclopedia of Statistical Sciences*, vol. 6, pp. 581–591, 1985. 21, 22

[47] 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. 21, 22, 51

[48] 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. 21, 22

[49] 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. 21, 22, 23, 25

[50] 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. 22

[51] 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. 22

[52] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face Spoofing Detection through Partial Least Squares and Low-Level Descriptors," in *IEEE Intl. Joint Conf. on Biometrics*, 2011. 22

[53] 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. 22

[54] 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. 22

[55] P. Geladi, "Notes on the History and Nature of Partial Least Squares (PLS) Modelling," *Wiley Journal of Chemometrics*, vol. 2, no. 4, pp. 231–246, 1988. 22

[56] H. Abdi, "Partial Least Squares Regression and Projection on Latent Structure Regression," *Wiley Int. Reviews: Computational Statistics*, vol. 2, no. 4, 2010. 22

[57] 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. 24

[58] J. Wright and G. Hua, "Implicit Elastic Matching with Random Projections for Pose-variant Face Recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 24

[59] B. Schölkopf, A. J. Smola, and K. R. Müller, "Kernel Principal Component Analysis," in *Springer Intl. Conf. on Artificial Neural Networks*, 1997. 25

[60] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K. R. Mullers, "Fisher Discriminant Analysis with Kernels," in *IEEE Neural Networks for Signal Processing Workshop*, 1999. 25

[61] R. Rosipal and L. J. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space," *Journal of Machine Learning Research*, vol. 2, pp. 97–123, 2001. 25

[62] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011. 25

[63] 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. 30, 55

[64] 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. 30, 55

[65] M. Riesenhuber and T. Poggio, "Hierarchical Models of Object Recognition in Cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999. 30, 32

[66] 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. 30, 57

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012. 30, 55

[68] 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. 30, 55

[69] C. F. Cadieu, H. Hong, D. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo, "The Neural Representation Benchmark and its Evaluation on Brain and Machine," *Intl. Conf. on Learning Representations*, 2013. 30

[70] 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. 30, 57

[71] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-wise Training of Deep Networks," in *Advances in Neural Information Processing Systems*, 2007. 30, 57

[72] 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. 30

[73] 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. 34, 56

[74] 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. 34

[75] A. Coates and A. Ng, "The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization," in *Intl. Conf. on Machine Learning*, 2011. 35, 37, 56

[76] 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. 37

[77] 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. 37

[78] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*. Springer-Verlag, 1995, pp. 23–37. 40

[79] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *Intl. Conf. on Machine Learning*, 2008. 40, 57

[80] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in Cognitive Sciences*, vol. 11, pp. 333–341, 2007. 55

[81] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, vol. 73, pp. 415–34, 2012 Feb 9 2012. 55

[82] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex." *The Journal of physiology*, vol. 148, pp. 574–591, 1959. 55

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536+, 1986. 57

[84] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997. 62

Appendix A

Running Example of our Preliminary Evaluation

In Figure A.1, we provide a running example of the identification scheme adopted in our preliminary evaluation. From top to bottom, the diagram starts with the person-specific representation of the gallery images $\mathbf{G}_{c,m}$ into the feature spaces $\mathbb{S}_c$, where c denotes the modeled persons in the training/gallery set, and m indicates which of the multiple samples of the person is being considered. In this example, we have two persons modeled with two gallery samples each. Thus, $c = \{1, 2\}$ and $m = \{a, b\}$. After representing the gallery samples in each person-specific feature space, we obtain samples $\mathbf{G}^c_{c,m}$, which means $\mathbf{G}_{c,m}$ represented in the feature space modeled for person c.

In order to recognize a probe $\mathbf{P}$, we represent it in each feature space $\mathbb{S}_c$, and the resulting $\mathbf{P}^c$ samples are correspondingly matched to the gallery. In this example, we match $\mathbf{P}^1$ to the samples $\mathbf{G}^1_{c,m}$ and $\mathbf{P}^2$ to the samples $\mathbf{G}^2_{c,m}$. The matchings are then ranked according to the dissimilarities and an identity is established by each *nearest-neighbor* classifier. Here we have two classifiers, one for $\mathbb{S}_1$ and one for $\mathbb{S}_2$. Finally, a voting scheme is done by considering the decisions of the classifiers, and the person in the gallery which has the most votes is taken as the probe identity.

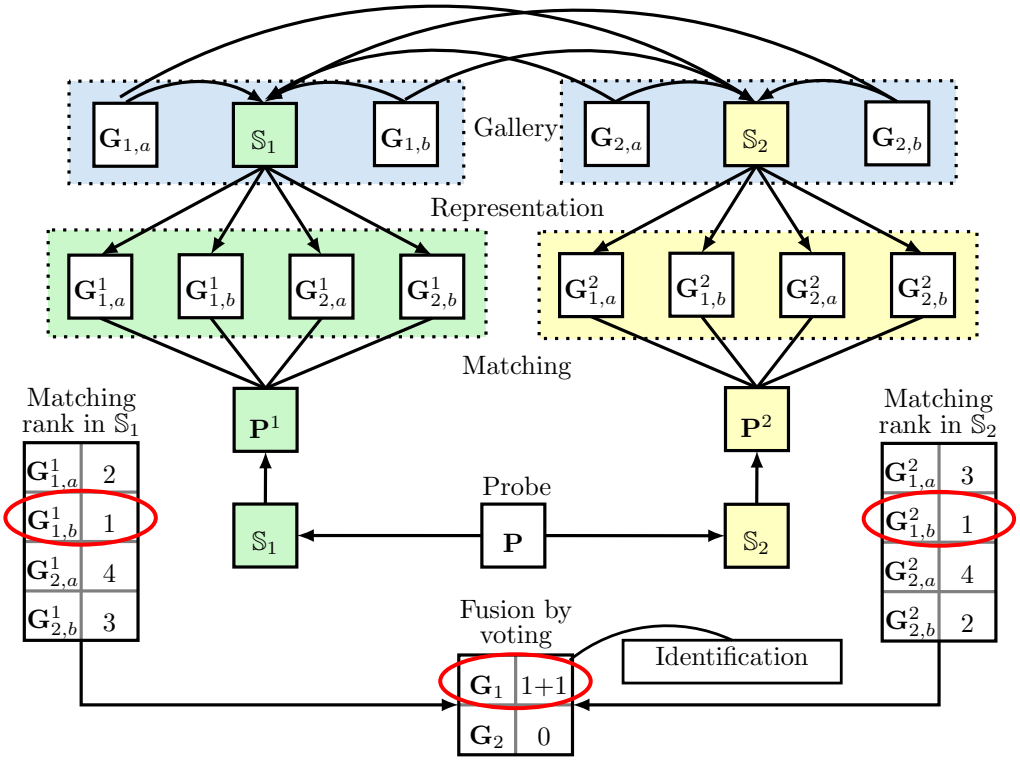


Figure A.1: Illustration of the identification scheme adopted in our preliminary evaluation, considering $c = \{1, 2\}$ persons in the training/gallery set, with $m = \{a, b\}$ samples each. From top to bottom, the diagram starts with the person-specific representation of the gallery $\mathbf{G}_{c,m}$ into the feature spaces $\mathbb{S}_c$. Such a representation results in $\mathbf{G}^c_{c,m}$, which means $\mathbf{G}_{c,m}$ represented in the feature space c. Given a probe $\mathbf{P}$, its representations $\mathbf{P}^c$ are correspondingly matched to the gallery. The matchings are then ranked and an identity is established by each classifier. Finally, a voting scheme is done by considering their decisions, and the person which has the most votes is taken as the probe identity.

Appendix B

Additional Results on Person-Specific Subspace Analysis

In addition to the core results on subspace analysis presented in Chapter 5, we also evaluated the approach on two additional visual representations and one additional dataset. The first additional representation is named HT-L3-1st and was taken from [16]. It can be thought of as a visual model slightly different from the L3+ model presented in Chapter 5. The second additional representation is, in turn, a blend of local binary patterns (LBP), histogram of oriented gradients (HOG), and Gabor wavelets (LBP+HOG+Gab), and was taken from [47] in order to test our method with a representation on which partial least squares (PLS) was already known to perform well.

The additional face dataset that we consider is Facebook100, which is similar in spirit to PubFig83. Indeed, a remarkably linear relationship between performance achieved on each set by a variety of algorithms has been reported in [16]. Both sets enable the investigation of face recognition methods where a considerable number of natural face images from the individuals is available. As advocated in Chapter 1, Facebook100 reflects the exact scenario on which learning person-specific representations is especially attractive, *i.e.*, social media. The reason why we omitted Facebook100 from our main results is because this dataset is private [16].

As we can observe in Table B.1, the results on PubFig83 with the additional representations are similar to the results reported in Chapter 5. Again, the only face subspace in which we could consistently get better results than RAW is PS-PLS.

For the Facebook100 dataset, we present in Table B.2 the performance obtained with the most competitive method of each category considered in Tables 5.1 and B.1. The results are similar to the ones obtained on PubFig83, where PCA representations performed most like RAW, LDA did better in V1-like+, and PS-PLS performed best across all representations.

Table B.1: Comparison of different face subspace analysis techniques on two additional visual representation applied on the PubFig83 dataset. In all cases, the final identities are estimated by linear SVMs.

Models	HT-L3-1st	LBP+HOG+Gab
RAW	87.66±.29	82.63±.28
Multiclass Unsupervised		
RP	85.61±.37	75.07±.37
PCA	87.50±.28	82.44±.34
Multiclass Supervised		
LDA	85.72±.33	83.40±.22
PLS	86.63±.35	83.02±.26
Person-Specific		
PS-PCA	52.65±.62	33.02±.39
PS-PLS	88.75±.26	85.42±.29

Table B.2: Comparison of different face subspace analysis techniques in the Facebook100 dataset.

Models	V1-like+	HT-L2-1st	HT-L3-1st
RAW	79.96±.19	85.81±.29	88.89±.25
PCA	79.81±.18	85.70±.29	88.88±.25
LDA	81.04±.29	83.07±.26	87.25±.29
PS-PLS	81.53±.25	86.84±.19	89.70±.25

Taken together, these results strengthen the use of person-specific subspace analysis via PLS in the unconstrained familiar face identification setting.

Appendix C

Scatter Plots from Different Subspace Analysis Techniques

In Chapter 5, we proposed a person-specific application of partial least squares (PS-PLS) to generate per-individual subspaces of familiar faces. By means of a straightforward evaluation methodology, we compared different subspace analysis techniques for modeling the problem. Extending Fig. 5.2(a), where we showed a scatter plot of training and test samples projected onto the first two projection vectors of a PS-PLS model, in this appendix we show these projections for three other subspace analysis techniques evaluated in our experiments, namely PCA, LDA, and PLS. As in Fig. 5.2(a), Adam Sandler's samples are in red.

The overall distribution of the points is in accordance to our expectations, where samples are spread out in PCA subspace, are more concentrated, apart with respect to the other classes, and Gaussian shaped in LDA subspace, and are also apart but less concentrated in PLS and PS-PLS. Due to its person-specific nature, we can observe a clear better separation of the samples in PS-PLS.

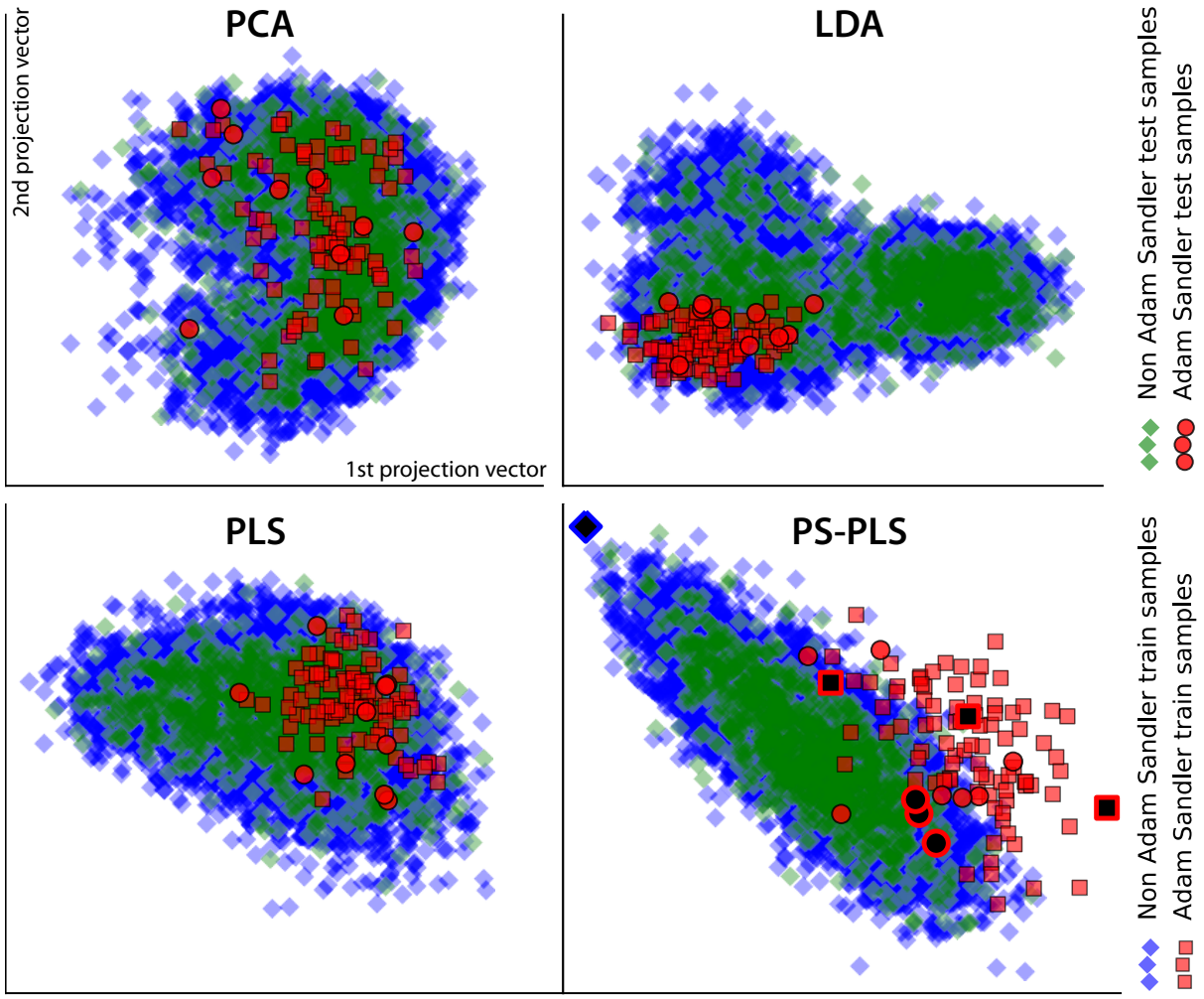


Figure C.1: Visualization of the training and test samples projected onto the first two projection vectors of each model. All models were obtained from the same training/test split using the V1-like+ representation.

Appendix D

Overview on Deep Visual Hierarchies

Humans have an impressive ability in recognizing faces, vehicle types, and a profusion of other objects without much effort. Fortunately, neuroscience has provided a number of important directions to computer vision researchers in their attempt to artificially reproduce these abilities. These directions come not only from recent research suggesting, for example, that the ventral visual stream of primates consists of a feedforward cascaded hierarchy that gradually "untangles" information about objects in the scene [80, 81]. These directions come also from seminal works like the one from Hubel and Wiesel [82], stating that the visual cortex is made by cells that are sensitive to small regions of the input space, called *receptive fields*, and that these cells are of two types; one that responds maximally to specific stimulus, known as *simple* cells, and another that account for local invariance to the exact position where the stimulus occurred, known as *complex* cells.

In fact, computer vision and machine learning researchers have been taking advantage of these findings for a long time. In the early 1980s, for example, Fukushima [63] proposed *neocognitron*, a self-organizing artificial neural network inspired in the cell types of Hubel and Wiesel [82], designed to extract robust signatures from visual patterns. With the same inspiration, Lecun *et al.* [64] proposed *convolutional neural networks* in late 1980s along with a procedure to discriminatively train them via backpropagation. Indeed, many other contributions have been made to computer vision literature in the past decades following the same principle of learning a visual hierarchy capable of representing high level concepts straight from image pixels.

Modern approaches like [15, 68, 67] often employ a sequence of well defined operations such as (i) linear *filtering* followed by nonlinear *activation* – mimicking simple cell behavior, (ii) local *pooling* – mimicking complex cell behavior, and (iii) local *normalization* – attempting to model competitive interactions among neurons. These operations can be thought of sublayers of a feedforward network with many *layers*. In Fig. D.1, we present the architecture of one hypothetical layer. Note in red the receptive fields of each

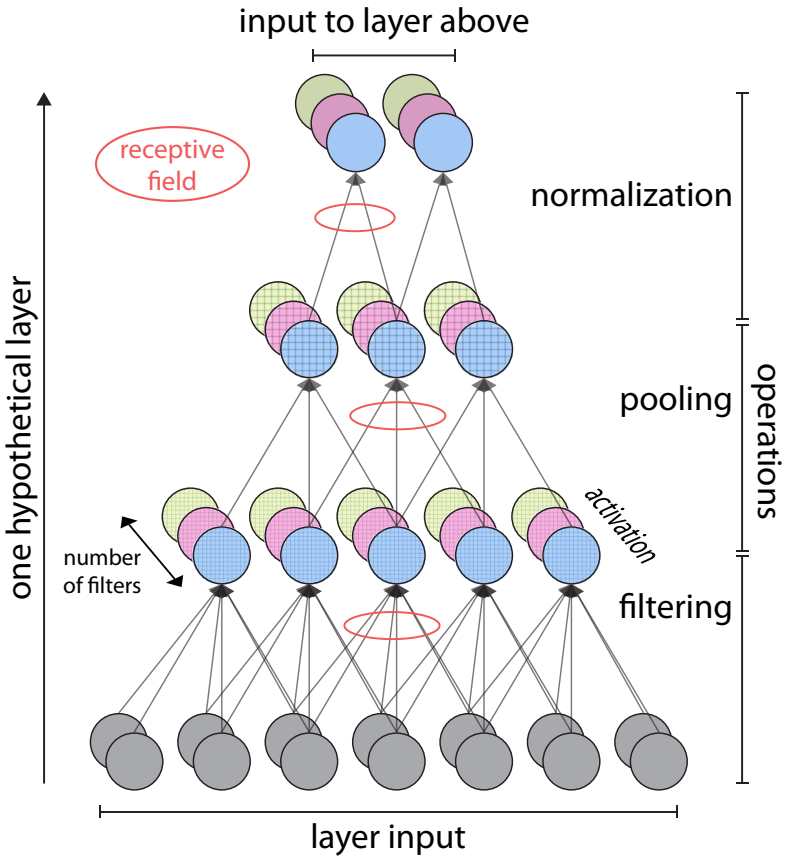


Figure D.1: Architecture of one hypothetical layer using three well-known biologically-inspired operations.

operation and how spatial resolution decreases as information flows in the network. The intuition of using deep visual hierarchies is that, by "correctly stacking" many of these biologically-inspired nonlinear operations, increasingly complex abstractions will emerge at top layers.

There are two important — and strongly related [75] — aspects to consider when designing deep visual hierarchies. One is how to determine the network architecture and the other is how to determine which stimuli the filters will maximally respond to.

The architectural details are usually referred to as *hyperparameters*, and define which operations should be involved, what are their receptive field, in what order should they be applied, how many layers should be used, how many filters should be considered in each layer, *etc.* While in many cases these hyperparameters are presented as tangential to the algorithm, recent work has shown that they are crucial to the method's performance [15, 75, 73].

The method employed to determine which stimuli the filters will maximally respond

to is often called the *training* procedure. One of the most common training procedures is backpropagation [83], which adjusts all filters in the network by minimizing the difference — and propagating it backwards — between the obtained and the desired representation in the topmost layer of the hierarchy. One problem of learning all filters at the same time is the huge amount of examples required. In Fig. D.1, one can think that a filter *weight*, *i.e.*, a parameter, needs to be learned for each arrow between the bottom sublayer and the sublayer above. It is not uncommon to have a network with tens of millions of such parameters. Therefore, this network would likewise require tens of millions of training samples in order to learn filters that would generalize the network behavior to new samples.

However, Hinton *et al.* [66] showed in 2006 that a particular form of probabilistic graphical model called *restricted Boltzmann machines* can be trained and stacked in a greedy manner, so that a bound on the probability of representing well the training data is increased at each layer. Since then, the term *deep learning* has been used to denote various other methods following the same principle of learning filters one layer after the other [71, 70, 79]. A key advantage of this layer-wise learning strategy is the alleviation of the aforementioned over-parameterization problem.

Appendix E

Scoring Best in the ICB-2013 Competition and the Applicability of Our Approach in the MOBIO Dataset

We were recently fortunate by scoring best in a competition on mobile face recognition organized as part of the prestigious *International Conference on Biometrics* [32]. This competition was carried out on the MOBIO database [38], which can be considered a relatively unconstrained dataset. In fact, in terms of user collaboration, the MOBIO dataset can be seen in-between the UND [21] and the PubFig83 [16] datasets. Most importantly, however, is the fact that this dataset also reflects a timely use case, which is face recognition in mobile devices.

In this appendix, we first present in Sec. E.1 the MOBIO dataset as well as its relevant aspects. Then, in Sec. E.2, we describe the performance measures that were used to evaluate the competitors. With this information, in Sec. E.3 we are able to report details about our winning method. After that, a thorough evaluation of our person-specific representation learning approach on this dataset is presented in Sec. E.4. Final remarks, lessons learned, and directions obtained with this experience are presented in Sec. E.5.

E.1 The MOBIO Dataset

The MOBIO dataset has 152 people with a female-male ratio of nearly 1:2 (100 males and 52 females). It is the result of an international collaboration, in which images from six institutions of five different countries were recorded in 12 distinct occasions for each

individual.[1] The dataset can be considered challenging in the sense that images were acquired without control over factors such as illumination, facial expression, and face pose. Moreover, in some cases, only parts of the face are visible.

For the competition, 150 out of the 152 individuals were considered. Based on the gender of the individuals, the evaluation protocol is split up into *female* and *male*. Still, for the sake of fairness, individuals in the dataset are divided into three subsets, namely the *training* set, the *development* set, and the *evaluation* set.

The training set has 50 individuals — 13 females and 37 males — with 192 images each and can be used for any purpose to aid the systems, from learning subspace models to leveraging score normalization. In addition, this is the only subset where gender can be combined according to the participant's needs.

The development set has 42 individuals — 18 females and 24 males — and can be used to tune the hyperparameters of the algorithm, *e.g.*, the number of projection vectors while learning subspaces, which similarity measure to use, *etc.* For each person in this set, there are five gallery images — which in the context of our method we call training images — and 105 test images. For each gender, participants were asked to submit a score file containing one similarity score between each test sample and each gallery person. For example, the score file related to the female protocol in the development set must contain $18\times(18\times105)$=34,020 similarity scores.

The evaluation set, in turn, is used to assess the final system performance. It has 58 individuals — 20 females and 38 males — with samples arranged in exactly the same way as the development set, *i.e.*, five gallery (or training) images and 105 test images. In order to disallow participants to optimize parameters on the evaluation set, test file names were anonymized and shuffled. Luckily, after the competition, the organizers released the test files with their original names. This way, we are now able to carry out experiments on our own and compare the performance of new approaches with the competition numbers.

In Fig. E.1, we present training and test images of four individuals in the evaluation set. While we can clearly see variation in pose, expression, and illumination, we can also observe that the individuals are — to some extent — collaborating with the image acquisition process. This is the reason we regard the MOBIO dataset as representing an intermediate scenario in terms of user collaboration (see Fig. 2.2). It is far from being as constrained as UND [21], but at the same time is not as "wild" as PubFig83 [16]. More importantly, however, is to observe the difference in appearance among the training and the test images. In fact, we can see that the five training images of each individual look quite similar. While this is a natural consequence from the fact that these images were recorded in the same session, this considerably diminishes the discriminative power of learning techniques operating on them. Moreover, such homogeneity in appearance is

[1] In particular for the competition, all images available were captured by mobile phones.

Figure E.1: Training and test images from the MOBIO evaluation set. We can observe that the dataset represents an intermediate recognition scenario in terms of user collaboration (see Fig. 2.2). It is not as constrained as UND (Fig. 3.1), but at the same time is not as "wild" as PubFig83 (Fig. 3.2). More importantly, however, is to observe the difference in appearance among the training and the test images. In fact, we can see that the five training images of each individual look quite similar. This is not aligned to the notion of familiarity that we pursue in this thesis, and considerably diminishes the discriminative power of learning techniques operating on them.

not aligned to the notion of familiarity that we attempt to approach in this thesis with PubFig83.

E.2 Performance Measures

During the competition, the systems were analyzed in verification mode and the performance metrics adopted by the organizers are based on compromises between false acceptance (FAR) and false rejection (FRR) rates. Indeed, what determine the relationship between these two measures is a threshold θ above which the system predicts that the matching images are from the same individual. By increasing θ, we decrease FAR and increase FRR. Conversely, by decreasing θ, we increase FAR and decrease FRR.

The main performance metrics used on the competition are actually known as equal error rate (EER) and half total error rate (HTER). In particular, EER was adopted to measure performance in the development set and HTER to measure performance in the evaluation set. For this purpose, a θ_{dev} is first computed to measure the EER on the development set and then is used to measure the HTER on the evaluation set. Formally,

$$\begin{aligned} \theta_{\text{dev}} &= \arg\min_{\theta} |\text{FAR}_{\text{dev}}(\theta) - \text{FRR}_{\text{dev}}(\theta)| \\ \text{EER} &= \frac{\text{FAR}_{\text{dev}}(\theta_{\text{dev}}) + \text{FRR}_{\text{dev}}(\theta_{\text{dev}})}{2} \\ \text{HTER} &= \frac{\text{FAR}_{\text{eval}}(\theta_{\text{dev}}) + \text{FRR}_{\text{eval}}(\theta_{\text{dev}})}{2} \end{aligned} \tag{E.1}$$

where the subscripts "dev" and "eval" denote values computed on the development and on the evaluation set, respectively.

As mentioned in Sec. E.1, both development and evaluation sets are split up into female and male subsets, and the systems are independently evaluated in each gender. For a given gender, θ_{dev} is obtained from the development set and used in the evaluation set of the same gender. Therefore, the main performance metrics considered in the competition were two EER values — one for each gender — and, likewise, two HTER values.

E.3 Our Winning Method

We started designing our system by aligning the images with the eye positions provided by the organizers, as we did for the UND and the PubFig83 datasets. Naturally, the visual representation of our choice was L3+, as we observed throughout the thesis that it achieves superior performance. By the time that the competition was running, there was a rule stating that *no parameter* could be learned on the evaluation set. Since this

Table E.1: Our initial systems. We can observe that learning a subspace model with LDA on the training set was fundamental in performance. Experiment A is the system whose scores we first submitted to the competition organizers, while experiment B is identical to A, but is does not use LDA.

system	LDA	matching	EER on dev. set		HTER on eval. set	
			female	male	female	male
A	yes	1-NN	**5.026**	**4.405**	**11.724**	**7.282**
B	no	1-NN	11.852	10.635	19.732	14.645

forbade us from learning person-specific models from gallery images, we had to recast our face recognition approach.

In the short timeframe that we had to put together a system meeting these conditions, we could experiment a few ideas before submitting our score files. Given that the training set was the only set that we could perform learning tasks, and that individuals in this set were different from the individuals in the development and the evaluation sets, we regarded this problem as a *transfer learning* problem.

In a first attempt, we tried to use deep person-specific filters learned from individuals in the training set to represent individuals in the other two sets. In accordance to the procedure presented in Chapter 6, we learned 100 person-specific filters for each individual (out of 50) and then, as an extension, we used AdaBoost [84] to select an optimal subset of filters performing best in the development set.

Another idea that occurred to us to leverage L3+ in this scenario was to perform *multiclass supervised* subspace learning on the training set, using the techniques of this type that we had previously evaluated in Chapter 5, namely multiclass partial least squares (PLS) and linear discriminant analysis (LDA). In this attempt, different from what we observed in Chapter 5 — where LDA and multiclass PLS performed equivalently — LDA showed to perform much better than multiclass PLS in transferring discriminative structure from the training set to the development set.

It turned out that, from the few ideas that we evaluated, our most effective approach by the submission deadline consisted of the ensemble of standard L3+ visual representations, LDA subspace analysis performed on the training set, and nearest neighbor predictions — considering the maximum score obtained while matching each test image to the five gallery images of each individual. This approach is presented in Table E.1 as system A, whose scores we first submitted to the competition organizers. Due to the importance of LDA throughout our experiments with the MOBIO dataset, we also present here system B, which is identical to A except that it does not use LDA.

A few weeks after submitting system A, we received a manuscript from the organizers with the description and performance of all systems submitted to the competition. From

Table E.2: Results obtained with the replacement of nearest neighbor predictions by one-versus-all linear SVMs. As we can observe, the use of linear SVMs did improve performance. However, while the performance of system D over B was substantially better, learning linear SVMs on the LDA subspace did not boost performance as greatly, as we can observe by comparing C with A.

			EER on dev. set		HTER on eval. set	
system	LDA	matching	female	male	female	male
A	yes	1-NN	5.026	4.405	11.724	7.282
B	no	1-NN	11.852	10.635	19.732	14.645
C	yes	linear SVM	**4.709**	**3.492**	**10.833**	**6.210**
D	no	linear SVM	7.196	6.786	15.655	8.747

that document, we first realized that our system had superior performance.[2] In addition, we also realized that a few other participants actually learned a discriminative binary model for each individual. They did so by considering *gallery images of a single individual as positive samples* and *images in the training set as negative samples*, repeating this process for all individuals.

This called our attention because, in our opinion, they were actually learning thousands of parameters from the evaluation set (even though using only gallery images), something that was clearly forbidden according to the guideline. Our reaction was to immediately replace our nearest neighbor classifier by a one-versus-all linear SVM for each individual, training them in the same way. As we can observe in Table E.2, the use of linear SVMs did boost the performance of our systems. However, while the performance of system D over B was substantially better, learning linear SVMs on the LDA subspace did not boost performance as greatly, as we can observe by comparing system C with A.[3]

In the end, the organizers accepted our arguments about the fact that the competition guideline was misleading, and allowed us to send them new score files from our slightly better system C, which ended up being the best performing single system of the competition [32].

The little boost in performance while using SVMs instead of nearest neighbor classification was somehow surprising to us. In PubFig83, for example, when we replace one by the other, the difference in performance is quite considerable, of over 30% in favor

[2]It is worth nothing that our system was considered by the organizers as belonging to the category of *simple systems*, in which predictions are made by a single classification engine. The other category, known as *fusion systems*, is related to systems that combine many visual representations with various classification engines to produce final matching scores. In any case, if we take the mean between HTERs on the evaluation set, our single system performed best than all other systems [32].

[3]It is also important to observe that all performance comparisons carried out during the competition were solely based on female and male EERs obtained on the development set. As mentioned in Sec. E.1, test file names in the evaluation set were encrypted at that time.

of SVMs in terms of identification rate. A more detailed analysis of the data, however, enables us to conjecture two possible reasons for this fact. First, as mentioned in Sec. E.1, the gallery/training images of each individual are quite homogeneous (see Fig. E.1), and this may not allow discriminative learning tasks to capture most informative features based on them. Second, the appearance of the 50 individuals in the training set — to which we train linear SVMs against — may not represent well the appearance of other individuals in either the development and the evaluation sets, which may also explain discriminative models performing under our expectations.

E.4 Learning Person-Specific Representations

Given the fact that we did not have the chance to learn person-specific representations by the time of the competition, in this section we present an evaluation of the two best performing representation learning techniques proposed in this thesis — namely person-specific partial least squares (PS-PLS) and deep person-specific models (Deep PS) — on the MOBIO dataset. As performance considering nearest neighbor (1-NN) and SVM predictions were close in systems A and C (Fig. E.2), in this section we always report results considering both prediction schemes.

We first evaluate how PS-PLS and Deep PS compare with LDA, which can also be seen as a representation learning method. In Table E.3, the resulting systems are presented as E, F, G, and H. We can clearly observe that neither PS-PLS nor Deep PS could beat systems A and C. While this is in opposition to our experiments on PubFig83 and Facebook100 (Chapters 5 and 6 and Appendix B) — where PS-PLS outperformed LDA and the advantage of Deep PS was conclusive — it also strengthen the conjecture presented in the previous section that (i) gallery/training images in the MOBIO dataset are quite homogeneous and that (ii) individuals in the training set may not represent well the appearance of other individuals in the development and the evaluation sets. Both of these issues may have impaired the process of learning person-specific representations in systems E, F, G, and H.

Another point that is clear to observe from Table E.3 is that LDA appears to be central in obtaining good performance in this dataset. Therefore, given that PS-PLS models can be learned from any kind of input in $\mathbb{R}^d$, we decided to further evaluate the construction of person-specific models with PS-PLS over LDA features. Moreover, we decided to slightly change the MOBIO protocol by considering a learning scenario closer to the scenario approached on UND, PubFig83, and Facebook100. To this end, we incorporated gallery images from the other individuals of the same set/gender as negative samples in the process of learning PS-PLS models. For example, when we train a person-specific model for a given female (out of 20) in the evaluation set, now we also include gallery images of

Table E.3: Comparison among LDA, PS-PLS, and Deep PS representation learning approaches. Neither PS-PLS nor Deep PS could beat systems A and C. While this is in opposition to our experiments throughout the thesis, it emphasizes that the MOBIO dataset and protocol is adverse for learning person-specific representations.

			EER on dev. set		HTER on eval. set	
system	Rep. Learning	matching	female	male	female	male
A	LDA	1-NN	5.026	4.405	11.724	7.282
C	LDA	linear SVM	**4.709**	**3.492**	**10.833**	**6.210**
E	PS-PLS	1-NN	12.121	12.033	19.509	13.848
F	PS-PLS	linear SVM	8.934	7.268	15.351	10.123
G	Deep PS	1-NN	9.101	5.784	12.395	10.064
H	Deep PS	linear SVM	6.878	4.873	16.454	8.679

Table E.4: Results obtained by incorporating gallery images in the process of learning person-specific representations with PS-PLS on top of LDA features. It is possible to observe from system I that the competition numbers are considerably improved in this setting of the MOBIO database. In addition, here we can also note from system J that representations learned with PS-PLS models consistently resulted in better performance.

	gallery included					
	PS-PLS		EER on dev. set		HTER on eval. set	
system	on LDA	matching	female	male	female	male
A	no	1-NN	5.026	4.405	11.724	7.282
I	no	linear SVM	3.181	2.656	8.377	4.931
J	yes	1-NN	**1.796**	**2.624**	**6.397**	**4.182**
K	yes	linear SVM	3.439	3.531	10.457	5.644

the other 19 females in the negative set.

These experiments gave rise to systems I, J, and K, as presented in Table E.4. While the baseline system A was not affected by this new learning strategy, the other baseline system C (the competition winner) was. Hence, we present system I as its replacement. In this new scenario, it is possible to observe that the competition numbers are considerably improved when comparing I with C (Table E.3). Here we can also note from system J that person-specific representations learned with PS-PLS models consistently resulted in better performance. Moreover, 1-NN predictions outperformed linear SVMs (system K) in this particular scenario. In our opinion, the performance of system J supports our initial guess that the MOBIO dataset — with its homogeneous gallery images and in its original protocol — represents a ill-posed problem for learning person-specific face representations.

E.5 Conclusions

Participating in the ICB2-2013 competition on face recognition was opportune in several ways. First, naturally, having produced a best performing system is the confirmation that we are grounded in good technology. Also, the fact that we could iterate over many ideas and rigorously evaluate them in a timely manner strengthen our work in general.

In addition, we learned a lot by evaluating our methods on the MOBIO dataset, which, even though reflects a presumably easier recognition scenario than PubFig83, has a different image collection process. While PubFig83 has a large pool of diverse gallery images and approach the operational scenario of face recognition in social media, MOBIO has only five homogeneous gallery images for each individual and approach the "one-time enrollment" operational scenario.

After several unsuccessful attempts and a slight modification in the MOBIO protocol, the multitude of systems evaluated in this Appendix culminated in the person-specific-representation-based system J, which was able to beat our ICB-2013 winning method. At the same time that this reassures our claim, it spontaneously suggest that we should carefully consider particularities of the operational scenario to which we target our systems.