Unsupervised Learning of Overcomplete Face Descriptors

Juha Ylioinas, Juho Kannala, Abdenour Hadid, and Matti Pietikäinen
Center for Machine Vision Research
University of Oulu
firstname.lastname@ee.oulu.fi

Abstract

The current state-of-the-art indicates that a very discriminative unsupervised face representation can be constructed by encoding overlapping multi-scale face image patches at facial landmarks. If fixed as such, there are even suggestions (albeit subtle) that the underlying features may no longer have as much meaning. In spite of the effectiveness of this strategy, we argue that one may still afford to improve especially at the feature level. In this paper, we investigate the role of overcompleteness in features for building unsupervised face representations. In our approach, we first learn an overcomplete basis from a set of sampled face image patches. Then, we use this basis to produce features that are further encoded using the Bag-of-Features (BoF) approach. Using our method, without an extensive use of facial landmarks, one is able to construct a single-scale representation reaching state-of-the-art performance in face recognition and age estimation following the protocols of LFW, FERET, and Adience benchmarks. Furthermore, we make several interesting findings related, for example, to the positive impact of applying soft feature encoding scheme preceding standard dimensionality reduction. To this end, making the encoding faster, we propose a novel method for approximative soft-assignment which we show to perform better than its hard-assigned counterpart.

1. Introduction

Face recognition has been investigated for several decades. It remains a major topic and has applications in a plethora of domains, the notable ones being security and surveillance, content-based image retrieval, and human computer interaction. Compared to other traditional biometric traits (e.g. fingerprints and iris), processing with faces has been claimed to be more acceptable among populations [22]. It is also more practical as collecting face images does not call for users to be in physical contact with sensors.

Among the famous early techniques in automatic face recognition are the Eigenfaces [39] and Fisherfaces [6] methods, both of them processing input faces holistically. Later, it was discovered that a more discriminative way of processing is rather locally oriented. Among the notable local methods include those like Local Binary Patterns (LBP) [1] and Scale Invariant Feature Transform (SIFT) [29]. Currently, there are plenty of recent developments [11, 19, 34, 12, 26, 36, 8, 5, 31, 37] thanks to challenging datasets and benchmarks publicly available for fair evaluations and comparisons. In general, the current top-performers are based on supervised convolutional neural networks [37]. However, to train such a variant, one needs hundreds of thousands or even millions labeled face images [37]. The problems related to obtaining and processing such big training data makes unsupervised methods an attractive option.

A characteristic termed overcompleteness has recently been identified as a valuable ingredient in building discriminative face representations [14, 12, 5]. In the given studies, overcompleteness often refers to using large codebooks, but also (in a less strict sense) to overlapping multiscale sampling strategies. To make it more comprehensive, in this paper, we investigate whether overcompleteness is also beneficial in the feature level, which is the case that has not been evaluated in face recognition so far, to our knowledge. Particularly, we are comparing two methods based on Independent Component Analysis (ICA). The first one is for learning a complete basis, whereas the second one for an overcomplete basis. With both of them, the result is a filter bank that we couple with the popular Bag-of-Features (BoF) approach. To elicit the possible improvement brought by overcompleteness, the literature suggests to apply dimensionality reduction [14, 12, 5]. Here we make a thorough investigation of a method based on Principal Component Analysis (PCA). Most importantly, we show that with overcomplete features and larger codebooks before applying PCA, softer methods become necessary in the feature encoding step. As the main obstacle of these softer methods is that they are more time-consuming, we present a fast approximate approach for soft-assignment coding based on adjancency matrices.

We evaluate our method in both face verification and in open-set face identification modes using the Labeled Faces in the Wild (LFW) benchmark with the recently updated and newly added protocols. We further perform cross-database experiments in constrained face identification using the Facial Recognition Technology (FERET) benchmark and in age estimation using the recent Adience benchmark. Finally, we show strong results in all of our experiments compared with the state-of-the-art.

2. Related Work

Since our focus is on unsupervised face descriptors, the following review is limited to studies that are of unsupervised methods in building representations for face and object recognition.

The very first steps in face recognition are face detection and further landmark localization [35]. Then there is face alignment [18] followed by the step for representing faces [19, 37], and finally, learning similarity metrics for matching the chosen representations [8, 7]. Here, we are focused on representing faces that has been shown to be a crucial component for achieving high performances in challenging unconstrained face recognition scenarios. According to the recent state-of-the-art studies, the recipe for constructing discriminative unsupervised face representations for challenging unconstrained conditions should contain the following ingredients: *(i) local feature extraction and coding*, *(ii) high-dimensionality*, and in the end, *(iii) compactness*.

Currently there are two directions for how to extract features and encode them. In detail, features can be either extracted from regions around detected facial landmarks or densely from every pixel location. The recent state-of-the-art studies demonstrate that using highly engineered and general purpose local descriptors such as LBP, HOG, or SIFT, one can construct a very discriminative face representation presuming the descriptor computations are centralized at or around face landmark locations [12, 9]. Without prior information about face landmarks, the descriptor computations are naturally accomplished on every pixel location. Evidently, in order to achieve the best possible result in this direction calls for descriptors that are either more redundant [5] or somehow more sophisticated [19, 14, 26, 37, 24].

High-dimensionality is not a key in itself to solve face recognition, but rather a consequence of the former step where features are extracted and encoded in a robust way. In [12], Chen *et al.* empirically showed that increasing the dimensionality of the representation has a positive impact on the accuracy. Their representation is based on first detecting face landmark locations (such as eyes, nose, and mouth corners) and then describing image regions at those locations using a multi-scale approach. They compared several local descriptors which all ended up to improved recognition accuracy while the feature dimension was increased by using a larger number of landmarks and image scales. Their best descriptor, called HighDimLBP, is based on 27 landmarks and 5-scale image pyramid from which they crop fixed-size image patches (at all landmarks) further encoding them using LBP. In turn, Cao *et al.* [9] demonstrated that by unsupervisedly learning feature encoders by increasing the size of the dictionary has a positive effect on the recognition performance. Their best representation was based on describing aligned facial components using LBP-like sampling and further encoding by using a random projection tree. They named the resulting representation as the Learning-based (LE) descriptor.

The findings of the studies above are in line with many interesting works in object recognition. For example, Coates *et al.* [13] showed that instead of choosing the feature, more weight should be put on finding out the most optimal extraction strategy, meaning what is the step-size (stride) and the size of receptive fields. Moreover, along with denser samplings and multiple scales, a factor that has been shown to have a positive impact on accuracy in object recognition has been the use of larger codebooks [10, 40]. In spite of these results, a careful design of features may still have a clear positive impact on the performance as it was demonstrated in [26, 37]. One of the promising approaches is to apply supervised [26] or unsupervised [24] learning to produce filters that are then used to produce meaningful features for later encoding steps.

The final step is often to compress the high-dimensional representation to a more compact form. This can be done either unsupervisedly (e.g. using WPCA) or supervisedly (e.g. LDA). Using either way has proven to provide extra boost to the face recognition performance in many studies [19, 26, 14, 5]. The first benefit of using PCA or LDA-based methods is the resulting reduced dimension of the final representation (also the utilized class information in LDA). The second benefit comes from the whitening part where the features projected along the components accounting for the highest amount of variability are equalized by scaling which has been shown to improve the matching process. In the supervised setting this scaling can be accomplished via a transformation called Within-Class Covariance Normalization (WCCN) [5].

Our proposed approach differs from the existing works from many aspects. Instead of learning dictionaries directly from pixel patches, like in [30] and [14], we use filter banks to first transform pixel neighborhoods into a more expressive domain. Unlike in [1, 19, 26, 9], we do not fix filter coefficients by hand, but we use unsupervised learning letting the data to fix them. We learn filters like in [24], but instead of natural images, we use face images as a training data. Moreover, instead of binary quantization [24], we use learning based vector quantization. Finally, we argue that, as a characteristic of face representations, overcompleteness has not yet been investigated in depth, but it is only briefly evaluated and discussed. At least, it has not been inspected in the feature level in its most formal meaning. Here, we evaluate overcompleteness in both feature and feature encoding levels. Finally, we show that the best results can be obtained using soft-assignments in the feature encoding step. To make it faster, we propose a novel approximate for soft-assigning that utilizes adjacency matrices.

3. Complete *vs* Overcomplete Features

In [24], Kannala and Rahtu showed that learning features from natural images using Independent Component Analysis (ICA) works very well in the task of building discriminative face representations. They applied ICA to learn a complete basis that was then used as a filter bank to transform input face image patches into features that were further binarized using coordinate-based quantization. To complement their findings, we make a step further and compare whether it is beneficial to learn an overcomplete set of filters instead of a complete one. With an overcomplete basis, we hope to produce more expressive and redundant features for the following encoding step.

Given $x \in \mathbb{R}^d$ of all raw pixel values of an image patch, our aim is to learn a filter bank $W \in \mathbb{R}^{n \times d}$ containing n linear filters $W^{(j)} \in \mathbb{R}^{1 \times d}$ stacked row by row, so that we can further produce local image features $f = Wx$, $f \in \mathbb{R}^n$. Both overcomplete and complete filter banks can be learnt using algorithms based on Independent Component Analysis (ICA).

Standard ICA. As described in [25], given a set of unlabeled data $[x_1, x_2, ..., x_m]$, $x_i \in \mathbb{R}^d$, with zero mean and unit covariance, the standard ICA can be defined as the following constrained optimization problem:

$$\min_W \sum_{i=1}^{m} \sum_{j=1}^{n} g(W^{(j)} x_i), \text{ s.t. } WW^\top = I \tag{1}$$

where g is a nonliear convex function (good for measuring sparsity, for example $g(\cdot) = \log\cosh(\cdot)$), $W \in \mathbb{R}^{n \times d}$ is the weight matrix, n is the number of components, and $W^{(j)}$ is one row (feature) in W. The orthonormality constraint $WW^\top = I$ is used to impose the components to be orthogonal. Using standard ICA, we can easily learn undercomplete and complete set of $W^{(j)}$, meaning a situation where $n \leq d$.

Reconstruction Cost for ICA (RICA). For producing an overcomplete set of $W^{(j)}$ (i.e. when $n > d$) the orthogonality constraint in (1) must be omitted. In [25], motivated by sparse coding and auto-encoders, Le *et al.* proposed a soft reconstruction cost for ICA for producing overcomplete features. RICA is based on the following unconstrained optimization problem:

$$\min_W \sum_{i=1}^{m} \sum_{j=1}^{n} g(W^{(j)} x_i) + \frac{\lambda}{m} \sum_{i=1}^{m} \delta(W, x_i), \tag{2}$$

where the penalty function $\delta(W, x) = \left\| W^\top W x - x \right\|_2^2$ measures the difference between the original sample and its reconstructed version. In general, it has been shown that solving either (1) or (2) using training data sampled from natural images yields features that are localized in space, frequency, and orientation, being similar to Gabor functions [21]. According to [25], RICA coincides with sparse autoencoders and sparse coding if one uses linear activations, soft penalties, and some other relaxed constraints.

In our experiments, we vary the number of learned features and always compare a filter bank produced with ICA with its overcomplete version produced with RICA. Moreover, we are always operating in a whitened and dimensionality reduced space accomplished with PCA. In detail, we first decompose W into two parts $W = UV$, where $V \in \mathbb{R}^{d' \times d}$ $(d' < d)$ contains both whitening and dimensionality reduction, and $U \in \mathbb{R}^{d' \times d'}$ (or $U \in \mathbb{R}^{vd' \times d'}$ in the v-times overcomplete case) which is produced using ICA (or RICA). Thus, we first reduce the dimension of our training sample vectors $z = Vx$, so that $z \in \mathbb{R}^{d'}$ with d' being equal to n. This procedure means we reduce the dimension of our training vectors to the length equal to the desired number of filters in the complete case. Strictly speaking, with ICA our features are always undercomplete with respect to the original sample space, but with respect to preprocessed training vectors, the features are complete and further overcomplete with RICA. Finally, the features are produced by $f = Uz = UVx$. By whitening and dimensionality reduction we reduce the possibility to learn noisy bases. For learning the basis based on ICA and RICA we used the methods described in [20] and [25], respectively.

4. Overcomplete Face Descriptor

Once we have learnt the filters, we propose to use them to produce features that are further used in learning higher dimensional descriptors more expressive in describing faces. To make the final descriptor tuned for different facial regions we learn a set of region specific overcomplete visual dictionaries that are used to map the input feature vectors to an expressive descriptor space. The final face descriptor is then formed by simply concatenating the resulting high-dimensional block-based descriptors. To achieve compactness, we project the resulting face descriptors into a lower dimension using the Whitening Principal Component Analysis (WPCA).

Feature Encoding. As usual, the idea is to partition the local feature space into informative regions resulting in an overcomplete dictionary (or codebook) of so called visual words. We use k-means clustering to construct our codebooks that we finally denote by a matrix $B \in \mathbb{R}^{d \times L}$ containing L visual words $b_j \in \mathbb{R}^d$ as colums. The codebook construction is done similarly for each facial region. For B, d is often less than L, but in the so called codebook overcompleteness, the case is $d \ll L$ [14]. Now, if we let $f \in \mathbb{R}^d$ be a local feature vector from a specific facial region, $u \in \mathbb{R}^L$ the assignment (or coding coefficient) vector of f, and $u^{(j)}$ the assignment with respect to word b_j, we can formulate several types of encoding strategies. In this work we are particularly interested in the *hard-assignment* and *soft-assignment* encodings, as they are among the fastest assignment methods [10]. As for other encoding schemes, it was shown that the localized soft assignment coding, which is utilized here, can perform comparably or even better to sparse and local coding schemes [28].

In hard-assignment (HA) coding, each feature vector f is mapped to a high-dimensional descriptor u with only one nonzero coding coefficient that is set to one in the standard case. The codebook entry or visual word that is selected for the feature is based on the shortest distance, i.e. its nearest neighbour (1-NN) in B. The major setback of HA is that even small variation in some coordinates of f may cause totally different assignments [40, 32].

The shortcomings of HA led to the development of soft-assigment (SA) coding [23, 28]. In this case, instead of indicating 1-NN with 1, each feature vector is encoded with a weighted set of k visual words using the k-NN rule. In detail, for each f the output is a high-dimensional descriptor with coefficients $u^{(j)}$ so that

$$u^{(j)} = \begin{cases} \alpha \exp(-\beta d(f, b_j)) & \text{if } j \in k\text{-NN}(f, B) \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where $d(f, b_j)$ is the distance between vectors f and b_j (usually Euclidean), β controls the softness of the assignment, and $\alpha = 1/\sum_{j=1}^{k} \exp(-\beta d(f, b_j))$ is the L_1 normalization coefficient ensuring the weights sum up to 1. Note that if β is set to 0 and k is set to 1, the strategy in (3) coincidences with HA.

The well-known drawback of SA coding is that its computational overhead compared with HA grows all the time with the size of the codebook, the dimension of the features, and the number of nearest neighbors sought. To make soft assignment coding faster, we argue that it does not make so much difference whether the assignment is done based on the distance between feature f and its nearest words b_j in B or between the visual word $b_i \in$ 1-NN(f, B) and its nearest words b_j in B (with b_i itself counted as the nearest one). If this is the case, SA can be accomplished via first applying HA and then multiplying the resulting sparse descriptor with a matrix containing the assignment weights.

Giving a codebook B, our fast approximate SA (FASA) coding method is based on a smoothing matrix $S_{\beta,k} \in \mathbb{R}^{L \times L}$ with elements

$$s_{\beta,k}^{(i,j)} = \begin{cases} \alpha \exp(-\beta d(b_i, b_j)) & \text{if } b_j \in k\text{-NN}(b_i, B), \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where L is the size of the codebook and others like previously. It is important to note now that $S_{\beta,k}$ does not depend on features f, but only on codebook elements b_j. This enables us to compute the smoothing matrix in advance so that run-time nearest neighbors search can be limited to the nearest one. Moreover, if we set the parameter k so that $k \ll L$ and used the coding given in (4), the result is that $S_{\beta,k}$ becomes sparse which makes it also memory efficient. In the end, FASA can be applied via one matrix-vector product, $S_{\beta,k}u$, where u is a hard-assigned coding coefficient vector.

Pooling. Given the coding coefficient vectors of all features in an image area, we pool the descriptors to obtain the final region-level representation $h \in \mathbb{R}^L$. In detail, we use sum-pooling in which the jth component of h is $h^{(j)} = \sum_{i=1}^{N} u_i^{(j)}$, where N is the total number of encoded features u_i in an image area. While using FASA coding, it is noteworthy that the matrix-vector multiplication can be done after pooling by $S_{\beta,k}h$.

Dimensionality Reduction. Once the block-based descriptors are computed we next concatenate them forming the representation for the whole face area. In our case, this yields a high-dimensional descriptor especially when large codebooks are in use. For having a more compact representation, we transform our high-dimensional descriptors into a lower dimensional space using the Whitening PCA (WPCA) transformation. Besides making our descriptor more compact, we will show that using WPCA also significantly improves the final recognition rates.

Furthermore, in our experiments, WPCA has an essential role in the respect that we use it to reveal the added accuracy of overcompleteness in features and codebooks as hinted in [5].

Face Matching. Three different similarity measures are used in our experiments depending on the representation and whether the scenario is unsupervised or supervised. Regardless of any circumstance, before dimensionality reduction, we always preprocess our raw face descriptors based on the Hellinger kernel [2].

For applying the Hellinger kernel, we first L_1 normalize the raw face descriptor vectors and then replace all coordinate values by their square roots. Finally, comparing face descriptors without dimensionality reduction, the distance can be measured as simply as [2] $d(p, q) = 2 - 2\, p^\top q$, where p and q are two properly preprocessed face descriptors. If we apply dimensionality reduction, we will use the Cosine distance which means that before distance computation we further L_2 normalize the compact descriptors projected by WPCA.

We perform supervised similarity calculations using the Joint Bayesian (JointBayes) [11] algorithm based on modeling the joint distribution of two input face descriptors. In JointBayes there is a prior according to which a face descriptor p is a summation of two variables $p = \mu + \varepsilon$, where μ is for identity and ε for within-person variation. Based on this prior, after solving the covariance matrices S_μ and S_ε, a closed-form solution can be derived for measuring the similarity of two input face descriptors p and q based on a log likelihood ratio test between two joint probabilities so that $r(p, q) = \log P(p, q|H_I) - \log P(p, q|H_E)$, where H_I is the intra and H_E the extra-person hypothesis, respectively. According to [11], the required S_μ and S_ε can be approximated by between-class covariance and within-class covariance matrices used in classic Linear Discriminant Analysis (LDA).

In both unsupervised and supervised scenarios, the final performances are measured by applying the flip-free augmentation [16]. That means, instead of direct similarity calculation between two input representations, we horizontally flip both images before feature extraction and calculate the average similarity between all possible four combinations of the two representations.

5. Experiments

We evaluate our method on the LFW benchmark using the updated protocol (LFW-*updated*) [17] and the novel supplementary protocol denoted as LFW large-scale (LFW-*ls*) [27]. For LFW-*updated*, we evaluate our method solely in the unsupervised mode, whereas for LFW-*ls*, we perform evaluations solely in the supervised mode coupling our unsupervised face descriptor with the state-of-the-art JointBayes [11] metric learning method.

To emphasize the efficiency of our method, we further evaluate our description construction algorithm on constrained (controlled) face identification using FERET [33], and finally, on age estimation in unconstrained scenarios using the recent Adience [15] benchmark. For benchmarking with these datasets, we use the filters and codebooks learnt from LFW images, but the WPCA projections are learnt based on the face descriptors of face images belonging to the dataset under evaluation. In addition, we do not utilize the flip-free augmentation in these experiments. Finally, we use linear SVM for the age estimation experiment.

Setup. LFW consists of 13,233 images of 5,749 people. The LFW-*updated* benchmark is divided into two disjoint subsets called *View1* and *View2*. *View1* is designed for algorithm development, whereas *View2* is a 10-fold cross-validation set with 6,000 face pairs for reporting the final accuracy. As a benchmark, LFW-*updated* is designed solely for face verification.

LFW-*ls* is a supplementary protocol for large-scale unconstrained face recognition under both verification and open-set identification modes. Besides enabling benchmarking in open-set face identification, the protocol provides a very large set of match and non-match pairs enabling statistically stable evaluations at lower false acceptance rates compared with its counterpart, LFW-*updated*. LFW-*ls* contains a development set and an evaluation set of 10 random partitions for reporting the final performance in a cross-validatory fashion.

For all LFW experiments, we use the LFW-a [41] distribution, where all the original LFW images are aligned using a commercial face alignment system. We crop the inner portion of these images of a size 150×81 pixels (see Fig. 3 (a)) and use blocks of a size 30×27 pixels with a horizontal and vertical overlap of 10 and 9, respectively. Each block is then separately coded using one of the feature encoding methods represented here and finally concatenated to form the final representation. We also fix the filter size to 11×11 and, before learning, project the resulting 121-dimensional training vectors first into a lower dimensional space. As the number of filters based on a complete basis (with ICA) is set to 16, the dimension of the reduced training vectors equals to that. With RICA, we learned a basis testing 2 and 3-times overcomplete versions compared to the one learned with ICA, yielding 32 and 48 filters, respectively. Filters and codebooks are learned block-wisely so that there will be a specific set of them for each facial region.

For experimenting with FERET, we are interested in the frontal profile images, which are divided into *fa*, *fb*, *fc*, *dup1*, and *dup2*. Each set has it own characteristics with regards to, for example, illumination and expression. Adience, in turn, is a large dataset composed of over 19,000 face images. The age estimation protocol there is based on eight age groups and the final performance is set to be reported using the mean accuracy (±standard error) over five folds. Further, we use the aligned Adience distribution. For constructing descriptors, we take the best performing combination of filters and codebooks learnt from LFW *View1* faces. We roughly aligned all FERET and Adience faces with the LFW-a coordinate frame relying only on eye locations (see Fig. 3 (a)). For FERET, we further applied a pre-processing step given in [38]. Finally, we found that using LFW faces in learning the WPCA projection does not yield the best possible performance while applied to FERET and Adience faces. Therefore, WPCA is trained using descriptors extracted from the faces of the benchmark under evaluation. To construct a general WPCA is of interest in our future studies.

5.1. Results

We started our experiments with *View1* according to LFW-*updated*. Our preliminary results are mainly for understanding whether, for the codebook constructions, it makes sense to use learnt filters to extract features rather than just to use raw pixels as features. For pixel features, we sampled a set of training image patches straightly from face regions (or blocks) and then applied k-means to learn a codebook for each of these blocks. For filter-based features, we convolved input face blocks with a set of filters (learned from the face image patches sampled from the same region) and then applied k-means to the resulting feature vectors. Regardless of the underlying features, we normalized the training vectors to unit L_2 norm for reducing the illumination changes and suppressing the scale effect [14]. Besides deciding between the underlying features (pixels, ICA, or RICA), it was of major interest to see the effect of using larger codebooks and the potential of learning the filters from face images rather than from some more general images (like in [24]). The results of Table 1 clearly demonstrate that using face images in filter learning (ICA and RICA) is valuable. Also, using larger codebooks, i.e. to have more overcompleteness in codebooks, enhances the performance, as hinted in the literature. However, it seems that the benefits of overcompleteness in features is not yet evident.

codebook	256	512	1024	2048	4096
pixels	.8157	.8237	.8325	.8397	.8463
ICA-16 (using natural images)	.8305	.8384	.8503	.8539	.8611
ICA-16	.8291	.8398	.8513	.8600	.8699
RICA-32	.8266	.8428	.8522	.8618	.8662
RICA-48	.8272	.8416	.8507	.8612	.8676

Table 1. Preliminary results on LFW *View1* reporting the AUC value. All results are based on HA (no WPCA).

To see whether feature overcompleteness truly benefits, we took the WPCA transformation and started gradually projecting the high-dimensional face descriptors into lower dimensional spaces. In Fig. 1 and 2 we can see that (including to the benefits of overcompleteness in codebooks) there are indications that it may also benefit in features. Interestingly, while using larger codebooks soft-assignments become necessary in order to maintain the performance levels. For SA encoded descriptors we varied k and β so that for 256-codebooks it was $k \in [5, 10:10:30]$ with $\beta \in [0.2:0.1:0.6]$ and for 4096-codebooks $k \in [5, 10:10:50]$ with β as above.

For the following experiments, we used mainly 4096-codebooks trained using features resulting from filtering with ICA and 2-times overcomplete RICA, and for both learning only from face image patches. For feature encoding, we used soft assignment and compared it to its fast approximate version for which the adjacency matrices were learned offline using the codebooks applying the best k and β (based on the SA evaluations on *View1*). For learning block-specific filter banks and codebooks, we sampled altogether 100,000 face image patches from the corresponding region of the faces in the training set. The length of the final descriptor finally yielded over 144K and 280K for 4096 and 10000-codebooks, respectively.

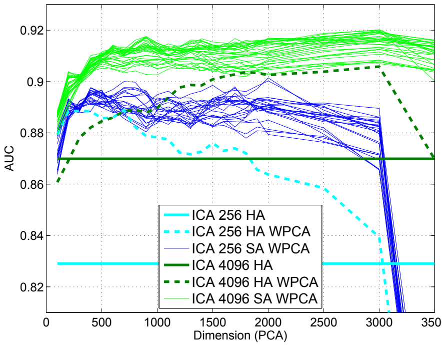

Figure 1. A plot of AUC values on *View1* with face descriptors using ICA features together with 256 and 4096 size of codebooks varying the feature encoding strategy and the parameter values.

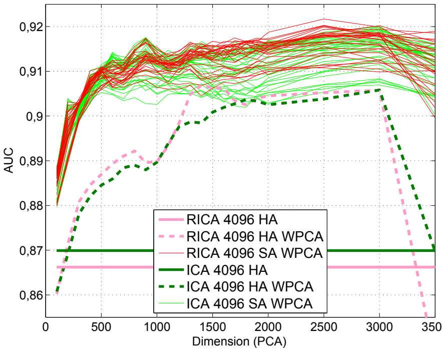

Figure 2. A plot of AUC values on *View1* with face descriptors using ICA and RICA-32 features together with 4096-codebooks varying the feature encoding strategy and the parameter values.

LFW-*updated*. We herein use the *View2* portion of the LFW-*updated* benchmark. Because of the particular setting of *View2*, everything is learned altogether 10 times. On every round, all filters, codebooks, and the WPCA projection are learned using only images belonging to other nine folds separate from the fold currently under evaluation. Before learning the block-specific filters and codebooks, we proceeded by evenly sampling patches from all face images in the current training set so that we finally had the desired amount of training data for each block. Besides 4096-codebooks, we also tested block-specific codebooks of 10,000 elements. After feature extraction, encodings, and concatenations the length of the descriptor is over 114K and 280K for 4096 and 10000-codebooks, respectively. The compact descriptor is then achieved by learning a WPCA projection from these descriptors. Following the protocol under unsupervised evaluation category, we report the performance of our methods by measuring the area under the average ROC curve (AUC) over 10 trials.

The obtained results are summarized in Tables 2 and 3. These results clearly indicate that our approach outperforms not only almost all state-of-the-art methods shown in Table 3 but also other recent methods including HighDimLBP, LE and LBP. All these methods are fairly evaluated using the same protocol. As MRF-Fusion-CSKDA [3] uses different kind of pose correction and a fusion strategy the results are not directly comparable to ours.

descriptor	raw	wpca-400	raw *ff*	wpca-400 *ff*
ICA HA	.8650	.8912	.8858	.9097
ICA SA	.8469	.9173	.8630	.9359
ICA FASA	.8251	.9088	.8412	.9251
RICA HA	.8693	.8912	.8906	.9095
RICA SA	.8372	.9189	.8513	.9368
RICA FASA	.8047	.9133	.8135	.9299
HighDimLBP	.7947	.9292	-	-
LE	.8064	.8905	-	-
LBP	.7782	.8819	-	-

Table 2. Results of LFW-*updated View2* under the unsupervised evaluation category. With SA, for both ICA and RICA, we set $\beta = 0.2$ and k as 15 and 40, respectively. For FASA, we set k and β as in the corresponding SA. *ff* stands for the flip-free augmentation.

descriptor	wpca-400	wpca-900	wpca-3000
RICA-32 4,096 SA	.9189 / .9368	.9259 / .9425	.9263 / .9462
RICA-32 10,000 SA	.9189 / .9372	.9264 / .9436	.9325 / **.9480**
HighDimLBP [12]	.9292	.9338	.9350
MRF-LBP (WPCA) [4]	.8994 (WPCA dim not reported in [4])		
PAF (WPCA) [43]	.9405 (WPCA dim not reported in [43])		
MRF-Fusion-CSKDA [3]	**.9894** (WPCA dim not reported in [3])		

Table 3. Final evaluation on LFW *View2* under the unsupervised evaluation category according to [17]. For both 4,096 and 10,000-codebooks, the parameters are set as $\beta = 0.2$ and k as 40 and 60, respectively. The first value is for direct distance calculations and the second (after slash and if given) for flip-free augmented.

LFW-*ls*. Like previously, we learn all filters, codebooks, and the WPCA transformation so that for each trial we use only those faces that belong to the current training set. Now, as we are using the JointBayes method, we also need to compute within-class and between-class scatter matrices on every trial. As there are 10 trials, we repeat these steps as many times. All parameters are fixed as in the previous experiments, but here we are using only 4096-codebooks. In face verification, we report the performance of our methods in terms of verification rates at FAR = 0.1% and 1%. For the open-set identification, we report the rank 1 detection and identification rate (DIR) at FAR = 1% and 10%. In both cases, we use the $\mu - \sigma$ measure [27]. The baseline results of HighDimLBP, LE, and LBP are from [27].

The obtained results in Table 5 and 6 indicate that using RICA features with SA gives better results than the current reported best one, namely HighDimLBP, in large-scale face verification, but slightly falls behind it in the large-scale open-set identification scenario. However, all the proposed methods perform better compared with LE and LBP.

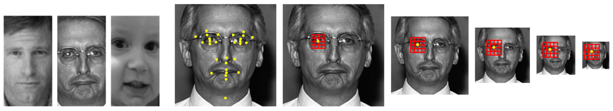

Figure 3. (a) Examples of the cropped LFW, FERET, and Adience faces used in our experiments. (b) Illustration of the fiducial points and image scales used in our HighDimLBP implementation (see details of the used landmark detection method and used image scales in [12]).

Cross-database evaluation. For both FERET and Adience evaluations we took our best performing model which was a combination of RICA-32 and 10K-codebooks coupled with SA encoding setting $k = 60$ and $\beta = 0.2$. For FERET, we learnt the WPCA projection using the gallery set (also known as *fa*) keeping the maximum amount of dimensions possible, i.e. $\#gallerySamples - 1$. For Adience, we learnt block-specific WPCA projections instead of a global one, which has been shown to perform poorly [15] in age estimation. Also, rather than fixing the number of dimensions, we evaluated the performance by keeping the 30, 40, and 50% of the variance for each block-based projection on every iteration (using 40% the length of the final compact descriptor was around 500 on average over all folds). For every iteration, we picked 2,000 face images to train a WPCA projection. Based on the results given in Table 4 and 5, we outperformed the state-of-the-art in FERET among single scale descriptors (no-fusion) and in Adience among the methods given in [15]. The HighDimLBP result here is based on our own implementation (see Fig. 3 (b)) based on the same landmark localization method used in [12] combined with uniform LBP codes setting $P = 8$ and $r = 3$.

descriptor	HighDimLBP	DFD [26]	LGXP [42]	G-LQP [19]	Ours
fb	99.4	99.4	99.0	**99.9**	99.6
fc	99.5	**100**	**100**	**100**	**100**
dup1	88.6	91.8	92.0	93.2	**94.9**
dup2	83.3	92.3	91.0	91.1	**93.6**

Table 6. Comparison to the state-of-the-art methods on FERET. All but LGXP uses WPCA for compression. LGXP uses supervised Fisher Linear Discriminant (FLD) approach. For both HighDimLBP and our proposed method, WPCA dim is set as 1195.

method	accuracies
LBP+FPLBP (PCA / raw / Drop-out) [15]	38.1±1.4 / 44.5±2.3 / 45.1±2.6
Ours (0.3 / 0.4 / 0.5)	47.1±2.9 / **48.1**±2.7 / 47.2±2.5

Table 7. Comparison to the current best methods on unconstrained age estimation using the Adience benchmark.

6. Discussion and Conclusion

The inclusion of overcompleteness has recently been argued crucial for building discriminative unsupervised face representations. In this paper, we investigated this statement from many aspects. Finally, we showed that a state-of-the-art description can be constucted by coupling overcomplete codebook learning with a well-defined set of overcomplete features. We compared our method with ones utilizing multi-scale descriptors and facial landmark-based com-

method	HighDimLBP	LE	LBP	ICA HA	ICA SA	ICA FASA	RICA HA	RICA SA	RICA FASA
FAR = 0.1%	.4166	.2331	.1418	.2571 / .2895	.3843 / .4133	.3393 / .3702	.2555 / .2906	.3919 / **.4234**	.3529 / .3839
FAR = 1%	.6585	.4660	.3139	.5313 / .5521	.6469 / .6609	.6086 / .6272	.5270 / .5523	.6444 / **.6641**	.6130 / .6332

Table 4. Face verification results following the LFW-*ls* protocol. The reported numbers are the mean verification rates (%) subtracted by the corresponding standard deviations over 10 trials. The first value is for direct distance calculations and the second (after slash) for flip-free augmented.

method	HighDimLBP	LE	LBP	ICA HA	ICA SA	ICA FASA	RICA HA	RICA SA	RICA FASA
FAR = 1%	**.1807**	.1126	.0882	.0546 / .0510	.1308 / .1515	.1108 / .1073	.0556 / .0541	.1514 / .1622	.1193 / .1210
FAR = 10%	.3263	.2073	.1661	.1845 / .2018	.2865 / .3023	.2486 / .2690	.1843 / .2034	.2899 / **.3337**	.2574 / .2739

Table 5. Open-set identification results at rank-1 following the LFW-*ls* protocol. The reported numbers are the mean detection and identification rates (%) subtracted by the corresponding standard deviations over 10 trials.

putations being able to perform on par using only single-scale features and without any use of facial landmarks. While WPCA has been earlier shown to be powerful, we demonstrated that using codebook-based methods with soft-assignments makes it even more powerful. We also proposed a novel method for approximative soft-assignment by using sparse adjacency matrices, but further analysis and experiments are needed to better understand its potential.

Based on our experiments, one may still afford to improve unsupervised face description process especially at the feature level. We base our claim on the observation which shows that without an extensive use of facial landmarks we were able to construct a descriptor that performed on par with methods like HighDimLBP and PAF that are the top-performing ones among the unsupervised landmark-oriented facial descriptors. Based on our FERET experiment, one may raise further questions whether landmark-oriented HighDimLBP method is the optimal one for processing with frontal faces. Our approach, in turn, outperforms the current state-of-the-art methods reported on FERET benchmark with frontal faces. We also applied our approach to age estimation showing promising results. We showed that applying dimensionality reduction is beneficial, but it seems it must be done block-wisely. In our future work, we are planning to make an in-depth comparison between global and block-wise linear projections used as dimensionality reduction for producing unsupervised face descriptors for age estimation.

Acknowledgement. The financial support of the Academy of Finland is gratefully acknowledged. The work was also partially supported by the FiDiPro program of Tekes.

References

[1] T. Ahonen and M. Pietikäinen. Image description using joint distribution of filter bank responses. *Pattern Recognition Letters 30(4):368-376*, 2009. 1, 2

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 4

[3] S. Arashloo and J. Kittler. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *Information Forensics and Security, IEEE Transactions on*, 9(12):2100–2109, Dec 2014. 7

[4] S. R. Arashloo and J. Kittler. Efficient processing of MRFs for unconstrained-pose face recognition. In *BTAS*, 2013. 7

[5] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz. Fast high dimensional vector multiplication for face recognition. In *ICCV*, pages 1960–1967, 2013. 1, 2, 4

[6] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 1997. 1

[7] Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. In *ICCV*, 2013. 2

[8] X. Cao, D. Wipf, F. Wen, and G. Duan. A practical transfer learning algorithm for face verification. In *ICCV*, pages 3208–3215, 2013. 1, 2

[9] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714, 2010. 2

[10] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011. 2, 4

[11] D. Chen, X. Cao, L. Wang, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, pages 566–579, 2012. 1, 5

[12] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, pages 3025–3032, 2013. 1, 2, 7

[13] A. Coates, H. Lee, and A. Ng. An analysis of single-layer networks un unsupervised feature learning. *Ann Arbor*, 1001:48109, 2010. 2

[14] 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 *CVPR*, 2013. 1, 2, 4, 6

[15] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Information Forensics and Security, IEEE Transactions on*, 9(12):2170–2179, Dec 2014. 5, 7

[16] C. Huang, S. Zhu, and K. Yu. Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval. TR115, 2007. 5

[17] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures, 2014. Technical report UM-CS-2014-003. 5, 7

[18] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller. Learning to align from scratch. In *NIPS*, pages 773–781, 2012. 2

[19] S. u. Hussain, T. Napoleon, and F. Jurie. Face recognition using local quantized patterns. In *BMVC*, 2012. 1, 2, 7

[20] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE TNN*, 10(3):626–634, 1999. 3

[21] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural image statistics*. Springer, 2009. 3

[22] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer-Verlag, 2007. 1

[23] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM*, CIVR '07, pages 494–501, 2007. 4

[24] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *ICPR*, pages 1364–1366, 2012. 2, 3, 6

[25] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In J. Shawe-taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *NIPS*, pages 1017–1025. 2011. 3

[26] Z. Lei, M. Pietikäinen, and S. Z. Li. Learning discriminant face descriptor. *IEEE TPAMI*, 36(2):289–302, 2014. 1, 2, 7

[27] S. Liao, Z. Lei, D. Yi, and S. Z. Lii. A benchmark study of large-scale unconstrained face recognition. In *IJCB*, 2014. 5, 7

[28] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, pages 2486–2493, 2011. 4

[29] G. D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157 vol.2, 1999. 1

[30] X. Meng, S. Shan, X. Chen, and W. Gao. Local visual primitives (lvp) for face modelling and recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 536–539, 2006. 2

[31] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *CVPR*, 2014. 1

[32] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8, 2008. 4

[33] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *IVC*, 16(5):295–306, 1998. 5

[34] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013. 1

[35] B. M. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *ECCV*, pages 78–93, 2014. 2

[36] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *ICCV*, pages 1489–1496, 2013. 1

[37] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 2

[38] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE TIP*, 19(6):1635–1650, 2010. 5

[39] M. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591, 1991. 1

[40] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE TPAMI*, 32(7):1271–1283, 2010. 2, 4

[41] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE TPAMI*, 33(10):1978–1990, 2011. 5

[42] S. Xie, S. Shan, X. Chen, and J. Chen. Fusing local patterns of Gabor magnitude and phase for face recognition. *IEEE TIP*, 19(5):1349–1361, 2010. 7

[43] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *CVPR*, 2013. 7