Improving Random Forests by correlation-enhancing projections and sample-based sparse discriminant selection

Marcus Wallenberg
Dept. of Electrical Engineering
Linköping University
Linköping, Sweden
marcus.wallenberg@liu.se

Per-Erik Forssén
Dept. of Electrical Engineering
Linköping University
Linköping, Sweden
per-erik.forssen@liu.se

***Abstract*—Random Forests (RF) is a learning technique with very low run-time complexity. It has found a niche application in situations where input data is low-dimensional and computational performance is paramount. We wish to make RFs more useful for high dimensional problems, and to this end, we propose two extensions to RFs: Firstly, a feature selection mechanism called *correlation-enhancing projections*, and secondly *sparse discriminant selection* schemes for better accuracy and faster training. We evaluate the proposed extensions by performing age and gender estimation on the MORPH-II dataset, and demonstrate near-equal or improved estimation performance when using these extensions despite a seventy-fold reduction in the number of data dimensions.**

***Keywords*-image classification;**

I. Introduction

Random Forests (RF) [2], [3] is a learning technique with very low run-time complexity. It has found a niche application in situations where input data is low-dimensional (tens or hundreds of dimensions) and computational performance is paramount. We wish to make RFs more useful for high-dimensional problems (tens or hundreds of thousands of dimensions), where deep learning [9], [18] is currently the method of choice. To this end, we propose two extensions to RFs: A feature selection mechanism called *correlation-enhancing projections* (CEP), and two discriminant selection schemes, *normalised sparse discriminant selection* (NSDS) and *sample-based sparse discriminant selection* (SSDS), for better accuracy and faster training.

Correlation enhancing projections work by first computing an unsupervised appearance clustering, and then selecting locally discriminative features using canonical correlation analysis (CCA) [1]. The result is a feature space where similarity better corresponds to response space similarity. The use of CCA allows us to do this simultaneously for disparate responses such as subject gender (a classification response) and age (a regression response), and thus the resultant projections are inherently well suited to feature sharing.

The proposed NSDS scheme provides automatic centering of projection data during training in high-dimensional spaces, reducing the need for threshold optimisation. The SSDS scheme is a simple, but effective means to speed up the search for discriminants which is the computational bottleneck for RF learning in high dimensional spaces. In the experiments, we evaluate the proposed extensions by performing age and gender estimation on the MORPH-II dataset [8].

A. Related work

In attribute estimation from facial images, state-of-the art has for quite some time been to use hand crafted features, e.g. [7], [17] coupled with support-vector machines (SVM) for classification or regression [13]. Recently however, features that are more data-driven have surpassed these in benchmarks [18], and this motivates us to also pursue this approach. One possible way to do this is to apply generic learned features from e.g. DeCAF [4], but as our goal is a highly efficient system, we want to simplify the feature computation to a minimum. We thus propose a two step feature selection, where the first one partitions the feature space using k-means clustering [11], [13], and the second step finds *correlation-enhancing projections* (CEP), by in each cluster performing canonical correlation analysis (CCA) [1] between the feature space and the response dimensions.

Our NSDS and SSDS schemes are related to the Forest-RC algorithm [2], and to the RANSAC algorithm used in geometric computer vision [5]. Just like Forest-RC, NSDS and SSDS find discriminants that are linear combinations of a subset of the feature dimensions. In Forest-RC, the feature dimensions are chosen randomly, and then random discriminants between these are checked to find a good split. In NSDS, a normalisation of the coefficients is performed, which serves to center the resulting projections. In SSDS splits are instead selected by first drawing a sample subset, and based on this, a subset of feature dimensions is then chosen. For classification, the minimal sample subset consists of just two samples, and this is the size we use. The two samples are drawn based on response; for e.g. a binary classifier, we draw one sample randomly from each class. Based on the sample subset, a discriminant projection

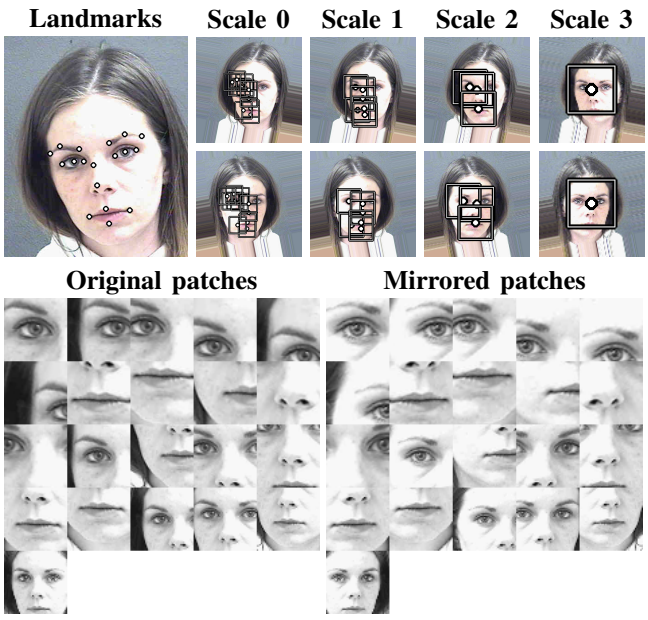

Figure 1: Facial landmarks and image patches used. Example from MORPH-II dataset (female, age 25). Top left: Original image with facial landmark points. Top right: extracted image patches and corresponding scale levels, original orientation (first row) and mirrored (second row). Bottom: resulting greyscale patches, original orientation (left) and mirrored (right). Artifacts at image borders are due to border replication in the rectification step.

can then be defined, as the hyperplane with best sample separation. We then proceed to sparsify this projection, by only allowing coefficients in a subset of the feature dimensions, and in this way obtain an efficient discriminant.

II. Application example: age and gender estimation

To test the proposed extensions, we perform age and gender estimation from facial images on the MORPH-II dataset [8], using the cross-validation folds from [14]. This is a fine-grained problem, where both high invariance and high discriminative power are required to pick up on the subtle differences needed for the estimation task. Changes such as differences in pose and illumination, as well as within-class differences such as differences in skin tone and facial geometry are difficult to handle for an estimator operating only on images. Such changes introduce *domain shifts* [16] which disrupt the neighbourhood structure of the feature space (two images of the same face in different poses or illuminations are not likely to have similar pixel values). The most common way to address this is to model these differences using a parametric model. This is a difficult problem, and may require highly specialised knowledge of the differences to be modelled [10]. Methods that do not rely on parametric models, such as [18], [7] typically use a very computationally expensive estimator (such as a CNN) [18] or, computationally expensive image features like BIF [7], [17]. We instead propose a combination of appearance clustering locally adapted correlation-enhancing projection operations, which do not require any problem-specific knowledge and are significantly cheaper to compute than the alternatives above. This means that although age and gender estimation is used as an application example, the technique is applicable to many other classification and regression tasks.

As input data, we use image patches as in [18] (see figure 1). To extract the patches, we first find facial landmarks, using the `pico` detector [12], and then rectify the facial region of interest to a canonical scale and orientation. Note that the detector used lacks points corresponding to two of the regions used in [18]. However, the remaining 21 regions are extracted and due to the amount of overlap we do not believe the final two regions play a decisive role in the estimation process. We also use patches of size 47×47 pixels (48×48 pixels were used in [18]). Half of these regions are extracted from the scale pyramid of the original rectified image, and half are extracted from a mirrored version. Thus, each image yields two sets of 21 patches. Each such set can be used as a training sample, resulting in $N = 46\,389$ feature dimensions in the input space of the estimator.

III. Appearance clustering and locally adapted correlation-enhancing projections

To find local clusters of similar appearance, we perform k-means clustering [11], [13] on the coarse-scale full-face patches of each training image. The result of such a clustering can be seen in figure 2. A small number of clusters ($k = 16$) is used, in order to ensure that each cluster is homogeneous enough to be represented by a single local feature, but large enough so that the feature selection process can be performed robustly. This number is highly dependent on the distribution of training data, and has not been optimised for this specific dataset. Once a set of k cluster prototypes has been found, means are subtracted and CCA is performed separately for each landmark patch in each cluster. As there are two response dimensions (age and gender), two CCA projections are obtained for each patch (see figure 3). When combined, these form the *correlation-enhancing projection* (CEP). With 21 landmark points, this results in a total of 42 projection values for each image in a cluster. In order to visualise the relative importance of different input feature dimensions (in this case corresponding to facial regions), the average weight magnitude is also shown in figure 3.

Since a novel image may belong to any cluster, and since we have observed that the CCA components of neighbouring clusters in many cases are similar, we represent all images using the CCA features obtained in all clusters. With $k = 16$,

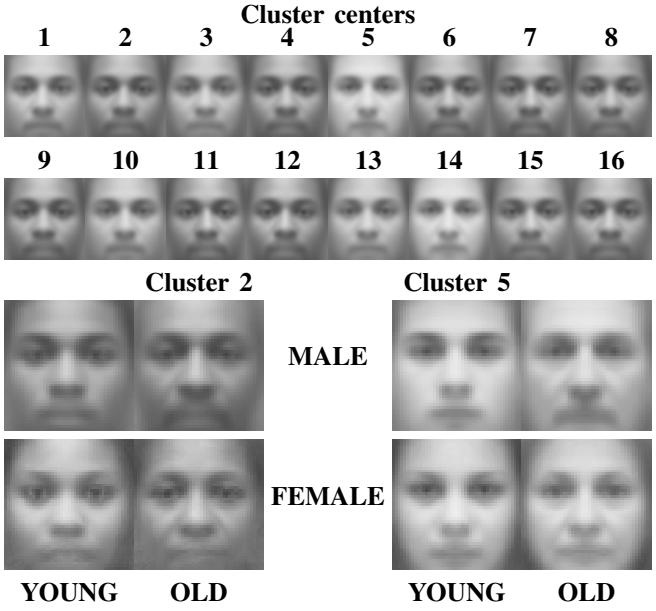

Figure 2: Clustering results. Top: full-face regions for each cluster center. Bottom: average full-face images of 10% oldest and youngest males and females in clusters two and five. Images were drawn from folds 1–4 of the MORPH-II dataset (omitting fold 0).

this results in a total of $N = 672$ feature dimensions per image (a reduction by a factor of approximately 70).

IV. Normalised sparse projection Random Forest with sample-based discriminant selection

Our baseline estimator is a special case of the Forest-RC algorithm [2]. It relies on linear combinations of S (where $1 < S \leq N$) feature dimensions as discriminants. However, the manner in which these discriminants are generated is not uniform (as in the original formulation), but engineered to simplify training. Typically $S \ll N$, representing a very sparse projection operation.

To optimise the internal nodes, we use a random test generation scheme. For each node, a maximum number M of tests are generated and evaluated using the information gain (mutual information), as is done in the standard Random Forest formulation [2], [3]. For the regression case, we use a separable Gaussian approximation of the left and right splits in order to express the resulting entropy in closed form. If a discriminant is found which separates the classes (in the classification case) or produces splits with a variance below a set threshold (in the regression case), node optimisation is terminated. If no such test is found, the best candidate among those evaluated is kept instead.

When selecting discriminant candidates at an internal node, feature space dimensionality affects the choice of

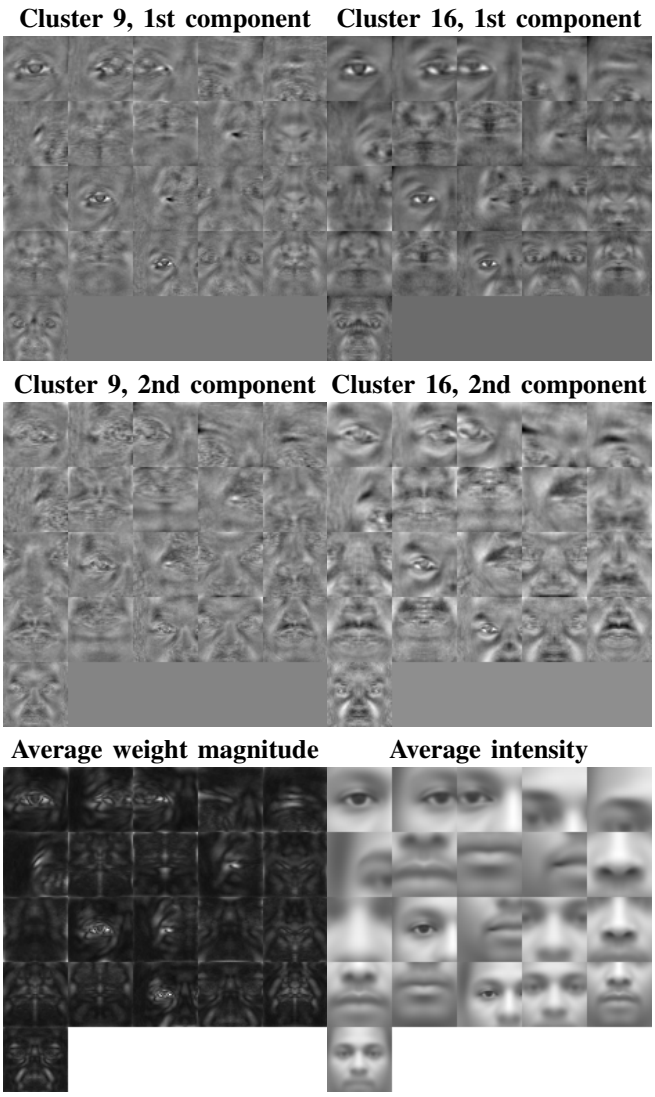

Figure 3: First (top row) and second (middle row) CCA basis vectors for clusters 9 and 16. Images were drawn from folds 1–4 of the MORPH-II dataset. (omitting fold 0). Bottom row: Average weight magnitude (left) and average intensity for reference (right).

method. The likelihood of a randomly selected sparse linear discriminant producing a useful split decreases with dimensionality. However, the computational cost of finding an optimal sparse linear discriminant for a minimal set of points also decreases with dimensionality. Therefore, we propose the following scheme: For high-dimensional data (such as image patches), sparse discriminants (S non-zero dimensions) are generated at random. The projection coefficients w_s are normalised such that

$$\sum_{s=1}^{S} w_s = 0, \quad \text{and} \quad \sum_{s=1}^{S} |w_s| = 1. \tag{1}$$

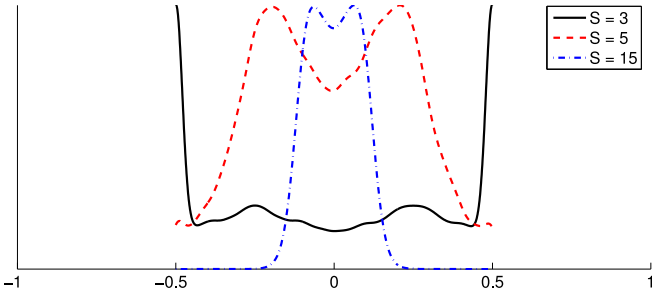

Figure 4: Normalised kernel density estimate (1000 samples, $\sigma = 0.02$) of relative coefficient distribution for different values of S.

This has the effect of centering the mean value of the projections at 0, which is then used as the threshold value except at the final split level of each tree, where the small number of remaining samples make threshold optimisation tractable. The absolute sum normalisation also preserves the range of the selected data dimensions. The resulting coefficient distribution for different values of S can be seen in figure 4. For specific choices of S we can also obtain other previous discriminants, for instance those used in [15], which are obtained for $S = 2$. We denote this selection method *normalised sparse discriminant selection* (NSDS).

For lower-dimensional data (such as the CEP features), discriminants are found in a manner similar to RANSAC [5]. For a minimal subset of two (different) sample points, the optimal discriminant is the hyperplane mid-way between the points with a normal parallel to the line connecting them. This discriminant is greedily sparsified by selecting the S largest-magnitude components of the normal vector, and is used as a candidate test for the node. In this setting, threshold optimisation is also only performed at the final split level. We denote this selection method *sample-based sparse discriminant selection* (SSDS).

V. Results and Discussion

In this section, we present results of the proposed extensions. The first section shows an evaluation of the CEP feature space, and the second shows results for Random Forest evaluation.

A. Feature selection

In order to evaluate the effect of the feature selection process described in section III, we perform a $5 - NN$ evaluation using Euclidean distance in the respective feature spaces (before and after applying the correlation-enhancing projection). Results of this evaluation are shown in table I.

As the table indicates, the local neighbourhood of a data point is more homogeneous after the CEP has been applied. We take this to mean that the application of the CEP simplifies the estimation problem, which should improve performance and simplify estimator training.

Method	Data dimensions	Gender error	Age MAE
Input space	46 389	$13.61 \pm 2.17\%$	8.05 ± 0.30 years
CEP	672	$7.24 \pm 1.14\%$	6.02 ± 0.20 years

Table I: Feature space evaluation using 5-NN. Standard deviations indicate results from five-fold cross-validation, using folds supplied in [6].

B. Random Forests

To evaluate the effect of CEP, NSDS and SSDS on the Random Forest, three variants were compared:

- **Naive Random Forest**: An example of the most basic Random Forest estimator, selecting a random feature dimension ($S = 1$) at each node and then brute-force optimising a threshold along that dimension. Input data are the original image patches, distributed such that each tree is trained on a subset of dimensions corresponding to a single landmark point.
- **NSDS**: Our variant of Forest-RC with normalised sparse projections ($S = 5$). Input data are the original image patches, distributed such that each tree is trained on a subset of dimensions corresponding to a single landmark point. A maximum number of 4096 discriminant candidates were evaluated at each node.
- **CEP-NSDS(5)**: The NSDS ($S = 5$), trained on the lower-dimensional CEP features. Input data are the CEP features, distributed such that each tree is trained on a subset of dimensions corresponding to a single data cluster. A maximum number of 2048 discriminant candidates were evaluated at each node.
- **CEP-SSDS(5)**: The SSDS ($S = 5$) with sample-based discriminant selection, trained on the lower-dimensional CEP features. Input data are the CEP features, distributed such that each tree is trained on a subset of dimensions corresponding to a single data cluster. A maximum number of 256 discriminant candidates were evaluated at each node.

In all cases, trees were grown to a maximum depth of 10, and the ensemble size was set as three trees per (non-mirrored) landmark (63 trees) in the first two cases, and three trees per data cluster (48 trees) in the last two cases. Sample bagging was performed by drawing (with replacement) 50 000 samples of each gender (for gender classification) and 2000 samples of each available age (for age estimation). Results of this evaluation are shown in table II.

The naive Random Forest, considering a single random data dimension at a time and performing brute-force threshold optimisation obtained a classification error rate of $32.05 \pm 1.34\%$ on the image patch data. Because the brute-force threshold optimisation is more computationally expensive for regression than for classification, age estimation was not performed in this case. At the set maximum depth, this variant is unsuitable for high-dimensional data. In order to

Method/trees/depth	Gender error	Age MAE	Iter.	Time
NSDS(5)/63/10	$4.83 \pm 0.77\%$	7.97 ± 0.19 y	4096	≈ 140 h/fold
CEP-NSDS(5)48/10	$5.08 \pm 0.91\%$	5.78 ± 0.04 y	2048	≈ 40 h/fold
CEP-SSDS(5)/48/10	$4.48 \pm 0.41\%$	5.29 ± 0.04 y	512	≈ 40 h/fold

Table II: Age and gender estimation performance for the different methods. Standard deviations indicate results from five-fold cross-validation, using folds supplied in [6]. For NSDS(5), the age estimation result is based only on the first two folds due to the prohibitive optimisation time for full cross-validation on the test machine.

handle these data, tree depth and number would have to be vastly increased. NSDS however, can produce good results even at this depth while using very sparse projections (5 non-zero elements out of $46\,389$) and also while performing no threshold optimisation, except at the final split level. For CEP-NSDS, age estimation results are significantly better than for NSDS. Gender classification performance, though slightly decreased, is still within one standard deviation of NSDS despite the massive reduction in the number of feature dimensions. For CEP-SSDS, both age and gender estimation are improved, despite using only 512 iterations at each node.

Timing results are from a machine with a quad-core Intel Xeon processor, running at 2.30 GHz. Feature selection is performed using all cores, while tree training is run using a single core for each evaluation fold. Although the feature selection takes time to perform (around 18 h on the test system), training time for the methods using CEP is still significantly shorter than for the image-space methods. Although the CEP methods use fewer discriminant candidates per node, the reduction in training time is not proportional to this reduction. This indicates that the CEP features do provide faster convergence. When using sample-based discriminant selection, the reduction in the number of iterations does not produce the corresponding reduction in training time, due to the higher computational cost of SSDS. However, at a corresponding number of iterations NSDS seldom produces usable results.

These results, though not state of the art (to the authors' knowledge, the best published results on the MORPH-II dataset are 1.5% error rate for gender classification [7] and 3.63 years MAE for age estimation [18]), indicate that the proposed extensions have the desired effect when dealing with high-dimensional problems. Since the parameters are not optimised for performance on the dataset used, best-case performance has yet to be determined.

VI. Conclusions and future work

We have proposed and evaluated extensions designed to make Random Forests better equipped to deal with high-dimensional problems including both feature selection and estimator training. In our experiments, the feature selection method based on CEP improves neighbourhood homogeneity of samples, and produces equal or superior estimation performance. In combination with the proposed discriminant generation schemes the rate of convergence is also improved, reducing the number of iterations needed and further improving results for age and gender estimation.

Future work includes optimisation schemes to automatically determine appropriate parameters of the clustering, CCA and Random Forest steps, as well as further investigation of the usefulness of the k-NN neighbourhood score as a measure of feature space quality. The effect of applying label- and parameter-based decision margin cost functions during RF optimisation will also be studied, as this has the potential of increasing the overall robustness and generalisation capabilities of the estimators. Furthermore, frequency weighting and a spatial weight function in feature selection could produce more complementary features by forcing them to be more localised around the landmark points.

Acknowledgements

The research presented in this paper was funded by Linköping University, and also in part by the Vinnova project "FaceTrack", grant no. 2013-00439.

References

[1] Magnus Borga. Canonical correlation: a tutorial. Technical report, Linköping University, 2001. http://www.imt.liu.se/people/magnus/cca/tutorial/tutorial.pdf.

[2] Leo Breiman. Random forests. *Mach. Learn.*, 45(1), 2001.

[3] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, 2012.

[4] Jeff Donahue, Yangqing Jia, and et al. DeCAF: A deep convolutional activation feature for generic visual recognition. ArXiv'13.

[5] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting, with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[6] Tobias Gehrig, Matthias Steiner, and Hazm Kemal Ekenel. Draft: Evaluation guidelines for gender classification and age estimation. Technical report, Karlsruhe Institute of Technology, 2011.

[7] Guodong Guo and Guownag Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *IEEE Face and Gesture*, April 2013.

[8] K. Ricanek Jr. and T. Tesafaye. MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2006.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS'12)*, 2012.

[10] Khoa Luu, Keshav Seshadri, Marios Savvides, Tien D. Bui, and Ching Y. Suen. Contourlet appearance model for facial age estimation. In *IJCB*, 2011.

[11] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[12] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. Fast localization of facial landmark points. Technical Report arXiv:1403.6888, arXiv, 2014.

[13] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[14] Karlsruhe Institute of Technology. FIPA - BeFIT - Proposed Benchmarks - Age Estimation. `http://face.cs.kit.edu/433.php`.

[15] Mustafa Özuysal, Pascal Fua, and Vincent Lepetit. Fast keypoint recognition in ten lines of code. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 0, 2007.

[16] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV'10)*, 2010.

[17] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[18] Dong Yi, Zhen Lei, and Stan Z. Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision ACCV'14*, 2014.

