USING SUPERVISED DEEP LEARNING FOR HUMAN AGE ESTIMATION PROBLEM

K. A. Drobnyh[a], A. N. Polovinkin[a]

[a] Lobachevsky State University of Nizhny Novgorod - (klim.drobnyh, alexey.polovinkin)@gmail.com

Commission II, WG II/10

KEY WORDS: Machine Learning, Age Estimation, Supervised Deep Learning, Active Appearance Model, Bio-Inspired Feature, Support Vector Machine

ABSTRACT:

Automatic facial age estimation is a challenging task upcoming in recent years. In this paper, we propose using the supervised deep learning features to improve an accuracy of the existing age estimation algorithms. There are many approaches solving the problem, an active appearance model and the bio-inspired features are two of them which showed the best accuracy. For experiments we chose popular publicly available FG-NET database, which contains 1002 images with a broad variety of light, pose, and expression. LOPO (leave-one-person-out) method was used to estimate the accuracy. Experiments demonstrated that adding supervised deep learning features has improved accuracy for some basic models. For example, adding the features to an active appearance model gave the 4% gain (the error decreased from 4.59 to 4.41).

1. INTRODUCTION

Automatic facial age estimation is a challenging task upcoming in recent years due to numerous applications in security systems, human-computer interactions, age-invariant person identification.

Using facial images might be the simplest way to estimate the age since human faces are considerably affected by the aging process. Therefore there are many articles that use facial images to solve this problem since 1999 (Kwon and Lobo, 1999).

Though the problem seems to be quite simple: the goal is to predict human age using only facial image, it remains a complex problem. The reason is that face skin is heavily determined by external factors, such as environment and lifestyle. Moreover, age perception can be changed by clothes / glasses / mustache.

This is a classic example of pattern recognition problem that can be stated as to classify a new unobserved sample into one of the predefined classes. In this case, classes are the ages and samples are the facial images.

One of the state-of-the-art approaches performing a pattern recognition task is deep learning. Deep learning is a set of algorithms that uses machine learning to find the best feature representation of the input data. Drobnyh (2016) used the unsupervised deep learning method based on K-Means clustering algorithm to solve the same problem.

A new supervised deep learning algorithm that uses random forest was proposed by Martyanov et al. (2012). In this paper, we propose to use the supervised deep learning approach mentioned above to improve an accuracy of the existing age estimation algorithms.

2. BASIC APPROACHES

The majority of approaches that give the best accuracy, according to Panis et al., 2016, are based on the active appearance model and the bio-inspired features.

2.1 Active appearance model

An active appearance model (Cootes et al., 1998) is a statistical model that describes both the shape and gray-level appearance of the object of interest. This model can be generalized to almost any valid case. A brief description of the algorithm is presented below.

First of all, we need to build an active shape model. Input data is a training set of images where each object is labeled with landmarked points.

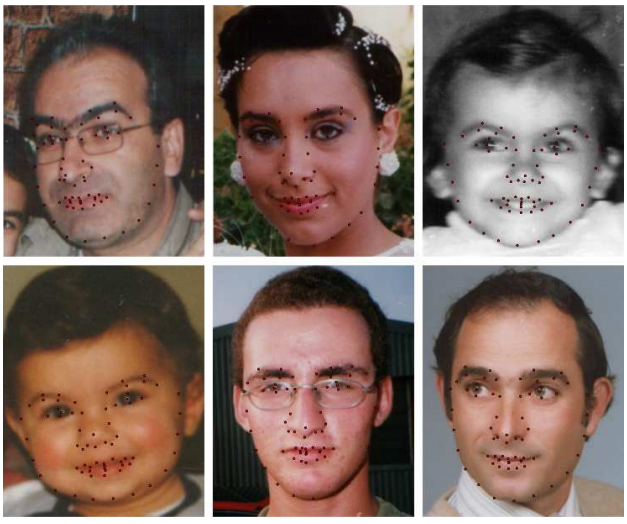

Figure 1. Labeled face example from FG-NET database.

Every set of that points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is a representation of shape. Each shape vector then can be formed in the following way:

$$\mathbf{x} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^\top \quad (1)$$

Then the shape model can be built. A shape of each object has to be independent of the position, orientation, and scale. Thus we need to align the shapes in the iterative way to minimize the sum of squared distances between each shape and the mean:

$$D = \sum |\mathbf{x}_i - \overline{\mathbf{x}}|^2 \tag{2}$$

To do that only rotation, scaling, and displacement operations are allowed. Next, principal component analysis (PCA) is applied to the set of shapes and we can finally obtain the entire shape model:

$$\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \tag{3}$$

where $\overline{\mathbf{x}}$ is a mean shape, $\mathbf{P}_s$ is a matrix containing eigenvectors of covariation matrix, and $\mathbf{b}_s$ is a set of the shape parameters, which we will use.

To build a statistical model of the gray-level appearance we should warp each sample to the mean shape (using a triangulation algorithm and bilinear approximation). An example of aligned textures is shown on the Fig. 2. They seem quite unusual, but that's a good way to ensure independence from pose. In order to obtain texture vectors **t**, we should collect all the gray intensity values from warped images over the main shape region. After that the texture vectors can be aligned to the mean in the same way as the shapes at equation 2 using described operations.

Figure 2. Aligned textures obtained from the AAM implementation.

As the next step, PCA is applied in order to get the entire texture model:

$$\mathbf{g} \approx \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{4}$$

Finally, we should concatenate vectors $\mathbf{b}_s$ and $\mathbf{b}_g$ and perform PCA once again to obtain appearance parameters.

2.2 Bio-inspired features

The bio-inspired features (BIF) (Mu et al., 2009) were created for approximate modeling of object recognition process in the cortex. The model contains alternated simple and complex layers creating increasing complexity as the layers progress from the primary visual cortex to inferior temporal cortex. The BIF model designed for the age estimation problem contains only one simple (S_1) and one complex (C_1) layers. The input is a gray-scale facial image.

The simple layer is created by convolving an array of Gabor filters (equation 5, 6 and 7) at 4 different angles θ_i and 8 pairs of different scales, according to the table in the article (Mu et al., 2009).

$$G_i(x,y) = \exp\left(-\frac{(X^2+\gamma^2Y^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}X\right) \tag{5}$$

$$X = x\cos\theta_i + y\sin\theta_i \tag{6}$$

$$Y = -x\sin\theta_i + y\cos\theta_i \tag{7}$$

The complex layer consists of MAX and STD operations. Firstly, each pair of scales are combined using MAX operator:

$$F_i = \max\left(x_i^{j1}, x_i^{j2}\right) \tag{8}$$

where x_i^{j1} and x_i^{j2} are results of convolution with a pair of Gabor filters (at s scale) and $\max$ – pixel-wise maximum operation.

Next, STD operation is performed:

$$std = \sqrt{\frac{1}{N_s \times N_s} \sum_{i=1}^{N_s \times N_s} \left(F_i - \overline{F}\right)^2}, \tag{9}$$

where $\overline{F}$ – mean value of convolved images in $N_s \times N_s$ neighborhood.

Finally, we should gather all the features (std) in a vector (that is our new representation for each image) and apply PCA to get a smaller dimensional representation of the features.

3. PROPOSED APPROACH

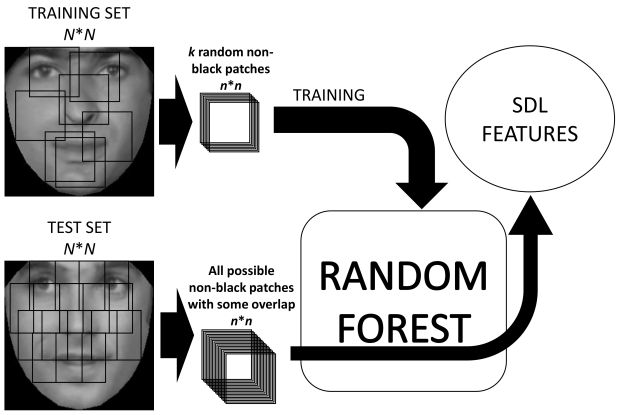

Figure 3. The supervised deep learning pipeline.

We propose to use the supervised deep learning features to improve prediction accuracy (Fig. 3). First of all, we should obtain a large set of small patches (for example, k patches from every sample). Let n be the size of the patches, every patch can be presented as an n^2 vector.

Next, the random forest (Breiman, 2001) with K trees is built. Vectors obtained above are the input, and classes (ages in our case) are the output. Let $T = \{t_i, i = 1, \dots, M\}$ be set of all terminal nodes of that forest.

Then we can calculate the new representation for each new image using listed algorithm (Fig. 4):

1. Extract all possible patches from a new image: $\left\{x_i, i = \overline{1, N}\right\}$
2. Push each patch x_i through the random forest in order to obtain the terminal nodes has been activated by it $\{T_i \subset T\}$
3. Calculate new representation H of length M using following formula:

$$H_i = \frac{|\{T_k, t_i \in T_k\}|}{\sum_j |\{T_k, t_j \in T_k\}|} \tag{10}$$

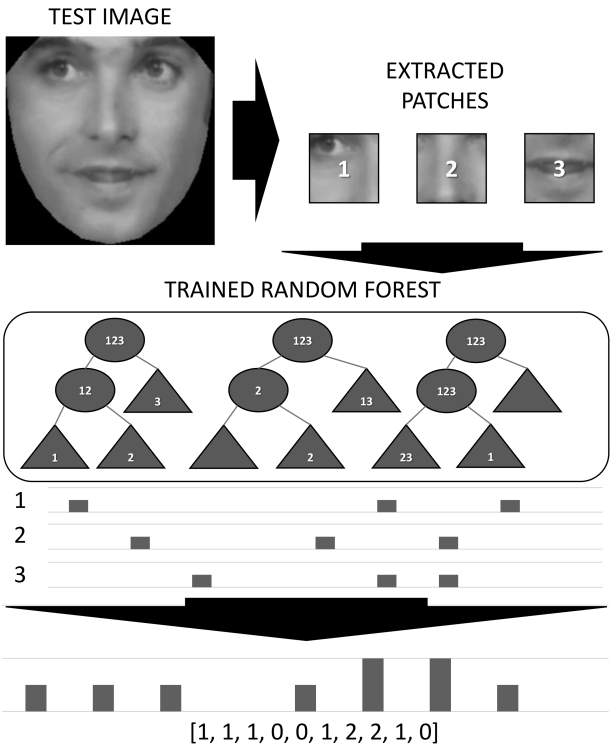

Figure 4. Histogram calculation.

The calculated histograms are the new feature representation.

4. EXPERIMENTS

4.1 Database

FG-NET (Panis et al., 2016) database was used to estimate the accuracy. The database contains 1002 images with wide variations of pose, expression, and lighting. There are 82 different subjects who contributed from 6 to 18 images with ages ranging from newborns to 69 years old subjects.

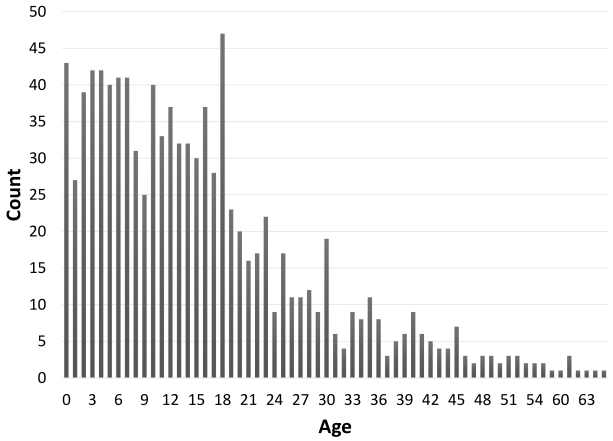

Figure 5. Age distribution in the FG-NET database.

FG-NET database also has 68-point annotations for the images (Fig. 1 shows an example).

4.2 Bio-inspired features

Enhanced bio-inspired features (Eldib and El-Saban, 2010) were implemented. The only difference from the original algorithm is the input data: 100×100 pixel aligned textures obtained from the AAM implementation instead of unprocessed images (Fig. 2 shows an example).

4.3 Supervised deep learning

200×200 pixel aligned textures that were obtained from the AAM implementation and smoothed by Gaussian filter (diameter = 9, σ = 3.2) were used as the input images. 1000 11×11 pixel patches from each image (from training set) were extracted and the random forest with 40 trees of depth 6 was built. Further, we created the histograms collecting all the possible patches from each image. Obtained histograms were concatenated with other features and then PCA was applied to reduce feature space dimension.

4.4 Classifier

To solve classification problem, we built a cascade of binary classifiers. Classifier $f_i(\mathbf{x})$ returns 1 if it decides that the subject on an image is older than i years and 0 otherwise. Then we can define the final classifier in the following way:

$$f(\mathbf{x}) = \sum_{i=0}^{68} f_i(\mathbf{x}) \tag{11}$$

Each classifier is an RBF SVM classifier (Cortes and Vapnik, 1995). We used cross-validation algorithm to determine optimal parameters for each classifier.

4.5 Error calculation

We evaluated the algorithms using the leave-one-person-out scheme. All images of each subject were used for testing, whereas all remaining samples formed the training data. Random forests were

also fit independently for each person. Mean Absolute Error (MAE, in ages) was used to measure the model accuracy. That scheme assures a much more stable evaluation since there were no subject dependencies between the training and testing sets.

4.6 Results

Results on the Fig. 6 show that using the supervised deep learning features can be used to improve an accuracy. With an active appearance model, the supervised deep learning features gave the 4% gain (an error decreased from 4.59 to 4.41). However, adding the features to the bio-inspired features didn't change an error (4.35). The reason is that both the bio-inspired features and the supervised deep learning approaches share similar ideas: they try to analyze images through edges, lines, and gradients.

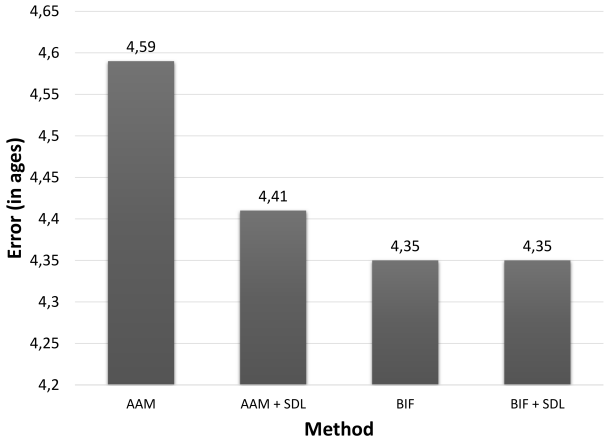

Figure 6. Results of the experiment.

5. CONCLUSION

In this paper, we proposed an innovative algorithm for human age estimation problem. The key idea of the algorithm is to use the supervised deep learning features to improve an accuracy of existing approaches. Two basic approaches were implemented: an active appearance model and the bio-inspired features. Experiments demonstrated that adding the supervised deep learning features may improve accuracy of some basic models.

REFERENCES

Breiman, L., 2001. Random forests. *Machine Learning* 45(1), pp. 5–32.

Cootes, T., Edwards, G. and Taylor, C., 1998. Active appearance models. In: *Proc. European Conference on Computer Vision*, pp. 484–498.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning* 20(3), pp. 273–297.

Drobnyh, K., 2016. *Using Unsupervised Deep Learning for Human Age Estimation Problem*. Springer International Publishing, Cham, pp. 443–450.

Eldib, M. Y. and El-Saban, M., 2010. Human age estimation using enhanced bio-inspired features (ebif). In: *ICIP*, IEEE, pp. 1589–1592.

Kwon, Y. H. and Lobo, N. D. V., 1999. Age classification from facial images. In: *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 762–767.

Martyanov, V., Polovinkin, A. and Tuv, E., 2012. Image classification with codebook based on decision tree ensembles. In: *Proc. 9th International Conference Intelligent Information Processing*, pp. 480–482.

Mu, G., Guo, G., Fu, Y. and Huang, T., 2009. Human age estimation using bio-inspired features. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 112–119.

Panis, G., Lanitis, A., Tsapatsoulis, N. and Cootes, T. F., 2016. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics* 5(2), pp. 37–46.