Component Biologically Inspired Features with Moving Segmentation for Age Estimation

Gee-Sern Jison Hsu, Yi-Tseng Cheng
National Taiwan University of Science and Technology
jison@mail.ntust.edu.tw, rouge@hotmail.com

Choon Ching Ng, Moi Hoon Yap
Manchester Metropolitan University
choonching5u@gmail.com, M.Yap@mmu.ac.uk

Abstract

We propose the Component Bio-Inspired Feature (CBIF) with a moving segmentation scheme for age estimation. The CBIF defines a superset for the commonly used Bio-Inspired Feature (BIF) with more parameters and flexibility in settings, resulting in features with abundant characteristics. An in-depth study is performed for the determination of the parameters good for capturing age-related traits. The moving segmentation is proposed to better determine the age boundaries good for age grouping, and improve the overall performance. The proposed approach is evaluated on two common benchmarks, FG-NET and MORPH databases, and compared with contemporary approaches to demonstrate its efficacy.

1. Introduction

Facial age estimation is one of the central concerns in face image and video analysis. Its application scope covers surveillance, access control, viewer discrimination and cross-age recognition [1]. Classification, regression and a hybrid of both are among common approaches for facial age estimation [2, 3, 4, 5, 6, 7]. A multi-linear regression approach with age features extracted by a subspace learning algorithm is proposed in [8]. This approach is trained and evaluated on a proprietary database, UIUC-IFP, instead of the common FG-NET [9] or MORPH [10], making it difficult to duplicate and compare to other methods. To learn a regression-based estimator in the presence of age label noises, a multi-instance regression algorithm is developed in [3]. This approach aims at facial age mining from web images and videos. It yields MAE (Minimum Absolute Error) 8.37 on FG-NET and 6.06 on MORPH, without taking any images from both databases as training samples. The hybrid approaches often use a classifier to segment faces into a few age groups, and estimate the ages of the faces in each group by regression. Most age boundaries, which refer to the ages at the boundaries between age groups, are postulated in an ad-hoc way without much interpretation, and are different one another in different works. We propose an approach with moving segmentation windows to better determine age boundaries.

Quite a few hybrid approaches come with a different number of age groups assumed at the classification phase. The following is a summary of approaches with different numbers of age groups with various age boundaries:

1. 2 Groups: Faces are classified into young and adult groups with age 20 or 21 as the boundary in [5], and each group is then processed by the SVR (Support Vector Regression) for age estimation.

2. 3 Groups: Three age groups are considered in [11, 4] with boundary ages at 10 and 20 in [12], 19 and 60 in [11], and 15 and 30 in [4]. The study in [12] offers a comparison of different classifiers with different settings, leading to the determination of classifiers appropriate for age estimation. Using the BIF (Bio-Inspired Feature) and manifold learning for face representation and SVM (Support Vector Machine) for age estimation, it is verified in [11] that age can be better estimated on smaller age groups of the same gender. In [4] five classifiers with facial landmark based features are combined using the majority rule for classification, followed by the RVM (Relevance Vector Machine) for regression. This approach shows 6.2 years in the MAE on the FG-NET database.

3. 4 Groups: Four age groups are considered in [13, 14] with 1, 16, 50 years as the boundary ages in [13], and 29, 49, 69 in [14]. A fuzzy LDA (Linear Discriminant Analysis) approach with Gabor features is proposed in

[13] for age classification of consumer facial images in uncontrolled conditions. It is concluded that the fuzzy LDA followed by quadratic regression function reveals a superb performance on uncontrolled images. However, because the uncontrolled images were collected from the internet and subjectively tagged with an age manually, the result can be hardly verified and duplicated.

4. Groups with constant age gaps: Age groups with a constant age gap are considered in [15] and [2], but with different features. The LBP (Local Binary Patterns) is used in [15] and the BIF is proposed in [2]. Two major groups and four subgroups are considered in [6, 7], where facial components are extracted as additional features processed by a SVM-based two-level binary decision tree (BDT) for classification and SVR for regression. The coarse-to-fine classification performed by the BDT and within-group regression makes the MAEs on four databases lower than many previous approaches. However, the authors have not provided specific information on the determination of the boundaries between age groups, making the duplication of this work difficult.

We propose the Component Bio-Inspired Features (CBIFs) extracted at component regions defined by facial landmarks to capture age-related characteristics. Compared to the BIFs which are popular age features [2, 11, 6, 7], the CBIFs embed more sophisticated and flexible settings on the model parameters, allowing more complex characteristics to be captured. The CBIFs define a superset of the BIFs, and BIFs can be considered a special case of the CBIFs. It is experimentally proven in this study that the CBIF outperforms the BIF on both FG-NET and MORPH databases.

The novelties of this study are twofold.

1. A simple yet effective approach using moving segmentation windows is proposed to better define age boundaries, which in the past were assumed in some ad-hoc way without much interpretation.

2. The Component Bio-Inspired Feature (CBIF) extracted at regions defined by facial landmarks is proposed to better capture facial age characteristics than the common BIF.

The rest of this paper is organized as follows: the extraction of CBIFs is presented in Sec. 2, followed by the moving segmentation windows in Sec. 3. Sec. 4 presents an experimental study on two popular benchmark databases, FG-NET and MORPH, including the performance comparison with contemporary approaches. As deep learning has recently demonstrated great success handling computer vision problems, we use the CBIF as the input to a convolutional neural network (CNN) and compare its performance with the support vector based classification and regression. The comparison is reported in Sec. 4 as well. A conclusion of this study is given in Sec. 5.

2. Component Biologically Inspired Features

The Component Bio-Inspired Feature (CBIF) defines a superset for the BIF and allows more flexible settings for capturing the appearance characteristics with different scales, orientations, locations and wavelengths. We extract the CBIFs from the local regions defined by the facial landmarks, which are automatically located using the Regressive Tree Structured Model (RTSM) [16]. The RTSM is selected because it handles both face and landmark detection in a unified model with sufficient speed and accuracy. We propose three forms of CBIF, namely the *component-based* obtained from each partition of the face, the *concatenated* that combines all partitions and the *truncated* with low-response components removed.

2.1. Extraction of Landmark-Oriented CBIF

Given an image patch, the extraction of CBIF requires a pyramid of convolution filters and an associated pyramid of pooling windows, which correspond respectively to the simple processing layers and complex processing layers commonly seen in the BIF settings. Each layer of the convolution filter pyramid is composed of a set of Gabor filters of neighboring sizes and same orientation. Specifically, one can define J_i Gabor filters for Layer-i, which can be written in the layer parameter $\omega_{i,j,k} = [s_{i,j}, \gamma_{i,j}, \theta_{i,k}, \sigma_{i,j}, \lambda_{i,j}]$,

$$G\left(x, y; \omega_{i,j,k}\right) = \exp\left(-\frac{\left(X^2 + \gamma_{i,j}^2 Y^2\right)}{2\sigma_{i,j}^2}\right)\cos\left(\frac{2\pi}{\lambda_{i,j}}X\right) \tag{1}$$

where $X = x\cos\theta_{i,k} + y\sin\theta_{i,k}$, $Y = -x\sin\theta_{i,k} + y\cos\theta_{i,k}$, and (x, y) denotes the coordinate of a point on the filter. The filter is of size $s_{i,j}$ with aspect ratio $\gamma_{i,j}$. $\omega_{i,j,k} = [s_{i,j}, \gamma_{i,j}, \theta_{i,k}, \sigma_{i,j}, \lambda_{i,j}]$, where $\theta_{i,k}$ defines the orientation, $\sigma_{i,j}$ defines the variation of the Gaussian component and $\lambda_{i,j}$ defines the wavelength of the sinusoidal component. Note that the size of the filter $s_{i,j}$ is not shown explicitly in (1) as it reveals how large the support of $G(\cdot)$ is. Although the settings in (1) allow different aspect ratio $\gamma_{i,j}$ for different size $s_{i,j}$ of the filter, we refrain the model complexity by assuming the importance along x- and y- directions are equal, making $\gamma_{i,j} = 1$ in all cases. Assuming K_i orientations considered at Layer-i, i.e., $k = 1, \cdots, K_i$, there are $J_i K_i$ convolution filters to be applied on this single layer.

The BIF in [2] is a special case of (1), with J_i selected as 2 and K_i selected as 8 for all 8 layers[1], i.e., $i = 1, \cdots, 8$, winding up 128 (=$2 \times 8 \times 8$) convolution filters overall. In such a setting the numbers of filters and orientations are the

[1] "Layer" in this context is the same as "band" in [2].

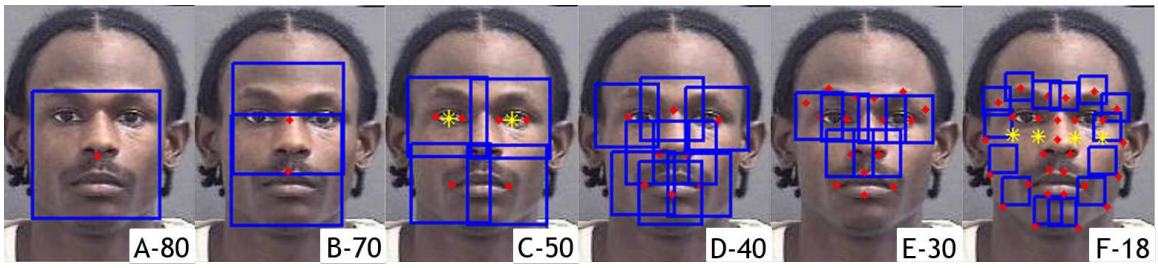

Figure 1. Landmark based face partition for the CBIF extraction. Each face is aligned to the eyes with the distance between the eyes normalized to 40 pixels. From Case A to F the number of partitioned components doubles, giving 1 to 32 components, respectively. Red dots are the RTSM landmarks used to define component regions. Most red dots are the centers of component regions, and so are the yellow asteroids, which are the mid points between a pair of red ones. Only a few components are shown in Cases E and F for better visibility. A-80 (B-70) denotes the component region in Case A (B) is 80×80 (70×70) pixels

same for all layers. The settings in (1) allow a different numbers of filters with different orientations in each stack, making it capable of capturing contents of different spatial frequencies and scales at different layers.

Given an image patch $I(m,n) \in I^{M\times N}$, each filter in $[G(x,y;\omega_{i,j,k^*})]_{j=1,\cdots,J_i}$, the stack of Gabor filters with neighboring sizes and same orientation (i.e., filters of sizes $[s_{i,j}]_{j=1,\cdots,J_i}$ and orientation θ_{i,k^*}), is convolved with $I(m,n)$, i.e.,

$$J(m,n;i,j,k^*) = G\left(x,y;\omega_{i,j,k^*}\right) \otimes I(m,n) \quad (2)$$

where $\otimes$ is the convolution operator. The next step is to extract the maximum of $[J(m,n;i,j,k^*)]_{j=1,\cdots,J_i}$ at each (m,n), i.e., the maximum of the J_i outputs from the stack of the filters of neighboring sizes and same orientation. This can be expressed as follows:

$$J(m,n;i,k^*) = \mathrm{MAX}_j\left([J(m,n;i,j,k^*)]_{j=1,\cdots,J_i}\right) \quad (3)$$

where MAX_j is the maximum across each stack. The above processing applies for all orientations $[k^*]$. Without loss of generality we write $J(m,n;i)$ as the result of the above convolution and max pooling, with k^* dropped as all orientations are processed in the same manner.

The corresponding layer of the pooling window pyramid, denoted as P_i, is composed of an $n_i \times n_i$ mask that scans $J(m,n;i)$ from left to right and top to bottom with stride step l_i. At each scan P_i computes the standard deviation of the patch on $[J(m,n;i)]_{m,n}$ covered by the $n_i \times n_i$ mask. The standard deviations from different patches are then concatenated into a vector $\mathbf{s}_I$,

$$\mathbf{s}_I = \left[STD\left(P_i(n_i,l_i), J(m,n;i)\right)\right]_i \quad (4)$$

where $STD(\cdot)$ denotes the standard deviations extracted from the patches on the i-th layer maxima $J(m,n;i)$ using the scanning mask defined by $P_i(n_i,l_i)$.

Different local responses are revealed when running the convolution filters and pooling windows of different scales and orientations across the image I. Large responses are observed at regions with similar spatial frequencies and scales as of the convolution filters; and insignificant responses are obtained at regions with spatial frequencies different from those of the convolution filters. It is found in our experiments that imposing a threshold on the local responses, i.e., on the elements of $\mathbf{s}_I$, can help the extraction of significant local responses and remove the insignificant counterparts. We have tested different thresholds imposed on $\mathbf{s}_I$ and selected the one with the best performance on the training set, and call it the *truncated* CBIF. See Sec.4 for experiments on the original and truncated CBIFs.

2.2. Landmark-Oriented CBIF Extraction

The CBIF features are extracted from the regions defined by the landmarks obtained using the Regressive Tree Structured Model (RTSM) [16]. The RTSM is chosen because it simultaneously solves the face detection and landmark localization, and it can handle faces with large head rotations. These two virtues make the RTSM different from most landmark localization algorithms, including the latest ones [17, 18], which require a face detector to localize faces first and only work for a limited pose range. The RTSM is composed of a coarse Tree Structured Model (c-TSM), a refined TSM (r-TSM) and a BSVR (Bidirectional Support Vector Regressor). The c-TSM is designed for fast detection of face candidates which are further processed by the r-TSM for precise landmark localization. The c-TSM is built on low-dimensional images with a small number of parts, and the r-TSM is built on high-dimensional images with more parts. Both are built following the TSM architecture [19]. The BSVR estimates the dense set of landmarks using the regression with r-TSM landmarks as reference inputs.

To extract the CBIF features from faces with large rotations, e.g., from some samples in the FG-NET [9], we locate the landmarks using the RTSM and perform the 3D reconstruction of the face using the approach in [20] that uses a single 3D face scan as the initial depth reference. The

CBIFs are then extracted from the frontal pose of the reconstructed face. There are 68 landmarks located using the RTSM, but we only use 32 of them. The landmarks are used to partition a face into same size local regions. The CBIFs $\mathbf{s}_I$ are extracted from each local region and concatenated to form the CBIF for the face. Experiments were carried out on different numbers of local regions, from 1 to 32, as shown in Fig. 1.

3. Moving Segmentation and Hierarchical Classification

Given a training set with CBIFs extracted, our approach consists of the determination of boundary ages using moving segmentation, and the determination of hierarchical configuration. Given a training set D_t and a validation set D_v with the youngest age y_m and oldest age y_M, the moving segmentation consists of the following steps:

1. Select the initial age coverage Δ_0 and two subsets, Young (Y) and Senior (S), from D_t with age segment $[y_m, (y_m + \Delta_0)]$ in the former and $[(y_m + \Delta_0 + 1), (y_m + 2\Delta_0 + 1)]$ in the latter. $(y_m + \Delta_0)$ is considered the age boundary. Train the binary classifier and compute the misclassification rate on D_v.

2. Move the Y and S subsets up with an additional year, i.e., the age segments become $[(y_m+1), (y_m+1+\Delta_0)]$ and $[(y_m + \Delta_0 + 2), (y_m + 2\Delta_0 + 2)]$, respectively. Repeat the training and computation of the misclassification rate in Step 1, and move on to the next Y and S subsets and repeat.

3. Increase the age coverage Δ_0 by an increment δ_1, $\Delta_1 = \Delta_0 + \delta_1$, and repeat Steps 1 and 2. Repeat for more increments δ_i.

4. Compute the weighted average of the misclassification rates at each age boundary over multiple age coverages Δ_i's considered. The weight exploited in this study is normalized $\sqrt{1/\Delta_i}$, which is obtained empirically.

5. Select the age y_0^* and age coverage Δ_0^* that give the local minimum weighted average of misclassification rate. This step ends up with a Y segment in $[y_m, y_0^*]$ and an S in $[(y_0^* + 1), y_M]$.

6. Select two more minimum weighted average points of misclassification rate so that YY (Young-Young), YS (Young-Senior), SY (Senior-Young) and SS (Senior-Senior) segments can be identified with age boundaries at $y_{1,y}^*$, y_0^* and $y_{1,s}^*$.

7. Same as Steps 6, but with more minimum weighted average points of misclassification rate so that it ends up with 8 segments, namely YYY, YYS, YSY, ..., SSY and SSS. If a segment is less than 4 years, it is merged with the shorter neighboring segment.

Table 1. Parameters Determined for CBIF: 4 layers of filtering and pooling pyramids, 2 Gabor filters each layer, the base σ_j and base λ_k are for the first Gabor filter, and both multiplied with Ratio factors right below for the second Gabor filter

	Complex layers		Simple layers		
	Pool Win. $l_p \times l_p$	Overlap s_p	Gabor fil. size s_l	Base σ_j Ratio	Base λ_k Ratio
Layer 1	6×6	3	4×4 6×6	1.6 1.5	2.0 1.5
Layer 2	8×8	4	8×8 10×10	3.2 1.3	4.1 1.2
Layer 3	10×10	5	12×12 14×14	5.0 1.2	6.3 1.2
Layer 4	12×12	6	16×16 18×18	6.9 1.1	8.6 1.1

According to our experiments on the FG-NET and MORPH, age grouping affects age estimation, and it must work with an appropriate number of age groups. More or less than the needed groups tend to degrade the performance. We have found that 4-group segmentation slightly outperforms 2-group segmentation, and both outperform 7- and 8-group segmentation with a clear gap in MAE. Experimental details are reported in Sec.4.

Given the age boundaries determined by the moving segmentation window, we found that the classification accuracy varies with different hierarchical configurations. We have run an exhaustive search for determining better configurations. Fig.2 shows several cases with 4-group segmentation that result in low MAEs. The comparison of these hierarchical settings is reported in Sec.4.

4. Experiments

4.1. Experimental Setup

Two benchmark databases, the FG-NET [9] and MORPH [10], were used in our experiments. The FG-NET database consists of 1,002 images of 82 individuals. The age of subjects in FG-NET ranges from 0-69 years, but over 50% of the subjects are between the ages 0 and 13. Our experiments on the FG-NET used the entire dataset and followed the subject-independent Leave-One-Person-Out (LOPO) protocol. The MORPH database is the largest publicly available longitudinal face database. It is a collection of mugshot images, including meta data for race, gender, date of birth, and date of acquisition. Our experiments were performed on MORPH Album-2 which contains 55,134 images of 13,000 individuals. Each face from FG-NET or MORPH was first converted to gray scale, and aligned to the eyes with the distance between the eyes normalized to 40 pixels. We performed the experiments on

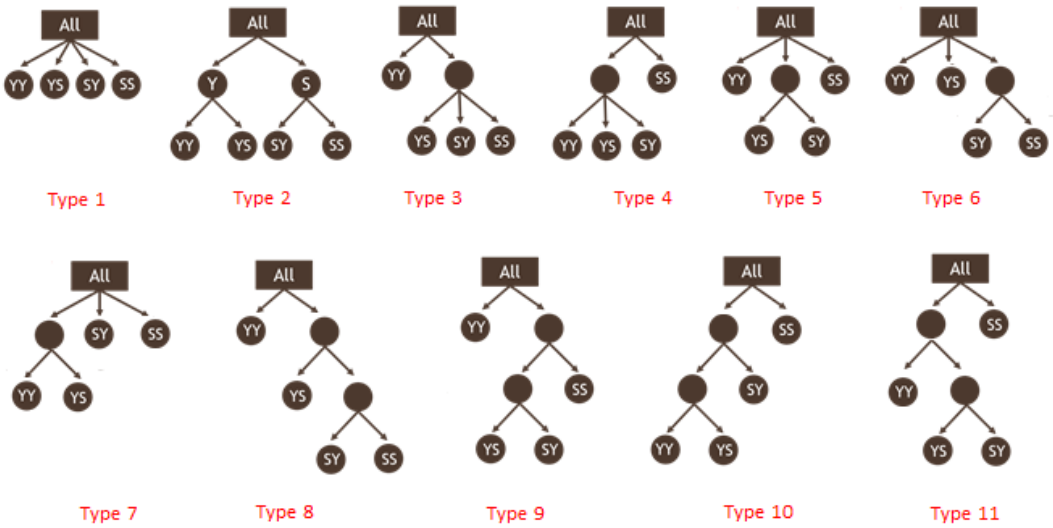

Figure 2. Configurations for the hierarchical combinations of binary and multiple classifiers, considering 4-group segmentation with segments YY, YS, SY and SS.

Table 2. MAEs on FG-NET with CBIFs of different partitions and the concatenated CBIF of all partitions, in both the original and truncated feature form

FG-NET	1	2	4	8	16	32	Concat.
Original	4.31	4.00	3.74	3.86	4.00	4.06	3.63
truncated	4.3	3.95	3.78	3.76	3.99	4.05	3.55

Table 3. MAEs on MORPH with CBIFs of different partitions and the concatenated CBIF of all partitions, in both the original and truncated feature form

MORPH	1	2	4	8	16	32	Concat.
Original	4.24	4.20	3.86	3.8	3.93	3.82	3.42
Truncated	4.22	4.00	3.79	3.65	3.88	3.80	3.39

MORPH with ten-fold cross-validation.

All experiments were run on Matlab upon a Windows PC with CPU 3.6 GHz and RAM 16 GB, and were designed to study the following issues:

- The CBIF settings: The estimation accuracy varied with two vital settings. 1) The landmark-based partition of the face: The face was partitioned into 1, 2, 4, 8, 16 and 32 local regions, as shown in Figure 1. 2) The CBIF parameters: Many parameters would alter the feature representation, including the number of layers n_l, the number of Gabor filters in each layer (i.e., J_i for Layer-i) and the sizes of the filters $[s_{i,j}]_j$ for Layer-i, the parameters σ_j and λ_j for each filter, the number of orientations K_i, the pooling window size l_p and overlap s_p when moving across. We allowed one to vary while keeping others as presumed, and selected the best as the presumed and repeat on the rest.

- The number of age groups, the associated boundary ages, and the configuration of hierarchical classification.

4.2. Results and Comparison

The above parameters and variables were determined in a recursive manner. A set of preselected CBIF parameters was used to determine the face partition and age boundaries, which were then used to revise the CBIF parameters. The revised CBIF parameters were then used to reselect the face partition and age boundaries. We started with an initial set of CBIF with 6-layer pyramids, 2 Gabor filters each layer, and 8 orientations $\times$ 5 scales for each Gabor filter; and ended up with 4-layer pyramids, 2 Gabor filters each layer, and 10 orientations $\times$ 5 scales for the Gabor filter. Table 1 shows the details of the determined CBIF parameters, and the following were the experiments conducted. Note that although the FG-NET and MORPH are different in ethnic background, image quality, number of samples and age range distribution, the CBIF parameters considered appropriate reveal similar variations in the associated MAEs.

Face Partition and Associated Feature Forms:

The MAEs with the six partitions (Figure 1) and the concatenated CBIF are shown in Table 2 for FG-NET and Table 3 for MORPH. The partition with 8 components outperforms other partitions on both FG-NET and MORPH, but the concatenated CBIF outperforms all partitions with a clear gap in the MAEs. Note that, however, the concatenated truncated CBIF outperforms the concatenated original CBIF, showing that the CBIFs are better extracted when the feature components with insignificant responses are removed. Table 4 shows the MAEs of truncated CBIF cases with different thresholds applied for removing the low-

Table 4. MAEs with truncated CBIFs obtained by applying different thresholds for removing CBIF components with low responses

	5	10	15	20	30	40
FG-NET	3.57	3.56	3.55	3.73	3.77	3.89
MORPH	3.44	3.43	3.39	3.46	3.56	3.69

Table 5. MAEs with different numbers of layers in the convolution pyramid for extracting CBIF

No. layers	1	2	3	4	5	6
FG-NET	4.30	3.91	3.73	3.55	3.93	4.27
MORPH	3.86	3.52	3.42	3.39	3.40	4.02

Table 6. MAEs with different numbers of Gabor filters in each layer

No. filters	1	2	3	4	6
FG-NET	3.56	3.55	3.75	3.83	3.95
MORPH	3.46	3.39	3.48	3.58	3.71

response CBIF components. The dimension of the original and truncated CBIF is 70740 and 55770, respectively, associated with MAE 3.63 and 3.55 and processing time 0.398 and 0.244 per face. In the following context, we stick to the concatenated truncated CBIF.

Numbers of Layers, Filters and Orientations:

The MAEs with different numbers of layers are shown in Tab.5. When the layers increases from one to four, the performance improves, observed on both FG-NET and MORPH test sets. However, the improvement halts when the layers exceed four, and it starts to degrade gradually for increasing layers.

The performances with different J_i, the number of Gabor filters with neighboring sizes and of the same orientation for the MAX extraction at each layer, are shown in Table 6. The case with two Gabor filters outperforms others on both benchmarks, although the difference from the case with a single Gabor filter on FG-NET is marginal.

The comparison on different numbers of orientations K_i is shown in Table 7. More orientations appear to yield lower MAEs, showing that the responses to the filters with more orientations reveal more age-related traits. However, when it exceeds 10, the MAE slightly increases.

Determination of Age Boundaries:

We selected the local minima in the weighted average of the misclassification rates obtained by running the moving segmentation windows with age coverages from 2 to 5 years. The result for FG-NET is shown in Figure 3, where the red line denotes the weighted average. Because the number of samples decreases sharply for ages over late 30s in FG-NET, the dotted lines show for the ages with insufficient (< 5 images) samples. The boundary ages, shown in little circles on the red line, are selected at 11, 28 and 38

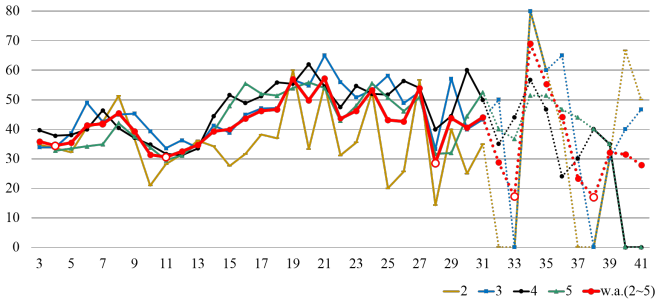

Figure 3. Age boundaries determined on FG-NET (age range: $0 \sim 69$) using the moving segmentation window with age coverages from 2 to 5. Red line shows the weighted average, and dotted lines are for insufficient number of training samples (< 5 images each age). The age boundaries are annotated with "o". Ages 11, 28 and 38 are taken as primary boundaries and 4, 33 are considered secondary boundaries.

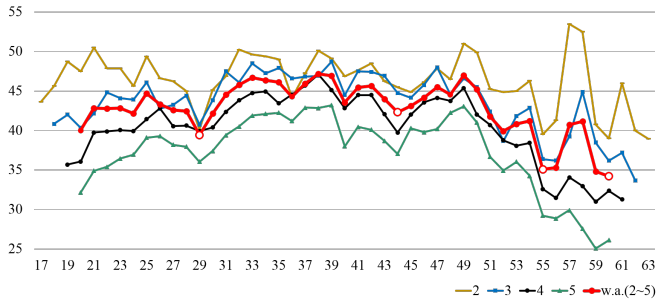

Figure 4. Age boundaries determined on MORPH (age range: $16 \sim 77$) using the moving segmentation window with age coverages from 2 to 5. Red line shows the weighted average. The age boundaries are annotated with "o". Ages 29, 44 and 60 are taken as primary boundaries and 55 is considered the secondary boundary.

Table 7. MAEs with different numbers of orientations in the Gabor filters

	4	6	8	10	12	18	24
FG-NET	3.95	3.89	3.63	3.55	3.71	3.89	3.76
MORPH	3.70	3.63	3.61	3.39	3.42	3.63	3.60

to segment the ages into YY (Young-Young), YS (Young-Senior), SY and SS. Since no local minimum is observed in the segment YS and the samples at the segment SS are insufficient for training, only 4 and 33 are selected to further segment the YY into YYY and YYS, and the SY into SYY and SYS, respectively. We compare the 4-group (YY, YS, SY and SS) and 6-group (YYY, YYS, YS, SYY, SYS and SS) settings. The boundary ages given by moving segmentation are determined completely on the training set. If the majority of the training set is within some age range, the boundary ages can be located within or close to that age range. This is also observed on MORPH, which has a sufficient number of samples across a wider age range.

Table 8. Comparison of Different Combination of Age Group Classification and Regression in MAE: Reg. Only refers to the case without age grouping; 2-group refers to ages grouped into Y and S; 4-group for YY, YS, SY and SS; 7-/8-group for YYY, YYS, ..., SSS wherever applicable

	Reg. Only	2-Group	4-Group	5-/6-Group
FG-NET	5.72	3.92	3.87	5.63 (6-Group)
MORPH	4.25	4.00	3.98	4.28 (5-Group)

Table 9. Comparison of Hierarchical Combinations of Binary and Multiple Classifiers: MAEs for the Five types shown in Fig.2

Type	1	2	3	4	5	6
FG-NET	3.75	3.55	3.62	3.57	3.66	3.61
MORPH	3.54	3.39	3.46	3.41	3.49	3.43

Type	7	8	9	10	11
FG-NET	3.69	3.59	3.53	3.63	3.58
MORPH	3.54	3.39	3.46	3.41	3.49

The moving segmentation applied on MORPH gives 44 parting Y from S; 29 parting YY from YS, 60 parting SY from SS. As no local minimum is observed in the segments YY and YS, only 55 is selected for parting SYY from SYS. The segment SS is not further segmented because of insufficient samples. We therefore compare the 4-group (YY, YS, SY and SS) and 5-group (YY, YS, SYY, SYS and SS) settings on MORPH. All boundary ages are shown in little circles in Figure 4.

Comparison on Age Groupings:

The experiments were carried out to compare the following cases: 0 (without age grouping), 2, 4 and 5 (or 6, if applicable) groups. The MAEs are shown in Table 8. The cases with no age grouping and 5-/6-Group are apparently outperformed by the 2-Group and 4-Group, and the 4-Group slightly outperforms 2-Group. This reveals that age grouping helps to improve age regression, but excessive grouping would degrade the performance.

Hierarchical Classification Configuration:

Given the boundary ages, different hierarchical settings on the combination of binary and multiple classifiers, as shown in Figure 2, yield different accuracy. Given the 4-Group segmentation, the eleven configurations shown in Figure 2 were tested. The MAEs are given in Table 9, showing that Type 2 performs the best.

Comparison with a Deep Learning Framework and Contemporary Approaches:

Deep learning has recently demonstrated great success in solving computer vision problems [23, 24]. It is of great interest to compare the performance of the CBIF on a deep learning framework with the performance using support vector based classification and regression. The CNN exploited in this comparison study is similar to the one used by Parkhi et al. [24] which shows superb performance for face recognition. It is composed of 9 blocks. Each of the first 5 blocks consists of 2 or 3 convolution layers and one max pooling layer, the 6th block is for dropout operation, followed by 2 fully connected blocks, and the last is the output block with softmax processing. See [24] for details. To be able to connect to the input layer of the CNN, the extraction of the CBIF follows the convolution with Gabor filters and MAX/STD pooling, as addressed in Sec. 2.1, but maintains the two dimensional feature matrix as the filter moving across each training sample. Each training sample is entered with its age as the output label, and the CNN is trained for classification.

Table 10. Comparison with Contemporary Approaches

Publication	FG-NET	MORPH
Guo et al.(2009) [2]	4.8	N/A
Luu et al.(2011) [21]	4.1	N/A
Chang et al.(2011) [22]	4.5	6.1
Kohli et al.(2013) [5]	3.9	N/A
Hu Han et al.(2013) [6]	4.6	4.2
Hu Han et al.(2015) [7]	3.8	3.6
Proposed (truncated CBIF)	3.38	3.21
Proposed (CBIF+CNN)	N/A	2.58

It is well known that the CNN requires a huge amount of training data to make the large number of parameters in the network converge. In case of insufficient data, its performance can be quite problematic and unreliable. Using the same evaluation protocols stated in Sec. 4.1, the CNN gives MAE 12.2 on the FG-NET and 2.58 on MORPH. As the FG-NET only has 1002 samples, the CNN is not considered a valid solution to handle dataset of this size. Figure 5 shows the comparison on the performance variation with the size of training samples, on the MORPH database. Although the CNN gives a better MAE on MORPH, it requires a sufficiently large number of data to guarantee the perfor-

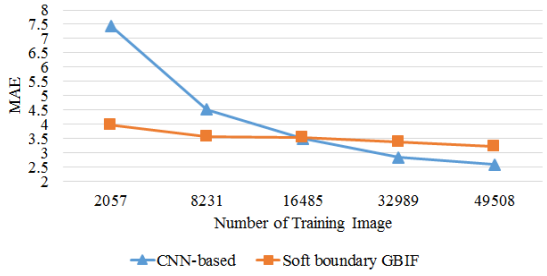

Figure 5. Comparison with the CBIF on CNN. The performance of the soft boundary framework, denoted in orange boxes, improves marginally as the number of training samples increases, showing that non-deep approaches can be effective for limited training data. Although the CNN, denoted in blue triangles, performs much better for cases with sufficient training data, it cannot handle limited training data.

mance. The performance using the proposed *non-deep* approach can handle cases with limited samples. The MAE appears to be steady as size goes beyond 8000.

Table 10 shows the comparison with contemporary approaches. The truncated CBIF outperforms all on both benchmarks. The performance, according to the extensive experiments reported above, is not just contributed by the CBIF along, but also the age boundaries determined by the moving segmentation. Although the CBIF with CNN performs exceptionally well on MORPH, it cannot handle the situation that only limited data is available for training, as the case with FG-NET.

5. Conclusion

We propose the CBIF (Component Bio-Inspired Feature) and the moving segmentation for age estimation. The proposed approach is verified on FG-NET and MORPH, and has demonstrated better performance than most of previous methods. A comparison with the CBIF on a deep learning network shows that the non-deep approach can be a powerful tool to handle problems with limited data, and this also raises an issue: can deep learning be made to handle limited data? The issue with ethnic and demographic properties in the dataset also deserves special attention. We have tested the trained-on-MORPH CBIF-CNN on the FG-NET, and the MAE is 7.6, way higher than that reported on the MORPH test set (2.58). The samples with large errors are often those whose ethnic backgrounds are different than those of the majority of the MORPH. All these issues are being studied in our lab, and will be updated with new findings when available.

References

[1] B.-C. Chen, C.-S. Chen, and W. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, 2015.

[2] Guodong Guo, Guowang Mu, Yun Fu, and T.S. Huang, "Human age estimation using bio-inspired features," in *CVPR*, 2009.

[3] B. Ni, Z. Song, and S. Yan, "Web image and video mining towards universal and robust age estimator," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1217–1229, 2011.

[4] Pavleen Thukral, Kaushik Mitra, and Rama Chellappa, "A hierarchical approach for human age estimation," in *ICASSP*, 2012, pp. 1529–1532.

[5] Sharad Kohli, Surya Prakash, and Phalguni Gupta, "Hierarchical age estimation with dissimilarity-based classification," *Neurocomputing*, vol. 120, pp. 164–176, 2013.

[6] Hu Han, C. Otto, and A.K. Jain, "Age estimation from face images: Human vs. machine performance," in *ICB*, 2013, pp. 1–8.

[7] H. Han, C. Otto, X. Liu, and A.K. Jain, "Demographic estimation from face images: Human vs. machine performance," *TPAMI*, vol. 37, pp. 1148–1161, 2015.

[8] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, 2008.

[9] "FG-NET Aging Database," http://www.fgnet.rsunit.com/.

[10] K. Ricanek and T. Tesafaye, "Morph: a longitudinal image database of normal adult age-progression," in *FG*, 2006.

[11] Guodong Guo, Guowang Mu, Yun Fu, C. Dyer, and T. Huang, "A study on automatic age estimation using a large database," in *ICCV*, Sept 2009, pp. 1986–1991.

[12] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *SMC*, vol. 34, no. 1, pp. 621–628, Feb 2004.

[13] Feng Gao and Haizhou Ai, "Face age classification on consumer images with gabor feature and fuzzy lda method," in *Int. Conf. Advances in Biometrics*, 2009, pp. 132–141.

[14] Guodong Guo and Xiaolong Wang, "A study on human age estimation under facial expression changes," in *CVPR*, 2012, pp. 2547–2553.

[15] Asuman Günay and Vasif V Nabiyev, "Automatic age classification with lbp," in *23rd Int. Symp. Comp. .Inf. Sci. (ISCIS)*. IEEE, 2008, pp. 1–4.

[16] G.-S. Hsu, K.-H. Chang, and S.-C. Huang, "Regressive tree structured model for facial landmark localization," in *ICCV*, December 2015.

[17] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *CVPR*, 2013.

[18] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *CVPR*, 2014.

[19] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[20] Ira Kemelmacher-Shlizerman and Ronen Basri, "3D face reconstruction from a single image using a single reference face shape," *TPAMI*, vol. 33, no. 2, pp. 394–405, Feb. 2011.

[21] Khoa Luu, K. Seshadri, M. Savvides, T.D. Bui, and C.Y. Suen, "Contourlet appearance model for facial age estimation," in *IJCB*, Oct 2011.

[22] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *CVPR*, 2011.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.