Age estimation from face image using Hybrid Representation for Deep Learning

Faten Ahmed Jebur Al-sudani[1], Hazim Abdulameer Fadhil Al-Afare[2],
[1,2] Department of Computer Systems Techniques, Institute of Administration Rusafa, Middle Technical University
[1]fatenahmed152@gmail.com,
[2]hazim8111979 @gmail.com

Abstract

New technique of Deep learning gave Face-age estimation higher results-accuracy. Proposed estimation technique of this paper presents a novel deep learning network for age estimation called Hybrid Representation Architecture (HRA). Where local, global and global-local branches are contained and jointly optimized for capturing multi-type features combined by complementary information. In each branch, the sub-network is designed for to extract features from independent region by employing a separate loss, where recurrent fusion is used for exploring correlations among them. We consider that different fashions of pose would cause misalignment in regions variation, thus an Aligned Region Pooling (ARP) operation is designed to extract aligned-region features. To compensate the demand of large-scale image datasets, a private image dataset regarding-age progression (named Web-FaceAge) is adopted with more than 120K face images, which are captured under different scenes providing wide range of ages. Benchmarking experiments conducted on five image datasets, such as MORPH, CACD, Chalearn LAP 2015, FG-NET and Web-FaceAge, show the significant precedence of proposed estimation technique against other state-of-the-art techniques.

Keywords: Face-age estimation, Web-FaceAge, Hybrid Representation Architecture, Aligned Region Pooling, Regions of Interest.

الخلاصة

التقنية الجديدة من خلال (deep Learning) تعطي دقة عالية في تقدير العمر من خلال ملامح الوجه.

تقنية التقدير المقترحة في هذه الورقة بأستخدام Deep Learning network لتقدير العمر تسمى بالطريقة الهجينة اذ ان التفرعات الخاصة والعامة تترابط بشكل امثل لألتقاط طبقات متعددة في الخصائص باستخدام متممات البيانات. في كل فرع الشبكة الفرعية صممت لاستخلاص الخصائص من المناطق غير المعتمدة من خلال توظيف الفقدان المعزول، حيث ان الاندماج المتكرر يستخدم لاستكشاف العلاقة فيما بينهم. نحن نعتبر ان اختلاف المظهر يسبب عدم تطابق في نقاط الاختلاف وعليه فأن مستوعب المناطق المتطابقة صممت بشكل لاستخلاص الخصائص منها ولغرض مكافئة الطلب لمجموعات الصور تم اعتماد مجموعة بيانات لصور خاص تحتوي على اكثر من ١٢٠ الف صورة للوجه والتي تم التقاطها تحت مشاهد مختلفة توفر نطاقا واسعا من الاعمار.

1. Introduction

Nowadays, in spite of considerable result-accuracy yielded in age estimation, it remains a challenge to reach the best possible accuracy regarding several factors such various genes, widely different face aging patterns or even different available datasets of face images. Most of published age estimation techniques depended on either extracting local features using face descriptors like BIF [1] or using the whole face for learning a global representation [2-7]. This work proposes a new technique as a suboptimal by learning local or global features, where depending only on global context can ignore significant details such wrinkles, glasses or beard [8]. On the other hand, depending only on local details can lose crucial information in addition to the hardness of achieving a robust description. Age estimation proposed scheme in this work is built considering local and global information, which is motivated by the human visualization system which combines between them to analyze captured image and extract the effective information [9]. To do so, this work proposes the new Hybrid Representation Architecture (HRA).

This architecture adopts local and global, global-local sections, in which, local section is learned using fine-grained local details such as wrinkles. Consequently, global-section learning is performed using coarse-grained global context, and integrated configuration is used to train global-local information. For each section, several sub-networks are used to extract region-related features (local, global and global-local). For full information-exploration among them and extracting reliable age-features, following strategies are adopted: in each sub-network, a separate loss function is forced for independent region-features learning, whereas single loss function is used by the final layer for guiding the entire network. Recurrent fusion is then employed considering Long Short Term Memory (LSTM) [10] which explores, in each branch, spatial connections between different regions within the face. Final strategy jointly optimizes all branches for capturing complementary information between them. Proposed HRA differs from previously published works [11 and 12] of learning local and global features, where correlation between different face-regions is ignored and integrated features learning is not involved.

As human face has considerable structure, specific Regions of Interest (ROI), such as nose, eyes, nose… etc., can be selected depending on face-landmarks. Many published studies adopted using ROI pooling for feature-generation from adequate face region [13 and 14]. However, it may be hard to deal with face-regions misalignment results from large variations of face poses. As a result, Aligned Region Pooling (ARP) is proposed to ensure extracting the suitable features from corresponding regions, which is performed using Affine Transform specifying sources and key-points. Such way, region-misalignment can be decreased to extract higher-efficiency features from related region.

Since deep learning methods need high number of images of training, the lack of such large-scale face datasets degrades developing deep-learning-techniques for age estimation. Yet, two large-scale face datasets are currently available, IMDB- CACD [15] and WIKI [2], with up to 100K age-related images. However, IMDB-WIKI images contain noisy information such as inaccurate age labels in some image, where some other images contain more than one face or even no face

[16]. Therefore, such images are adopted in pre-training stages rather than in evaluation. CACD dataset is relatively small, where all images are captured for 2000 celebrities only with small age ranges (14- 62), where the range of MORPH II ages (16, 77) handles older ages. Therefore, a new large-scale image dataset is collected regarding age progression that contains high number of images, subjects, diverse scenes and wide age ranges.

Contributions of this work include:

i. To solve the issue of local and global features, this work proposes using HRA network to learn both of local details, global context and the integrated configuration. Due to the best of our knowledge, no previous technique integrated all of these cues in age estimation.
ii. This work proposes ARP pooling method to aligned candidate region features, which can be affected by region misalignment due to large variations of face poses.
iii. Proposing new, privately collected, age dataset (Web-FaceAge) with high number of images, wide age ranges and accurate identities.
iv. Benchmarking with other literatures adopted five datasets of face image, encouraging results are yielded.

2. Proposed Methodology

General diagram of the proposed technique is illustrated in Fig. 1. Age estimation scheme is built depending on ResNet-18 [17]. It is constructed of three sections: local section, global section and local-global section. Each section contains three sub-networks for features-extraction from different regions (local and global).

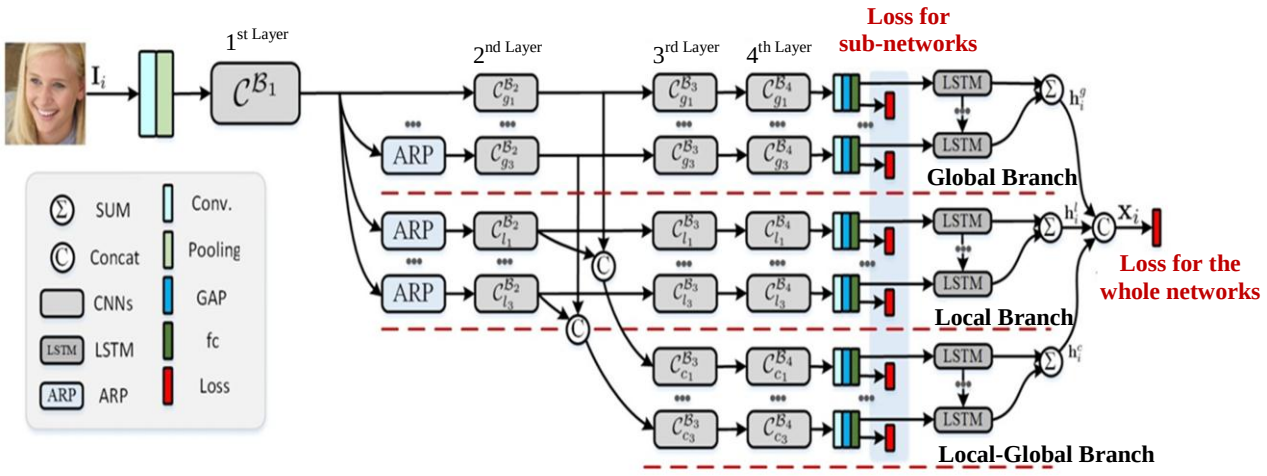

Figure 1: Proposed HRA architecture Deep-Learning age estimation, in which, Local, Global and Local-global sections are illustrated.

Global regions in this methodology represent the entire-face regions, while local regions represent areas around face organs (eyes, mouth, forehead and nose). Image flipping symmetry of the face are considered in methodology stages (training and testing), where for experimental

example, left eye local is considered as local regions. Selected local and global regions of this methodology are illustrated in Fig. 2.

Firstly, all branches are shared with low-level features, where the knowledge is transferred between network branches to eliminate over-fitting risk. Secondly, aligned-region features are generated using proposed ARP technique to avoid misalignment cases result large variations of face poses. Moreover, recurrent fusion is employed in each branch for exploring the dependency and spatial correlation between different regions of the face. Finally, proposed technique combines all branches together and optimizes them jointly, which generates reliable features with discriminative power.

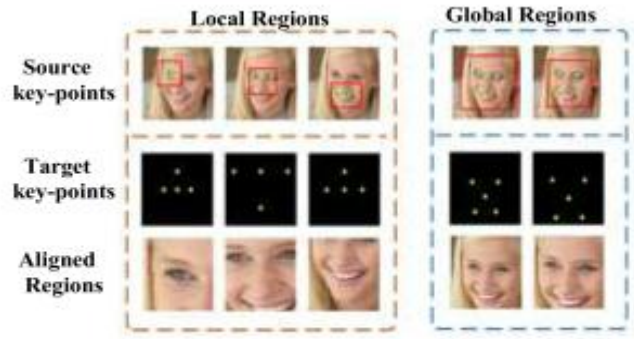

Figure 2: Illustrates aligning face

2.1 Generating Aligned-Region Features

Although some input images are aligned, sub-region parts may have misalignment. To handle such case, ARP is adopted for feature extraction of aligned region. The ARP technique adopts using Affine transformation, depending on some source and target points. In this work, source key points are denoted as $R^s=\{p_1, p_2 \dots p_m\}$, where p_i represents the i^{th} source point. Consequently, target key points are denoted as $R^t=\{q_1, q_2 \dots q_m\}$, where q_i represents the i^{th} target point. Parameter matrix M_A of the Affine-transformation T (.;MA) is constructed based on {R^s , R^t,} via the minimum values of the Least Squares Error (LSE):

$$LSE = \sum_{p,q} \|q_i - T(p_i; M_A)\|^2 \quad (1)$$

Firstly proposed method calculates the parameter matrix (MA) using (source and target) key points (as in Fig. 2) before using the block C^{B1} to produce the input features X^S. which is used to calculate the aligned feature vector X^t as ($X^t = T(X^s; M_A)$). Normalization to the [−1, 1] range is applied to the coordinates (height and width) of feature maps from source and target [18], then (X^t) is adapted to a specific matrix resolution (56×56 in this work) before feeding them into the next blocks. In this paper, ARP module is used for five aligned region features, which include two global ones (one of them is larger than the other with 1.2 times), while the remaining three local regions keep central nose and mouth with left eye. Global features, as example, are aligned based on two mouth corners, one nose tip and two eyes key points. Obviously, the first set of

global features is represented using the entire C^{B1}features since the face is previously aligned.

2.2 Feature Extraction

The input feature vector for local and global branches is employed using either features from aligned region C^{B1}or the raw ones, while the input vector of local-global branch uses a sequence of features from local and global region, in the second block layer, to learn the integrated configuration. Specifically in global-local branch, the block $C_{C_k}^{B_3}$ takes the feature sequence of $C_{g_k}^{B_2}$and $C_{l_k}^{B_2}$. Both of local and global information are jointly considered to help the network in capturing complementary information.

In local or global branches, sub-networks contain three residual blocks, while there are only two branches for sub-networks within the local-global branch. The task of last residual in each sub-network $C_{g_k}^{B_4}$, $C_{l_k}^{B_4}$ or $C_{c_k}^{B_4}$ is generating high-dimensional features for each of the (512) channels. It is considered in the increasing of posterior-layers parameters such as the Fully Connected layer and the Long Short Term Memory LSTM unit leading to the case of over fitting. As a results, convolutional layers of (128) kernels only, a Global Average Pooling (GAP) are employed in addition to the Fully Connected (FC) layer containing 128 neurons only, whereas last residual block of the sub-network is following. Then, a vector of 128 features is extracted from each sub-network. For more clearance $z_i^{g_k}, z_i^{l_k}$ and $z_i^{c_k}$ refer to feature vectors extracted from I_i in the k^{th} sub-network of the local, global and local-global branches respectively.

2.3 Recurrent Fusion

A sum operation and LSTM network are adopted in each branch to detect the correlation between features from different region. Taking the global branch as an example, the encoded hidden state can be represented as:

$$\boldsymbol{h_i^{g_k} = LSTM\left(z_i^{g_k}, h_i^{g_{k-1}}\right)} \quad \textbf{(2)}$$

Where: LSTM is the unit encoding hidden state ($h_i^{g_k}$) considering ($z_i^{g_k}$) features with prior hidden ($h_i^{g_{k-1}}$) state as an input. Then the final form of global features (h_i^g) is computed by summing all hidden features as:

$$h_i^g = \sum_k h_i^{g_k} \quad (3)$$

Correlated regions and spatial dependency are explored using LSTM Recurrent fusion. Discriminative features for age estimation scheme can be extracted by learning topological contextual information of different regions. Similarly, the final form of local and local-global features (h_i^l, h_i^c) respectively can be also obtained. Fusing different features (h_i^l, h_i^g and h_i^c)

provides different information about studied face regarding age estimation, e.g., (h_i^l) mainly refers to the local detailed attributes, whereas (h_i^g) refers to the global face-structure. The local-global form (h_i^g) provides integrated configuration for both of local and global information. Consequently, they are concatenated in fused form to provide robust and efficient features xi = con (h_i^l, h_i^g, h_i^c). Due to best of our knowledge, this form of feature fusion is not proposed in any other work regarding facial age estimation literatures.

2.4 The Loss Function

Many previous techniques [2, 3 and 4] trained their networks by applying the loss function at the final level only. However, it is difficult to provide independent learning process for extracted features from different regions. Therefore, this work employs multiple-objective loss functions for the training of the proposed HRA. An individual objective loss function is enforced in each sub-network applying same label constraint, where each sub-network learns extracted discriminative-features independently. In addition, the final layer follows the cross entropy classification loss for guiding the entire network for extracting efficient and robust features for final age estimation. The Posterior Probability $p_j(x_i)$ is adopted as the probability measure to assign each extracted features (x_i) to the related j^{th} class due to:

$$\boldsymbol{p_j(x_i) = \frac{exp(W_j^T x_i)}{\sum_v exp(W_v^T x_i)}} \quad (4)$$

Where w_j represents j-class parameters regarding softmax classifier. Similarly, posterior probabilities $p_j(z_i^{lk})$, $p_j(z_i^{gk})$ and $p_j(z_i^{ck})$ are defined for extracted features z_i^{lk}, z_i^{gk} and z_i^{ck} of the j^{th} class, respectively. As a result, the entire training process of losses is represented as:

$$\boldsymbol{L = -\frac{\sum_{i=1}^{n} log\left(p_{yi}(x_i)\right)}{n} - \frac{\sum_{i=1}^{n}\sum_{\vartheta\in\theta}\sum_{k=1}^{3} log\left(p_{yi}(x_i^{k\vartheta})\right)}{n}} \quad (5)$$

Where *n* represents the size of the training dataset, *yi* represents age label i^{th} image, and ϕ is the set of local (*l*),gloval (*g*) or *local-global* (*c*) branches. Finally, this work employs a refinement of the expectation before the final layer for predicting the specific age. It is referred to as: $y'_i = \Sigma_j j \times p_j(x_i)$.

Above Eq. 5 can be re-written as $L = -A - B$ where (A) refers to the whole network, while (B) refers to the sub network. Note that, employing the classifiers before the sub-network assists the network training only, while in the test stage, only the last classifier is used to predict the final age.

3. Web-FaceAge Dataset

The experiments of this work are conducted on privately collected face-image dataset. Dataset images are collected from different resources such as daily image project on YouTube and

downloaded images from search engines, such as Google, Yahoo, Opera and Bing. When a specific URL contains multiple images, single image is downloaded only. The studied face is detected from all images using standard face detection method depending on faster r-cnn [19]. Age labels are collected based on different tools published age labels, some standard face application such as Face++ API [20], AuthenMetric [21] and opinions of more than 20 academic students and staff. In the case of different estimated labels, the average of them is adopted. Image collection consumed more than two months to be accomplished.

To avoid image duplication, the standard pHash technique is adopted by the author to represent each image using 64-bits binary vector. Hamming distance is used to measure images similarity, and two images are considered duplicated if the result is smaller than threshold (6), where one of them is removed randomly. Collected database, in this work, contains about 125k face images with their age labels. Large varieties are ensured between collected image samples regarding ethnicities, head poses, camera angles and illumination. Age labels vary over wide range (0-93), as in Fig. 3, where the biggest portion of them is for children and young ages. As same as most of age estimation works, dataset image are divided into 20% of the images are kept for test and the remaining 80% are used for training. Fig. 4 illustrates age distributions adopted in both of training and test datasets.

Figure 3: Samples of dataset images and their age distribution

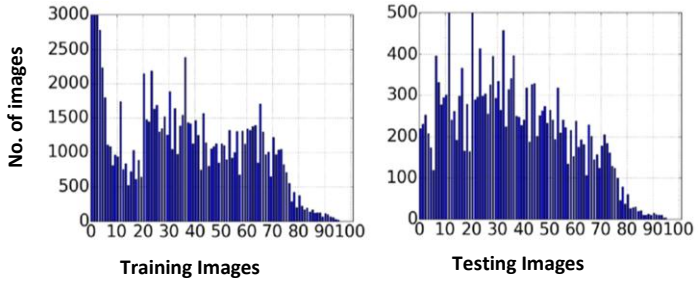

Figure 4: Samples of dataset images and their age distribution

4. Experiments

4.1. Datasets

To evaluate proposed estimation scheme, it is benchmarked with five datasets, IMDB-WIKI [2], CACD [15], MORPH II [22], FG-NET3 and Chalearn LAP 2015 [23]. The collected Web-FaceAge and IMDB-WIKI datasets are previously been discussed in Section 3. While other datasets are:

MORPH II. It contains more than 55k face images. In this work, three types of evaluation are employed regarding training and testing datasets, where in the first type, 80% of dataset images are used for training against 20% for testing as same as done by [3]. In the second type, same division applied in the first type is applied on sub-sets (races) to avoid race-interference following [2 and 4], where (5493) images chosen from Caucasian faces, then, they are randomly divided into two subsets (80% training and 20% testing). In the third type, MORPH II images are divided into non-overlapped S1, S2 and S3 inspired by [4] techniques. Each sub-set of them is used for testing against the training using the sum of the other two sub sets.

FG-NET, it contains only 1,002 images captured for different 82 subjects, and the standard Leave-One Person-Out (LOPO) technique [2] is used for evaluation. f

CACD, dataset images are collected for about 2k celebrities with about 160k images of. Following Tan et al. [4], dataset is divided such that images for 1800 celebrities are used for training, 80 for validation and the remaining 120 for testing.

Chalearn LAP 2015, it is a considerable dataset in apparent applications of age estimation. It contains more than 4.5 images apparent age-label and the deviation. This dataset is divided into three subsets, 2476 training images, 1079 test images and 1136 validation images. For evaluation of this work, experimental settings are applied following [24].

IMDB-WIKI, it is common dataset pre-training techniques because of its noises. Since this work is an age estimation application, non-face images and some of multiple-face images are ignored. The remaining 30k images are adopted for pre-training process.

4.2. Experiments and Settings

In pre-processing stage, image alignment is performed based on two landmarks, eye-center and upper lip, and to extract features with large vector size regarding ARP, all images are re-sized to 256 × 256 size as an input. As it is previously proved [5], studied images are augmented using scaling, random horizontal flipping, rotation to be translated to training stage. In testing stage, the image is combined with the flipped copy to be fed into the network, and then final prediction results from average prediction. On the other side, this work employs random horizontal flipping to evaluate Web-FaceAge and CACD datasets due to the big size of training datasets of them. The training of all networks is pre-trained using the ImageNet to be optimized with Nesterov momentum used by SGD. The initialization of the network is set to 0.0005 for weight decay, to 0.9 for momentum and to 0.01 for learning rate, which is reduced, in training iterations, by a factor of 10. For results evaluation of this work, ϵ-error and Mean Absolute Error (MAE) are adopted.

4.3. Ablation Studies

In this part, both of proposed Web-FaceAge and the standard MORPH II datasets are selected for effectiveness validation on learning the local and global features, recurrent feature fusion, ARP and the losses on sub-networks.

Learning the Local and Global Features: Analysis and learning results of local, global and local-global features are illustrated in Table 1. First row explains baseline methods using the plain method (Res18+EX) of ResNet-18 proposed by [2] and ignoring Local, Global or Local-Global. Results enhancement is obvious using Local, Global or Local-Global, where they outperform yielded results using baseline method. This demonstrates the efficiency of all three proposed branches, where multiple sub-networks are adopted with repeated fusions. In addition, lowest error rate (MAE) among the three branches is yielded using the local-global branch, where it is shown that the integration of learning between both local and global extracts features with higher efficiency than using either local or global learning alone. Joining the learning of local and global branches achieves lower error rate than using local-global branch. Yet, features-learning by employing all of Local, Global and Local-Global provides better performance due to the lower error rate values. It illustrates the efficiency of jointly learning the local details, global context, besides the integration of their features.

Table 1: Analyzing & learning local (L), global (G), and local-global (L-G) features on standard MORPH II and collected Web-FaceAge datasets.

Method	MORPH II			Web-Face Age
	20%- 80%	Sub 20%- 80%	S1, S2 & S3	
Res18+EX [24]	2.34	2.72	2.70	3.74
L	2.19	2.63	2.68	3.75
G	2.09	2.61	2.62	3.64
L-G	2.03	2.56	2.55	3.59
L & G	1.99	2.55	2.53	3.59
L, G and L-G	1.91	2.49	2.45	3.55

4.4. Aligned Region Pooling (ARP): by the replacement between ARP and ROI pooling, nose-tip landmark is adopted to be the center to generate a global region. Yet for generating the local regions, central points of left eye, mouth, and the middle point between two-eye center and the nose tip. As in Fig. 5, it's clear that replacing ROI pooling by ARP leads the performance to drop on both of the two used face datasets, which indicated that ARP yield higher performance than using ROI for extracting features with higher efficiency.

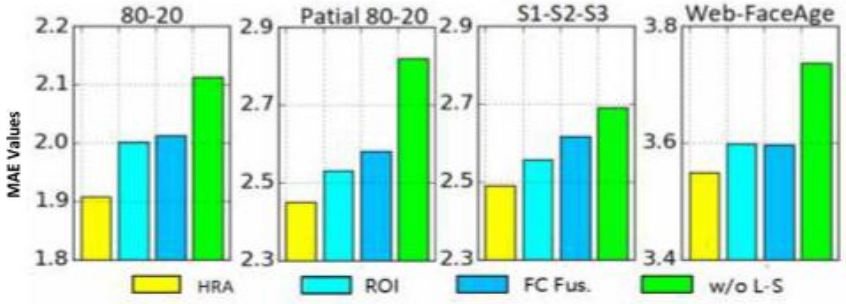

Figure 5: Yielded results of MAE using MORPH II dataset. (ROI) color referes to replace ARP by ROI pooling. While (FC Fus.) means using FC Fusion in HRA, and w/o L-S refers to train the network regardless losses functions

4.5. Benchmarking with State of Art:

Applied on the collected Web-FaceAge dataset, results yielded by HRA are benchmarked between the firstly applied Res18+EX [24] and the adaptive form AgeED [16] based on essential values n = 9 and ρ = 2. Due to the illustrated results in Table 2, proposed HRA estimation scheme yielded better performance with the lowest MAE value (3.55). Comparing with baseline application (Res18+EX), HRA provides promising improvement, where it reduces MAE value by 0.19.

Table 2: Benchmarking with AgeED and Res18+EX based on MAE

AgeED	Res18+EX	**HRA**
3.93	3.74	**3.55**

Using MORPH II.: Illustrated in Table 3, explain that proposed HRA provide considerable enhancements on previously yielded results from state-of-the-art techniques using all 80%-20%, partial 80%-20% and S1-S2-S2 protocols, where the achieved MAE values where of 1.91, 2.49, 2.45 respectively. Note that previous works [2, 16 and 27] and the previous work [5] are pre-trained using IMDB-WIKI and MS-Celeb-1M respectively, while the proposed estimation scheme was not pre-trained using any additional dataset. Despite that, proposed HRA is still outperforming other methods even with additional pre-training. Due to the best of our knowledge, this work conducts the first technique yielding less than 2 years for MAE value using 80%-20% protocol even though fine-tuning is not adopted for face datasets.

Using FG-NET & **CACD**: Table 4 provide another comparison for proposed method against state of art techniques applied on both of **FG-NET** and **CACD** images. In **FG-NET** also, proposed estimation scheme provides an encouraging enhancements on the performance of the compared techniques. Yielded MAE value was reduced to 3.72, where no additional datasets are used for pre-training. But due to the small size of this dataset, pre-training the network with additional large datasets provides additional decrement on MAE value down to (2.59). These results illustrate that proposed estimation scheme provide low levels of error rate even using small scale datasets. On the other hand, using noisy images decreases the accuracy and increases the error rate, yet, applying the proposed estimation scheme on CACD face dataset reduces MAE value to be 4.35.

Table 3: Benchmarking state of art based on MAE

Method	**20%- 80%**	**Sub 20%- 80%**	**S1, S2 & S3**
OR-CNN [3]	3.34	-	-
CasCNN [25]	3.30	-	2.93
DEX [2]	-	3.25	-
ARN [26]	3.01	-	-
AgeED [16]	-	2.93	2.86
SSR-Net [27]	-	2.52	-
M-V Loss [7]	2.41	-	-
DRFs [28]	2.17	2.91	-

ThinAgeNet [5]	1.97	-	-
Res18+EX [24]	2.33	2.72	2.69
HRA (Proposed)	**1.91**	**2.49**	**2.45**

Table 4: Using FG-NET and CACD for benchmarking state of art based on MAE

Method	FG-NET	CACD
SOFTMAX [29]	-	5.19
DEX [2]	4.63	4.79
CasCNN [25]	-	5.22
AgeED [١6]	4.34	4.68
M-V Loss [7]	4.10	-
ARN [26]	-	-
DRFs [28]	3.85	4.64
Res18+EX [24]	4.26	4.51
HRA (Proposed)	**3.72**	**4.35**

Using Chalearn LAP 2015: In this case ϵ- error and MAE error rate of the results are reported regarding the validation set against ϵ-error of the results only regarding the test set. Shown in Table 5, conducted comparison depends on estimation techniques used IMDB-WIKI dataset for pre-training. The proposed estimation scheme outperforms other techniques regarding both of test and validation sets. Proposed HRA scheme provides, in validation set, 3.05 for MAE valued and 0.27 for ϵ-error, and in the test set, HRA achieves 0.252 for ϵ-error, which the lowest value among the others.

Table 5: Benchmarking state of art based on MAE

Method	Validation		Test
	20%- 80%	Sub 20%- 80%	
Res18+EX[24]	3.29	0.29	0.27
ThinAgeNet [5]	3.14	0.28	-
ARN[26]	3.15	-	-
AgeED [١6]	3.21	0.28	0.26
TinyAgeNet [5]	3.43	0.31	-
CVL ETHZ[2] (**rank 1**)	3.52	0.28	0.27
ICT-VIPL [31] (**rank 2**)	3.34	0.29	0.27
WVU CVL [32](**rank 3**)	-	0.31	0.30
HRA	**3.05**	**0.27**	**0.25**

4.6. Discussions

Different types of classifiers were applied in this work to explain the effects of local, global and local-global features on the results. Yielded MAE values by all classifiers are stated and illustrated in Fig. 6. HRA denotes proposed classifier, whereas L-*i*, G-*i*, LG-*i* denote the i^{th} classifier regarding local, global, and local-global branches, respectively. It's obvious that highest

MAE values were yielded using local features since such features have local effects regardless the others, which gives global features higher performance relatively. They handle wider effects since the represent wider effects of age progression [33]. The local-global features yielded performance since they consider both of local details and wider age progression signs. Combining all of these types in the proposed HRA yielded the best possible results for MAE where considering all of these types provides features with higher efficiency and discrimination.

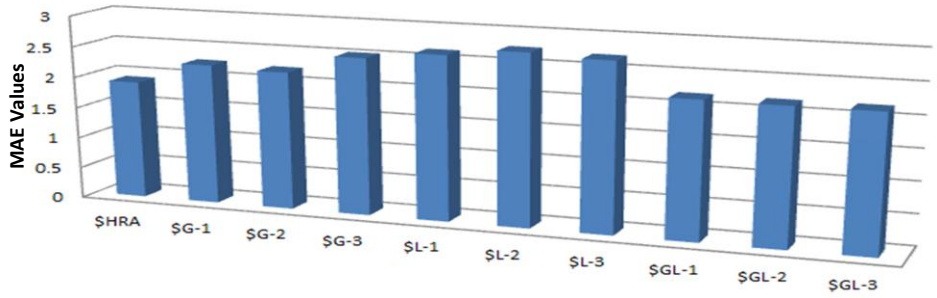

Figure 6: Yielded results of MAE for different classifiers using MORPH II dataset and 80- 20 protocol

Extracted features are visualized for proposed HRA and the sub-components local, global and local-global regarding t-SNE [30], and the behavior of these features is shown in Fig. 7. Features distribution shows that, in similar ages, generated compact features by the proposed HRA (*e.g.*, points with black (or brown) color in the red (or blue) box, respectively) are more than the generated by the nearest work to the proposed technique [4]. Due to the proposed HRA, features that extracted from same age preserve being closer as they ate more robust in age estimation applications.

5. Conclusion

This work provides an individual work that jointly learns the proposed network for the local details, global context and integrated local-global features, called Hybrid-Representation Architecture (HRA), for estimating a person's age from his face image. Moreover, proposed HRA provides an integration of novel components, like ARP, recurrent fusion, and it employs separate losses on each sub-network for capturing robust features. It is shown due to the extensive experiments that significant enhancements are provided by the proposed estimation scheme regarding the state-of-the-art published works.

References

[1] Guodong Guo and Guowang Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In FG, 2013.

[2] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. IJCV, 2018.

[3] Zhenxing Niu, Mo Zhou, Le Wang, Xin- bo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In CVPR, 2016.

[4] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Li Stan Z. Attention based pedestrian attribute analysis. IEEE TIP, 2019.

[5] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. IJCAI, 2018.

[6] Kai Li, Junliang Xing, Chi Su, Weiming Hu, Yundong Zhang, and Stephen Maybank. Deep cost- sensitive and order-preserving feature learning for cross- population age estimation. In CVPR, 2018.

[7] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In CVPR, 2018.

[8] SALIH, Ghalib A. Salman; SULONG, Gazali. Topographical Features for Senior Adult Age Estimation. Journal of Soft Computing and Decision Support Systems, 2015, 2.6: 20-26.

[9] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi- loss classification. IJCAI, 2017.

[10] Sepp Hochreiter and Ju¨rgen Schmidhuber. Long short-term memory. Neural computation, 1997.

[11] Dong Yi, Zhen Lei, and Stan Z Li. Age es- timation by multi-scale convolutional network. In ACCV, 2014

[12] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi- labeling learning and optimal temporal fusing. In CVPR, 2017.

[13] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. Grouping attribute recog- nition for pedestrian with joint recurrent learning. In IJ- CAI, 2018.

[14] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In ACM MM, 2017.

[15] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age- invariant face recognition and retrieval. In ECCV, 2014.

[16] Zichang Tan, Jun Wan, Zhen Lei, Ruicon- g Zhi, Guodong Guo, and Stan Z. Li. Efficient group-n encoding and decoding for facial age estimation. IEEE TPAMI, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog- nition. In CVPR, 2016.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In NIPS, 2015.

[19] Shaoqing Ren, Kaiming He, Ross Gir- shick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[20] https://www.faceplusplus.com /

[21] http://www.authenmetric.com /

[22] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In FG, 2006.
[23] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baro´, Jordi Gonzalez, Hugo J. Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cul- tural event recognition datasets and results. In ICCVW, 2015.
[24] Tan, Z., Yang, Y., Wan, J., Guo, G., & Li, S. Z. Deeply-learned Hybrid Representations for Facial Age Estimation. In IJCAI (pp. 3548-3554), 2019.
[25] Jun Wan, Zichang Tan, Zhen Lei, Guodong Guo, and Stan Z. Li. Auxiliary demographic in- formation assisted age estimation with cascaded structure. IEEE TCYB, 2018.
[26] Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Anchored regression networks applied to age estimation and super resolution. In ICCV, 2017.
[27] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stagewise regression network for age estima- tion. In IJCAI, 2018.
[28] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L. Yuille. Deep regression forests for age estimation. In CVPR, 2018.
[29] Zichang Tan, Shuai Zhou, Jun Wan, Zhen Lei, and Stan Z Li. Age estimation based on a single network with soft softmax of aging modeling. In ACCV, 2016.
[30] Laurens van der Maaten and Ge- offrey Hinton. Visualizing data using t-sne. JMLR, 2008.
[31] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. Agenet: Deeply learned regres- sor and classifier for robust apparent age estimation. In ICCVW, 2015.
[32] Yu Zhu, Yan Li, Guowang Mu, and Guodong Guo. A study on apparent age estimation. In ICCVW, 2015.
[33] Salih, Ghalib A. Salman; SULONG, Ghazali. Geometrical Age Effects on Child Face for Age Estimation. 2015.