A Practical Study on Quality Evaluation for Age Recognition Systems

Chuanqi Tao, ①② Hao Chen, ② Tiexin Wang ①, Jerry Gao ③, Wanzhi Wen ④

① College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

② School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

③ Department of Computer Engineering , San Jose State University, USA

④ School of Computer Science and Technology, Nantong University

Correspondence to: taochuanqi@nuaa.edu.cn

Abstract—Face recognition system is a widely-used intelligent application nowadays. Existing recognition system evaluation methods primarily focus on recognition rate, i.e., the correct result. However, current research seldom focuses on the quality evaluation of face recognition systems. They seldom consider accuracy or the quality of recognition. To address this issue, this paper proposes several quality factors for evaluation. In addition, corresponding metrics for diverse quality factors are illustrated. Moreover, the paper presents an experimental study on a realistic non-trial face age recognition system using the proposed quality evaluation method. The study result shows the proposed method is feasible and effectiveness in quality evaluation.

Keywords: ***face recognition, age recognition application, quality factors, quality evaluation***

I. INTRODUCTION

Face recognition is one of the commonly-used intelligent systems in people's daily life. During the past ten years, face recognition technology has been widely focused and studied. Thus, face recognition systems are fast developing, from the feature points extraction or face matching in early years to face authentication, face age recognition and prediction and face dynamic tracing in recent times. Face recognition usually consists of the following steps: face image collecting, image preprocessing, image feature extraction, image processing (such as matching, recognition, and etc.). Facial age recognition has been concentrated and becomes a hot research issue recently in face recognition, which mainly realizes the judgment and prediction of age of a face picture. Some well-known age recognition systems have been used widely such as Microsoft's latest application--HOW-OLD and Alibaba cloud face age recognition API. When a user uploads photos, HOW-OLD will be able to recognize the gender and age of the given image. It works through learning and training of massive data in cloud, using detecting and identifying to tell the age of the image. HOW-OLD is mainly done by face detection, gender classification and age detection. Face detection is the basis of the other two technologies, and age detection and gender detection conduct the classification problem in the process of machine learning. This relates to the facial features of the portrait, the collection of learning data, the establishment of a classification model and model optimization. HOW-OLD analyzes the 27 points on people's faces to draw conclusions. These points are key nodes of the face, such as the pupil and the corner of the eye. Many researchers proposed optimize algorithm of facial age recognition. However, there are few quality evaluations for facial age recognition. The commonly-used metrics is recognition rate, which only tells the number of successful recognition. Age recognition system is a complex application that allows and tolerates errors (error is deviation from actual and expected value in software engineering). For example, assuming the real age of a face picture is 20 years old, if the recognition result is between 18 and 22, we may say that the system has a good recognition, rather than it must be recognized as 20-years-old exactly. Thus, how to measure the quality of face age recognition is an important issue for those kinds of systems.

Based on our recent research and survey, this paper proposes a set of reference quality factors for quality evaluation of face age recognition system. In addition, an experimental study is performed to discover the quality defects and problems of face age recognition. The remaining part of the paper is structured as follows. Section II is the related work. Section III presents the proposed quality factors and their measurements for facial age recognition systems. Section IV shows our experimental study and comparison with existing approach. Conclusions and future work are summarized in the end.

II. RELATED WORK

There are increasing quality problems resulting in erroneous testing costs in enterprises and businesses. According to IDC [1], the Big Data technology market will "grow at a 27% *compound annual growth rate* (CAGR) to $32.4 billion through 2017". In our previous work, the issue of quality assurance and validation for big data and applications was preliminary discussed [2, 3].

In the field of facial age recognition, most of the researchers focus on recognition algorithms. Du proposed a facial age estimation method based on sparsity constrained non-negative matrix factorization [4]. Yu proposed an age recognition method based on fusion error correcting output coding [5], which is a kind of SVM multi class classifier and integrates the error correcting output coding. Zhu proposed a 3D facial age recognition algorithm based on multi classifier fusion [6]. There are more proposed different age estimation

DOI reference number: 10.18293/SEKE2017-176

methods, such as [7-11].

Facial age recognition becomes a hot research domain in pattern recognition, so there is little evaluation method of facial age recognition. When evaluating the effect of the algorithms, researchers usually use the *average absolute error* to evaluate its quality [4, 7, 8, 9, 10], or the group recognition rate [5, 6, 8, 11]. The *average absolute error* is the average value of the absolute difference between real age and recognition age. The group recognition rate means the ratio of the occurrence when real age and recognition age are in the same age group.

Age recognition system not only need consider recognition rate or pass rate, but also the error between real age and recognition age. In the field of Agricultural, water conservancy, weather and economy, some error factors are used to evaluate the error, such as average absolute error, relative error [12-15]. But these quality factors haven't been used systematically and adaptively to evaluate an age recognition system. In Physics and Mathematics, there are also some error theory and error factors [16][17].

III. QUALITY FACTORS OF AGE RECOGNITION

Based on our investigation and analysis, we find that the *average absolute error* has some limitations in quality evaluation. For instance, when we recognize an image with 10-years-old in real as 20 years old, and recognize an image which is 70-years-old in real as 80 years old, the average absolute error is both 10 years old, but actually, recognize 10 as 20, the error rate is up to 100%, while recognizing 70 as 80, error rate is only 15%. Thus only using the *average absolute error* to evaluate the quality of facial age recognition is not reasonable. The *group recognition rate* used to evaluate age recognition is also not fully available, since the age recognition is a fault-tolerance application. When we define groups, there exists unavoidable boundary. For example, we define two group intervals as (10, 20] and (20, 30]. Both groups are not available in left boundary while available in right boundary. When the real age falls on the boundary just right, such as 20, if the system recognizes it as 19, then the result belongs to interval (10, 20]. However, when the system recognizes it as 21, it belongs to the interval (20, 30]. Thereby evaluation system may think there is an error. As we know, both 19 and 21 is an acceptable recognition result of real age 20.

Thus, only using average absolute error or the recognition rate cannot measure the quality of the age recognition system effectively. It is easy to ignore the hidden quality problems and is not conducive to the improvement of the system. The error between the facial age recognition results and actual real results, determines the accuracy of face recognition. Combing the error theory in theoretical physics and mathematics with the study of features of age recognition system, we propose several quality evaluation factors for age recognition systems shown in figure1.

The quality factors are mainly composed of two parts: recognition rate and accuracy rate. They are illustrated as follows.

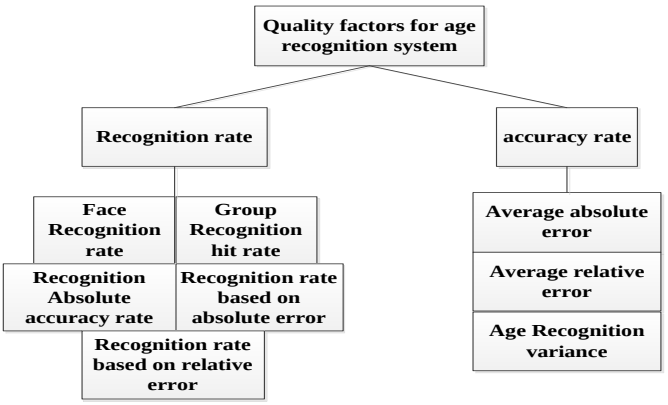

Figure 1 Quality factors for age recognition system

Recognition rate

The recognition rate refers to the ratio of recognition age to the actual age. The recognition rate is a reflection of the capability of age recognition such as the recognition ratio in different age groups with fault tolerance. Overall, the recognition rate reflects the correctness of an age identification system. Here we propose five quality calculating indexes for recognition rate. They are listed as: (1) the face recognition rate; (2) the group recognition hit rate; (3) Recognition absolute accuracy rate; (4) the recognition rate based on absolute error; and (5) the recognition rate based on relative error. The detailed calculation metrics for the five calculating indexes are explained as follows.

(1) Face recognition rate. Recognition rate is affected by a number of external factors, such as the angle of face, the size of face, and multi-faces in one image. There are more reasons that may cause the failure of face recognition in practical use. Here we adopt formula $R1 = NP/N$ to measure face recognition rate, where NP refers to the passed recognition cases (regardless of recognition accuracy), i.e., the system recognize a face and give its corresponding age, and N refers to the total cases that have faces without image quality problems.

(2) Group recognition hit rate. It refers to the ratio of the recognition cases that in the correct age group. For instance, we divide age range into several groups, with interval such as 10-20, 20-25, and etc. To avoid the boundary problem as discussed before, we assume if the real age is just at the boundary, both the left and right age group are considered to be correct for the recognition age. A good age recognition system should be able to make the recognition rate as high as possible. We use formula $R2 = NH/N$ to measure group recognition hit rate, where NH refers to the cases that hit the correct age group, and N refers to the total cases that pass the recognition system.

(3) Recognition absolute accuracy rate. It describes the ratio of the cases that the recognition age equals to the true age. In this metric, we use the formula $R3 = NC/N$, where NC means the cases that the recognition age is equal to the

DOI reference number: 10.18293/SEKE2017-176

actual age and N refers to the total cases that pass the recognition system.

(4) The recognition rate based on absolute error. It refers to that the error is divided with a fixed real number. The *recognition rate based on absolute error* represents the ratio of the cases that the error between real age and recognition age is smaller than the fixed real number as threshold. When the fixed real number threshold is zero, the value is the same as the index (3) above. The given formula is $R4 = COUNT(abs(age1 - age2) < e) / N$, where *age1* means the real age, *age2* is the recognition age, *e* represents the fixed real number threshold, and *N* refers to the total cases that pass the recognition system.

(5) The recognition rate based on relative error. This index shows that the error is divided based on the percentage of actual age. Since *recognition rate based on absolute error* cannot avoid the effect of the age size, we need to define new metrics for recognition rate. For instance, when we recognize an image with 70 years-old in true as 80 years-old, and recognize an image which is 20 years-old in true as 30 years-old, though the absolute error is both 10, it is obvious that recognizing 20 as 30 is more intolerable than recognizing 70 as 80, since the error for those two cases is 50% and 15% respectively. The *recognition rate based on relative error* is the ratio of the cases that the error between real age and recognition age is smaller than the percentage of actual age as threshold. The formula is given as $R5 = COUNT(abs(age1 - age2) / age1 < e) / N$, where *age1* means the real age, *age2* means the recognition age, e means the percentage, and N refers to the total cases that pass the recognition system.

Accuracy

Accuracy is a measurement of recognition error between recognition age and real age. If the 19-years-old is recognized as 20, from the recognition group rate, this age recognition is successful, because the error is small, but this does not mean there is no error. A good age recognition system should ensure a high recognition rate and a high accuracy with low error. The accuracy is divided into three indicators: *average absolute error*, *average relative error*, and *age recognition variance*. The explanations and their metrics are shown as follows.

- Average absolute error. It refers to the average absolute value of the error between real age and recognition age. The lower the value is, the closer the actual age and the age of recognition is, and the better the recognition quality is. The formula is given as $R5 = \sum abs(age1 - age2) / N$, where *age1* means the real age, *age2* means the recognition age, and N refers to the total cases that pass the recognition system.

- Average relative error. As explained in the quality factor #3 in recognition rate before, here we use average relative error to avoid the side-effect of age size. The formula is given as $R6 = \sum \frac{abs(age1 - age2)}{age1} / N$, where *age1* means the real age, *age2* means the recognition age, and *N* refers to the total cases that pass the recognition system.

- Age recognition variance. The absolute error and relative error of recognition cannot reflect the overall performance of the age recognition system. A big error may be evened out due to other small errors. Age recognition variance is more sensitive to the larger error point. The variance of the identification is defined as $R7 = \sum \frac{(age1 - age2) * (age1 - age2)}{N}$, where *age1* means the real age, *age2* means the recognition age, and N refers to the total cases that pass the recognition system.

IV. AN EXPERIMENTAL STUDY

In Section III, we discuss the quality factor and the calculation methods for quality evaluation of the age recognition system. In this section, we verify the effectiveness of the proposed quality factors and calculations by a realistic experimental study.

A. Study Object

The study object selects "Face Recognition Technology - Alibaba Cloud Computing Services Facial Age Recognition API" provided by Alibaba in china as the research object. The *base64* encoding of images is submitted to APIs, and the system returns with recognition result. The experiment data sets are selected from the wiki_crop.tar in the open face dataset IMDB-WIKI. There are total 52444 face data, and 10K images are selected randomly as experimental data sets.

B. Study Result Analysis

The study result of *age—cases* distribution is shown in figure 2, where the X axis represents age and the Y axis is the number of cases. The minimum age is 7 years old while the oldest is 98 years old, with 1 years old as the age interval. The actual age cases are marked in blue while recognition age cases are marked in orange.

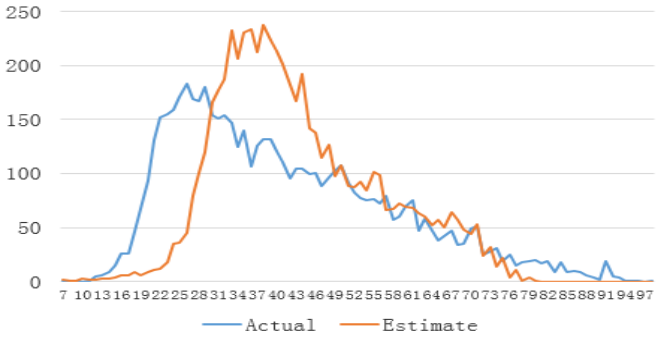

Figure 2 Age-Cases distribution

It can be seen from the figure that the actual age is primarily distributed in the age interval 22-37, and the recognition age is mainly distributed in the 30-40 age group, thus the peak of the age group is offset. Take the group of age

DOI reference number: 10.18293/SEKE2017-176

30-40 as an example, the number of age recognized as 30-40 is higher than the actual number. There are three possible reasons: (1) the system recognizes those under 30 years-old to 30-40; (2) the system recognizes those higher than 40 years-old to 30-40; and (3) it is also possible that the some of those belongs to 30-40 are recognized out of this range, but a number of people whose age is less than 30 or older than 40 are recognized as 30-40 age group, resulting the higher number.

The result of actual age—recognition age distribution is shown in Figure 3. The X-axis denotes the real age and the Y-axis denotes the average recognition age. The minimum age is 7 years old while the oldest was 98 years old with 1 year old as the age interval. The actual age cases are marked in blue while the recognition age cases are marked in orange. As shown in the Figure, the deviation is found as follows: Those between 0-30 years old are often recognized older than their actual age. Those between 70-90 years old in actual are often recognized younger. Age between 30-70 years old in actual are nearly the same as the recognition age. Overall, the error for cases in 0-30 and 70-90 groups is high while the error for cases in 30-70 groups is much lower.

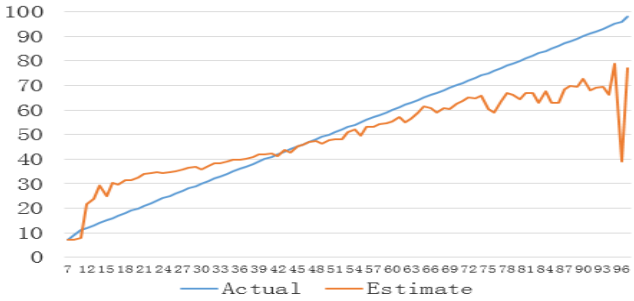

Figure 3 Actual age—Recognition age distribution

Now we analyze the result based on the proposed quality factors. The results of each quality factor that belongs to recognition rate are illustrated as follow.

(1) Face recognition rate. Among these 10000 image cases, 6081 cases are recognized as face images with given recognition age, and 3919 are recognized neither as face images nor be given recognition age. Through artificial identification, we find that the failed 3919 cases are divided into three categories: (a) image quality problems exist, as there is no face in the image, accounted for 33%; (b) though there do have faces in the image, the recognition face is too small, or face is profile, accounting for 54%, among which, profile faces account for 12% and small faces account for 42%; and (c) there exist images with front face are clear enough, accounting for 13%. In addition, among the passed 6081 cases, 5861 cases are recognized as single face while 220 cases are recognized as multi faces. Through our identification, we found that in the 220 multi-face images, there only 16% of the total images indeed have multi faces, and the other 84% only have single face while the system recognized it as multi-face. In summary, the number of cases that have faces without image quality problems is 6590, among which the face recognition rate is 88.92%. The study results show that the age recognition system cannot fully recognize. The possible reasons are as follows.

-The defect of age estimate algorithm;

-The amount of train data set that the age recognition provider use is not enough;

-The train data set that the age recognition provider use does not have universality or representativeness;

-The quality of the provided train data set has problems;

(2) Group recognition hit rate. Here we use the 6590 cases which passed the system. The result is shown in Figure 4.

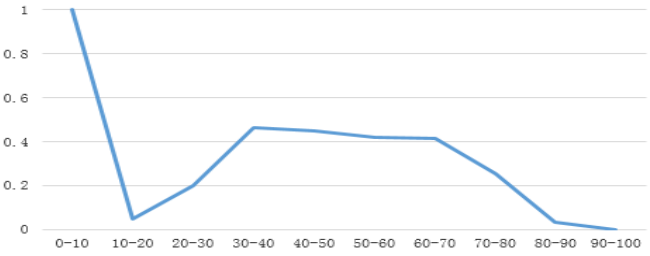

Figure 4 group recognition hit rate

From the figure we discover that the lowest recognition rate falls in the 10-20 age group and 80-100 age group while the higher recognition rate is in 30-70 group.

(3) Recognition absolute accuracy rate. As shown in Figure 5, the group with the lowest recognition accuracy is group 10-30 and group 80-100 while the higher recognition rate group is 30-70.

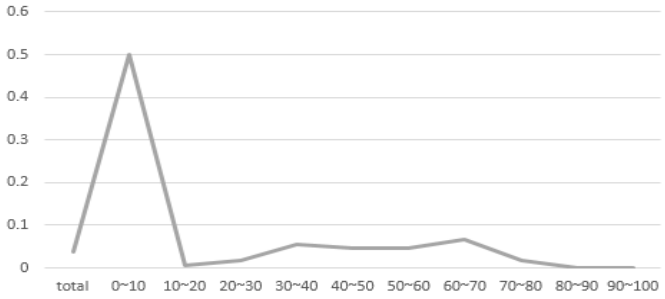

Figure 5 recognition absolute accuracy rate

(4) Recognition rate based on absolute error. The study result is shown in Figure 6.

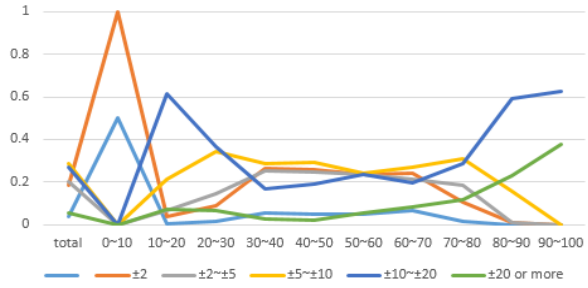

Figure 6 recognition rate based on absolute error map

From the figure we can see that, cases which have absolute error lower than 2 years and 2-5 years both account for 20%.

DOI reference number: 10.18293/SEKE2017-176

Cases with absolute error higher than 20 years account for 5%, and the others both account for 25%. In different age groups, absolute errors show differences. For instance, in the group of age 0-10, the errors are no more than 2 years; in the group of 30-60, the errors are usually between 0-5; and in the age group of 70-100, errors are usually much higher than other age groups.

(5) Recognition rate based on relative error. The result is shown in Figure 7. From the figure we can see that, cases which have relative error lower than 20% account for 45%. Cases which have relative error higher than 40% account for 20%. Others account for 10% or less. In different age groups, relative errors have differences. For example, in the group of 10-20, the errors are usually above 40%; in the group of 30-60, the errors are usually as low as 5%-15%. In the group of 70-80, though the absolute error is large, the relative error is small.

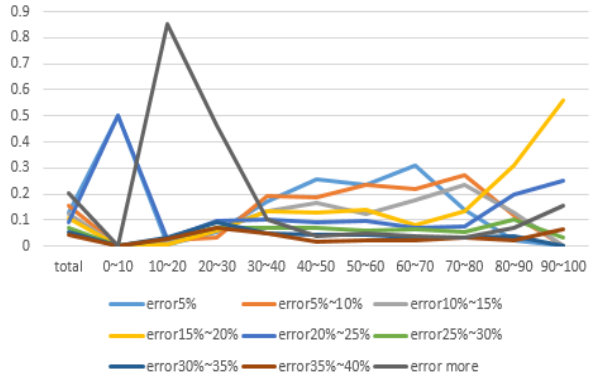

Figure 7 recognition rate based on relative error

The results of each quality factor of *recognition accuracy* are listed as follows.

(1) Average absolute error. The result is shown in Figure 8. We can see that the total average absolute error is about 9 years. Among these, the minimum average absolute error is at the section of 30-40 years, at about 6. The maximum average absolute error is in the group of 90-100. Therefore, the system works well on middle-aged person and works bad in the young and aged faces.

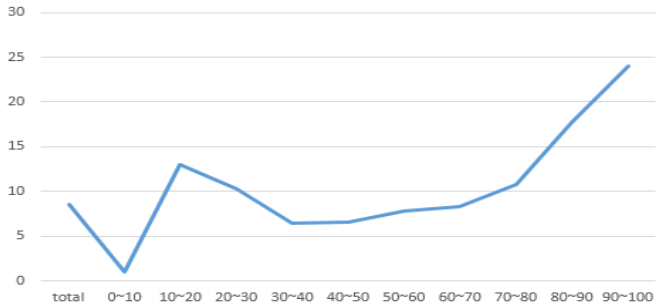

Figure 8. average absolute error map

(2) Average relative error. As shown in Figure 9, the total average relative error is about 25%. Among these, the average relative error in age group 40-60 is around 15% and the average relative error in group 10-20 is up to 75%.

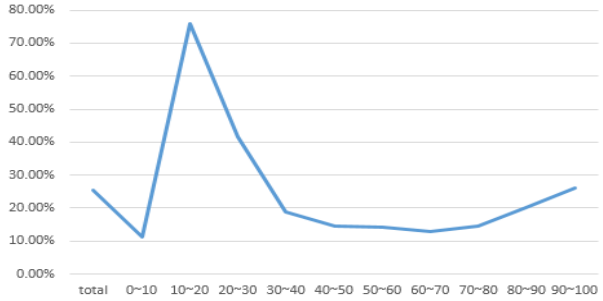

Figure 9 average relative error map

(3) Age recognition variance. As shown in Figure 10, variance can catch big errors instead of evening out them. We can see that the variance in group 10-20 and 70-100 are quite large, while the variance in the middle age is smaller.

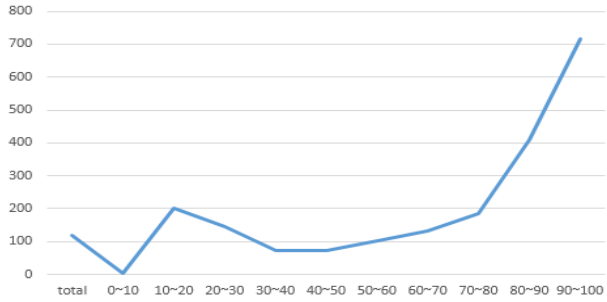

Figure 10 age recognition variance

C. Study Conclusions

Through the experimental study, we discover several defects of the chosen face age recognition system. They are listed as follows

(1) In some cases, the system cannot recognize some of the face images, let alone identify their age.

(2) The system recognizes some single people images as multi people images in some cases.

(3) If the face is too small, or face is profile, the recognition may fail sometimes.

(4) Regarding the recognition rate of absolute error, the ratio of those have errors less than 5 years (excluding completely correct identification) account for about 50%, thus the system has a good recognition ability.

(5) For recognition rate of relative error, the ratio of those have error larger than 40% account for too much and these usually occurs in the age group of 10-20 and 80-100. The system works poor in the young and the aged groups.

(6) In terms of absolute error, 30-50 years have the best performance, while the average absolute error of the aged is higher than 20%.

(7) According to the relative error, 30-50 year group shows the best performance, while the worst performance exists in the 10-20 age group.

D. Comparison with Other Evaluation Methods

DOI reference number: 10.18293/SEKE2017-176

Compared with *average absolute error*, such as the proposed methods in [4, 7, 8, 9, 10], to some extent, *average absolute error* can reflect whether the age recognition system shows good or bad. However, in the 10-20 age group of our experiment, the *absolute error* is not the highest. Nevertheless, in some cases, the *average absolute error* is more than 10 years. Compared with the actual age, the error is high up to more than 50% or even 100%. Therefore, the *average absolute error* is not enough to reflect the problem in this case.

Compared with the *average relative error* work- It can reflect the age recognition system is good or bad in some cases. In our study, the *average relative error* is not the highest in the 80-100 age group, and the *average relative error* of 10-20 and 20-30 groups are higher. However, there exist cases that the system recognizes a 90-year-old people to 70 or even younger. Only using this metric might lead to ignoring of the problems that the system works badly on the aged group.

Only using the recognition group hit rate, such as the methods in [5, 6, 8, 11], the *recognition group hit rate* to a certain extent reflects the performance of the age recognition system, However, it can only reflect the hit number, not the recognition quality and recognition accuracy. In addition, only using age group hit may ignore the error problems of those in the middle of an age group.

E. Study Limitations

There still existing some limitations of our study. They are listed below.

- The magnitude of experiment data is 10K. It can be expanded to millions or more in future work;
- The network or the net speed may affect the response time of the system;
- More quality factors and quality indicators can be proposed to evaluate the error between real age and estimate age;
- We do not take the quality of the picture into account. The noise or bad quality of the picture may affect the experiment result.

V. Conclusions and Future Work

This paper presents a practical study on a realistic recognition system based on the proposed quality evaluation method. The study results show that compared with the single metric such as *recognition rate* or *absolute error*, the proposed method performs better in finding quality issues existed in the age recognition system. In the future work, the quality factors in this paper can be extended to evaluate other intelligent systems such as prediction applications and recommendation applications.

References

[1] M. R. Wigan, R. Clarke. Big Data's Big Unintended Consequences[J]. Computer, 2013, 46(6):46-53.

[2] J. Gao, C. L. Xie, C.Q. Tao. Big Data Validation and Quality Assurance - Issuses, Challenges, and Needs[C], IEEE International Symposium on Service-Oriented System Engineering, 2016, pp 433-441.

[3] C.Q. Tao, J. Gao. Quality Assurance for Big Data Applications– Issues, Challenges, and Needs[C], the 28th International Conference on Software Engineering and Knowledge Engineering. 2016, pp 375-381.

[4] J.X. Du, Q Yu, Q.M. Qu. Facial age estimation method based on sparse constraint non negative matrix factorization[J]. Journal of Shandong University (SCIENCE EDITION), 2010, 45 (7): 65-69.

[5] M.S. Yu, A.Q. Zhu, X.M. Xie. Facial Age Recognition [J]. Co mputer and Modernization (in Chinese), 2013 (11): 210-213.

[6] Y. Zhu, J.L. Chang. 3 D Facial Age Recognition [J]. Computer Simulation, 2009, 26 (5): 248-250.

[7] X. Geng, Z.H. Zhou, K. Smithmiles. Automatic Age Estimation Based on Facial Aging Patterns [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2007, 30(2):368-368.

[8] Y. Dong, Y. Liu, S. Lian. Automatic Age Estimation Based on Deep Learning Algorithm [J]. Neurocomputing, 2015, 187:4-10.

[9] G. Guo,Y. Fu, C.R. Dyer, et al. Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression[J]. IEEE Transactions on Image Processing, 2008, 17(7):1178.

[10] M. Ozaki, W. Motokawa. Dental Age Estimation by Two Computer Methods: Fuzzy Logic and Neural Network [J]. Biomedical Fuzzy & Human Sciences, 2000, 6:13-17.

[11] G. Guo, G. Mu, Y. Fu, et al. Human Age Estimation using Bio-inspired Features [C], IEEE Conference on Computer Vision and Pattern Recognition , 2009, pp 112-119.

[12] K. Kase, A. Makinouchi, T. Nakagawa. Shape Error Evaluation Method of Free-form Surfaces [J]. Computer-Aided Design, 1999, 31(8):495-505.

[13] T.S. Lee, E.E. Adam. Forecasting Error Evaluation in Material Requirements Planning (MRP) Production-Inventory Systems [J]. Management Science, 1986, 32(9):1186-1205.

[14] C. Lee, E. Choi. Bayes Error Evaluation of the Gaussian ML Classifier [J]. IEEE Transactions on Geoscience & Remote Sensing, 2000, 38(3):1471-1475.

[15] C. Cui, B. Li, F. Huang, et al. Genetic Algorithm-Based Form Error Evaluation [J]. Measurement Science & Technology, 2007, 18(7):1818.

[16] Q. Zhou. The Application of Markedness Theory in Error Evaluation: An Experimental Study [J]. Journal of Changchun University of Science & Technology, 2009.

[17] P. Jiang, J.Y. Zhao, J. Wei. Error Theory and Data Processing [M]. National Defence Industry Press, 2014.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China under Grant No.61402229, 61502233, and 61602267; the Open Fund of the State Key Laboratory for Novel Software Technology (KFKT2015B10), the Postdoctoral Fund of Jiangsu Province under Grant No.1401043B, and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant no. 15KJB52003.

DOI reference number: 10.18293/SEKE2017-176