Preliminary Studies on a Large Face Database MORPH-II

G. Bingham, K. Kempfert, B. Yip, J. Fabish, M. Ferguson, C. Nansalo, K. Park, R. Towner,
T. Kling, Y. Wang, and C. Chen

ABSTRACT. In this paper, we consider the statistical learning on a large face database MORPH-II. First, we present a detailed summary of the inconsistencies in the non-commercial release of the MORPH-II dataset and introduce the steps and strategy taken to clean it. In addition, examples of prior research that made use of the uncleaned data are briefly introduced and the potential implications on their results are discussed. Next, we propose a new subsetting scheme for the longitudinal face aging database MORPH-II. Our subsetting scheme is intended to overcome the unbalanced racial and gender distributions of MORPH-II, while ensuring independence between training and testing sets. Our automatic subsetting scheme can be used for various face analysis tasks, including gender classification, age prediction, and race classification.

1. Introduction

MORPH is one of the largest publicly available longitudinal face databases (Ricanek and Tesafaye, 2006). Since its first release in 2006, it has been cited by over 500 publications, as determined by our Google Scholar search. Multiple versions of MORPH-II have been released, but the 2008 non-commercial release will be used and referred to as MORPH-II in this paper. The MORPH-II dataset is a collection of 55,134 mugshots with longitudinal spans, taken between 2003 and late 2007. For each image, the following metadata is included: subject ID number, picture number, date of birth, date of arrest, race, gender, age, time since last arrest, and image filename. Details are shown in Table 2.2. It includes many images of individuals that were arrested multiple times over this five year span. This gives the data a longitudinal aspect that has made it very useful. There are about 4 images per subjects on average. Because of its size, longitudinal span, large number of subjects, and inclusion of relevant metadata, MORPH-II becomes one of the benchmark dataset in the field of computer vision and pattern recognition. It has been used for a variety of face recognition and demographical analysis. In particular, the MORPH-II dataset is widely utilized in research on gender (Han et al., 2015) and race classification (Guo and Mu, 2010b), as well as age estimation (Antipov et al., 2017; Jana and Basu, 2017; Niu et al., 2016; Liu et al., 2015; Guo and Mu, 2010a) and age synthesis (Fu et al., 2010). Notice that there are large pose, lighting and expression variations alone with occlusion in this database.

However, inconsistencies in the records of age, gender and race are uncovered in our exploratory data analytics on the MORPH-II database. According to the local police department, most of the data gathered for mugshots is self-reported, and technology has helped tremendously in being able to verify the information reported. To our best knowledge, no previous work with MORPH-II dataset has acknowledged these inconsistencies, which can be critical in the demographical analysis such as gender and race classification, as well as age estimation and age synthesis. We

Received by the editors May 10, 2018.

2010 *Mathematics Subject Classification.* 62P30; 68U20.

Key words and phrases. MORPH-II, Inconsistency, Cleaning Methodology, Subsetting Scheme.

believe this preliminary study can provide a general guide to the data validation and data cleaning process for future studies on MORPH-II. Accordingly, the first goal of this paper is to provide a thorough explanation of the inconsistencies in MORPH-II and to explicitly detail our cleaning methodology.

In general, test errors can be estimated by the cross-validation technique in order compare the performances between different models. In order to prevent the information leaking, that is, the algorithm learns the identity of a subject from the training set rather than the proposed algorithm itself, one need to make sure that all images of each individual subject must be placed into either the training set or the testing set, but not both. Moreover, the folds should be selected in such a way that distributions of age, gender and ethnicity in each fold should be as similar to the distribution of the full dataset as possible. Therefore in general, researchers manually split the full dataset into non-overlapping folds, based on the predetermined proportion of age, gender, and race. However, a majority of individuals in MORPH-II were arrested multiple times over this five year span, each of those subjects contains more than one image in the database. Additionally, the MORPH-II is highly imbalanced towards black male subjects. With such challenges, Guo et al. (Guo and Mu, 2010a,b, 2011) proposed an evaluation protocol which has been adopted by many studies on MORPH-II. However, it can be tedious and time consuming to manually split the full data into folders with all the criteria mentioned above. Therefore, the second goal of this paper is to investigate how to automatically split the MORPH-II into non-overlapping folds according to evaluation protocol proposed by Guo et al.

The organization of this paper is laid out as follows: Section 2 presents our preliminary study on the inconsistencies and cleaning in MORPH-II. Our proposed automatic subsetting scheme is considered in Section 3. Conclusions are drawn in final section of this paper.

2. Inconsistencies and Cleaning in MORPH-II

We present a detailed summary of the inconsistencies in the non-commercial release of the MORPH-II dataset and introduce the steps and strategy taken to clean it.

2.1. The Original Data

In order to clearly define the inconsistencies in MORPH-II, we first present a summary of the database, drawn from the whitepaper attached to the 2008 non-commercial release (see (MORPH-II) for more details).

TABLE 2.1. Number of Images by Gender and Ancestry (n=55,134)

	Black	**W**hite	**A**sian	**H**ispanic	**O**ther	**Total**
Male	36,832	7,961	141	1,667	44	46,645
Female	5,757	2,598	13	102	19	8,489
Total	42,589	10,559	154	1,769	63	**55,134**

The gender and race distribution for MORPH-II is listed in Table 2.1, with metadata details provided in Table 2.2.

TABLE 2.2. Metadata of MORPH-II

Variable	Information
id_num	6-digit subject identifier (with leading zeros)
picture_num	subject photo number
dob	date of birth (mm/dd/yyyy)
doa	date of arrest (mm/dd/yyyy)
race	(B, W, A, H, O)
gender	(M or F)
facial_hair	not recorded (NULL)
age	integer age ($\lfloor doa - dob \rfloor$)
age_diff	time since last arrest (days)
glasses	not recorded (NULL)
photo	image filename

2.2. Inconsistencies

As the first step of data validation and data management in our preliminary study, we would like to discover the number of unique individuals, the number of unique female, and also the number of unique male in the MORPH-II dataset. By using the 6-digit subject identifier *id_num*, it was found that the true number of unique individuals in the MORPH-II dataset is 13,617. On the other hand, it was found that there were 11, 459 unique males and 2,159 unique females. It implied that the total number of distinct subjects by gender is now 13,618, suggesting that there may be an individual listed as both male and female.

With the discovery of inconsistency in gender, we suspect that there may be inconsistency in the race and age. Repeating the same procedure for race produces, similar results are shown in Table 2.3. Clearly, the total number of distinct individuals 13,658 in Table 2.3 does not agree with the true number of unique individuals of 13,617, illustrating the extent of the inconsistencies in the dataset.

TABLE 2.3. Number of Distinct Individuals by Race and Gender

	Black	**W**hite	**A**sian	**H**ispanic	**O**ther	**Total**
Male	8,838	2,070	49	517	15	11,489
Female	1,494	634	6	30	5	2,169
Total	10,332	2,704	55	547	20	**13,658**

In this section, we will discuss the reasons for these discrepancies in gender and race and further investigate similar inconsistencies in date of birth. Notice that some individuals were arrested multiple times over this five year span, with an extreme case that one individual was arrested 53 times. Thus many subjects have multiple entries in the MORPH-II dataset. Recall that most of the data gathered for mugshots is self-reported, with technology used to verify the information reported. Some of these subjects have more than one gender, race, and/or birthdate reported across their database entries. This may cause critical problems when trying to use the MORPH-II images to build facial demographic systems, such as age estimation, or race classification. The inconsistencies among gender, race and birthdate are summarized in Table 2.4. Note that for the 457 subjects with only one entry in the dataset, there is no way to check whether the reported information is correct or not.

TABLE 2.4. MORPH-II Inconsistencies by Attribute

Attribute	Number of Subjects
Gender	1
Race	33
Birthdate	1,779

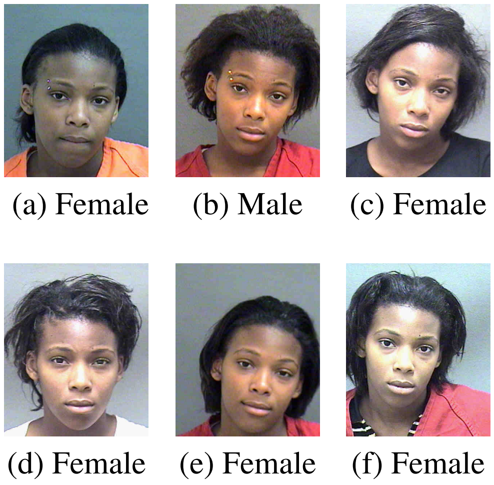
2.3. Cleaning Process

In the following section, we will discuss the methods that are used to resolve the inconsistencies in MORPH-II in details.

2.3.1. Cleaning for Gender Inconsistency

From the previous discussion, it is noted that there is only one subject with inconsistent gender in the database. It turned out that an individual was listed as both male and female in the data entries, as shown in Figure 2.1. Since 5 of the 6 images were marked as female, and this person does appear to be a female, we changed picture (b) in Figure 2.1 to female.

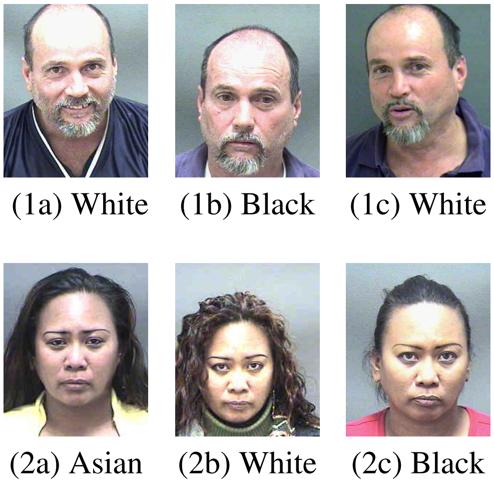
FIGURE 2.1. Gender Inconsistency



(a) Female (b) Male (c) Female

(d) Female (e) Female (f) Female

FIGURE 2.2. Race Inconsistencies



(1a) White (1b) Black (1c) White

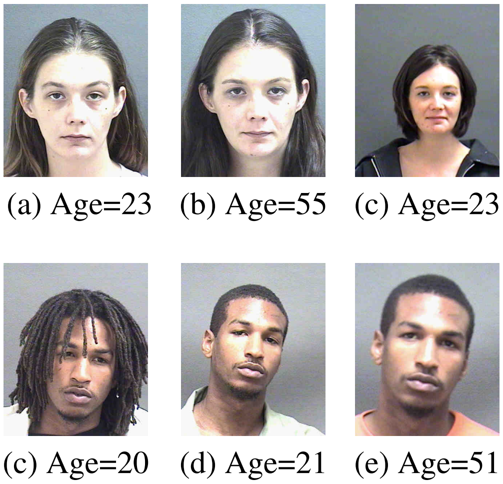
(2a) Asian (2b) White (2c) Black

2.3.2. Cleaning for Race Inconsistencies

There are 33 subjects with 132 images in MORPH-II with inconsistent race. In order to best determine race in the MORPH-II dataset, human perception was utilized. To reduce personal bias, a group of researchers were trained by the following steps: First, literature on race classification was carefully selected and reviewed, including (Guo and Mu, 2010b), (Guo and Mu, 2010a), and (Fu et al., 2014). The literature outlines the significance of eyes and nose and the insignificance of features such as skin tone. Researchers were also made aware of possible bias from the other-race-effect: the tendency to more easily recognize faces of one's own race. The most popular and effective methods of perceiving race were summarized to create race perception guidelines. Next, researchers were trained on human race perception with correctly labeled images in MORPH-II. The images of Asians and Hispanics were focused on as these made up the majority of the misclassifications.

After reviewing literature and being trained on race perception, the researchers attempted to identify the race of the subjects in question. To start, any individual with a clear majority of images belonging to one race was identified as this majority race. For example, the first subject in Figure 2.2 has 24 images which are classified as White, while 1 image is classified as Black. Therefore this subject is classified as white and (1b) is changed to White. In cases without a clear majority, human race perception was applied to perceive the race of the individual by using the information gathered from the literature and the training acquired from the rest of the dataset. In this process, multiple perspectives were also considered to eliminate as much bias as possible. Finally, in the case that the race of the individual is unclear or not enough information is available, the subject is identified as the Other race category. For example, the second subject in Figure 2.2 is identified as Other because she does not clearly exhibit only one race.

FIGURE 2.3. Worst Inconsistent Birthdates



(a) Age=23 (b) Age=55 (c) Age=23

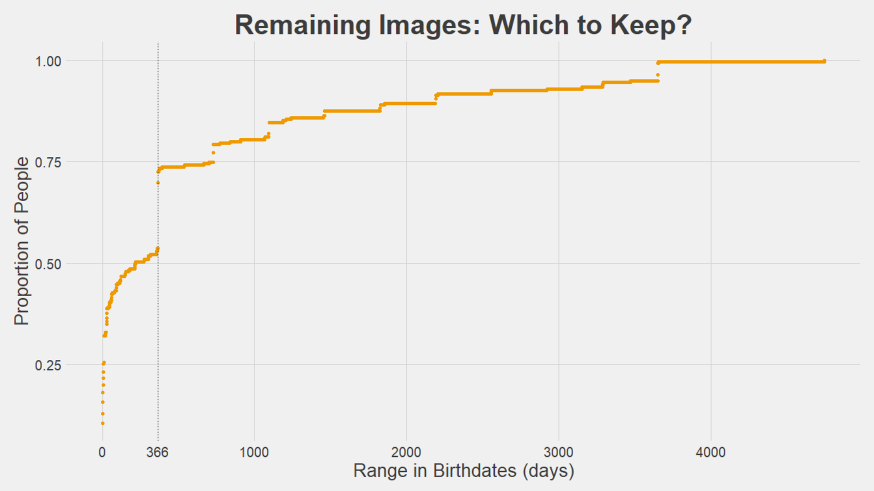
(c) Age=20 (d) Age=21 (e) Age=51

2.3.3. Cleaning for Birthdate Inconsistencies

There are 1,779 subjects in MORPH-II with inconsistent birthdates. 1,524 of the 1,779 subjects were identified and resolved with a simple majority, much like with person 1 in Figure 2.3. However, the remaining 255 subjects pose additional problems. For some of them, their birthdates are in a multiway tie. For others, there is no majority, or their birthdates differ by several years. This makes it difficult to choose one birthdate over another.

Figure 2.4 shows the proportion of subjects' birthdates that are within a given range in days. For each subject whose birthdates differed by no more than one year (366 days to include leap year), we calculated the mean birthdate and assigned this date to all images of this subject. This strategy

FIGURE 2.4. MORPH-II Inconsistency by Age



was applied to 185 subjects. The remaining 70 subjects with a total of 230 images were set aside as *Not For Training*. In general, previous studies have used the floor age instead of the decimal age. For example, this means that a subject who is actually 20.6 years old is recorded as simply 20 years old. Thus, it is reasonable to use the mean birthdate when the range is within one year.

2.4. Multiple Versions of Cleaned Datasets

After cleaning MORPH-II of gender, race, and birthdate consistencies, three cleaned datasets were created:

- *morphII_cleaned_v2* - same as original dataset (morph_2008_nonCommercial.csv), but with dob, race, and gender inconsistencies corrected.
- *morphII_go_for_age* - individuals with unidentifiable birthdates are removed from the above dataset. This leaves all the images with consistent age information that are ready for training and testing age estimation models.
- *morphII_holdout_for_age* - images with unidentifiable birthdates (greater than 1 year difference in the inconsistent birthdates).

Two new variables are created for each of the above datasets, shown below in Table 2.5. The corrected column takes a value between 0 and 8 representing what changes were made to a given entry. The indicators and their associated meanings are explained in Table 2.6.

It is noted that there are N=55,134 data entries in the dataset of *morphII_cleaned_v2*, while N=54,904 in *morphII_go_for_age*, and N=230 in *morphII_holdout_for_age*. That means there are only 230 images with unidentifiable birthdates, with greater than 1 year difference in the inconsistent birthdates.

TABLE 2.5. New Variables Created

Variable	Information
corrected	indicator (0-8)
age_dec	decimal age ($dob - doa$)

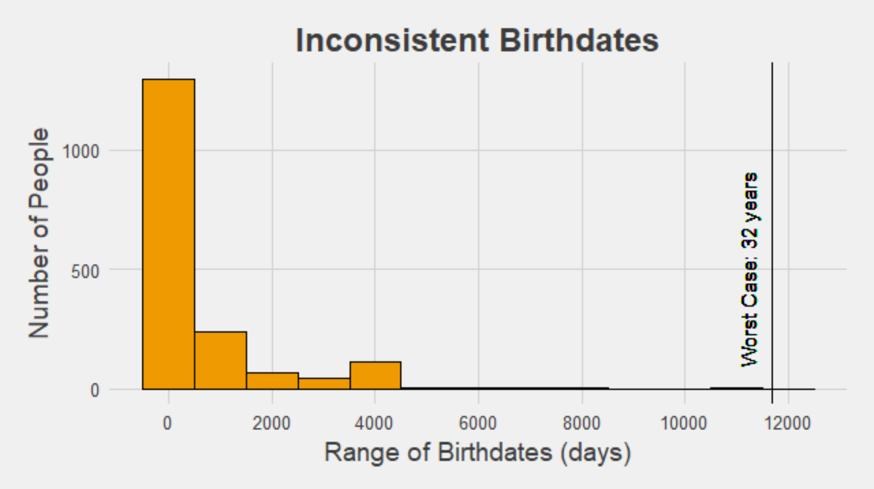
TABLE 2.6. Indicators for new variable *corrected*

Indicator	Information	# of images
0	no change	**52,414**
1	dob - majority	1,906
2	dob - averaged	515
3	dob - uncorrectable	230
4	race - majority	11
5	race - perception	22
6	race - too difficult to tell, assigned to Other	33
7	more than 1 change	2
8	gender corrected	1
		Total = 55,134

2.5. Research Based on Uncleaned MORPH-II Data

A substantial amount of research has been done on the MORPH-II dataset. Unfortunately, when researchers report, for example, that the total number of subjects in the dataset is 13,618 (when it is actually 13,617) or that the number of males classified as "Other" is three (upon further inspection one of these three has inconsistent race), this indicates that data used in such research were not properly cleaned. Without discrediting the important contributions that have been made, such research outcomes could be more accurate if the data were cleaned properly.

FIGURE 2.5. Inconsistent Birthdates Summary



There will not likely be an enormous impact on model performance for gender or race prediction, because the number of gender and race inconsistencies is relatively small. However, age estimation models may see an increased Mean Absolute Error (MAE). Figure 2.5 shows the summary for inconsistent birthdates. The worst case is a subject whose reported birthdates are 32 years apart (see Figure 2.3). In some cases, subjects' birthdates change so that in the dataset their reported age decreases with time. This could significantly affect models concerned with age estimation or age progression.

3. Automatic Subsetting Scheme for Evaluation Protocol

The MORPH-II database is the largest publicly available longitudinal face database with over 55,000 face images. It contains about 77% of Black faces, 19% of White, and the remaining 4% of Hispanic, Asian, Indian, and Other.

3.1. Background

In the computer vision field, the performance of a learning algorithm is usually measured in terms of prediction error. In most real-world problems, the prediction error is unknown, and thus must be estimated. It is important to choose an appropriate estimator of the prediction error. In general, test errors can be estimated by the cross-validation technique in order compare the performances between different models. In order to prevent the information leaking such that the algorithm learns the identity of a subject from the training set rather than the proposed algorithm itself, one need to make sure that all images of each individual subject must be placed into either the training set or the testing set, but not both. Moreover, the folds should be selected in such a way that distributions of age, gender and ethnicity in each fold should be as similar to the distribution of the full dataset as possible. Therefore in general, researchers manually split the full dataset into non-overlapping folds, based on the predetermined proportion of age, gender, and race.

It is challenging that the majority of individuals in MORPH-II were arrested multiple times over this five year span, each of those subjects contains more than one image in the database. Additionally, the MORPH-II is highly imbalanced towards black male subjects. Therefore, Guo et al. (Guo and Mu, 2010a,b, 2011) proposed an evaluation protocol which has been adopted by many studies on MORPH-II. Their work is based on a previous version (with unknown release date) of the MORPH-II noncommercial dataset, which here we will refer to as MORPHpre. MORPHpre and MORPH-II are very similar, but there are some minor differences. Most images are the same, but there are 2 images in MORPH-II that are not included in MORPHpre. MORPHpre includes an additional race category Indian, in addition to the 5 races in MORPH-II: White, Black, Asian, Hispanic, and Other. Beyond the specific differences mentioned, MORPH-II is an updated, cleaner version of MORPHpre. This information is briefly summarized in Table 3.1.

TABLE 3.1. Differences between MORPHpre and MORPH-II

	Number of Images	**Race Categories**	**Other Qualities**
MORPHpre	55,132	W,B,H,A,I,O	older version
MORPH-II	55,134	W,B,H,A,O	newer version; cleaner

For comparative purposes, the experimental design by Guo and Mu from (Guo and Mu, 2010a,b, 2011) will be first summarized as follows. About 66.75% of images in MORPHpre are of black males, while only about 4.72% of images are of white females. Guo and Mu utilize a subsetting scheme to overcome such disproportionate distributions of racial and gender groups in MORPHpre. They denote the whole dataset as W and randomly select a subset S of 21,060 white and black faces of both genders from ages 16 to 67. Only white and black subjects are included in S, because other races have too few images for use in the training set. In S, half the images are white, and the other

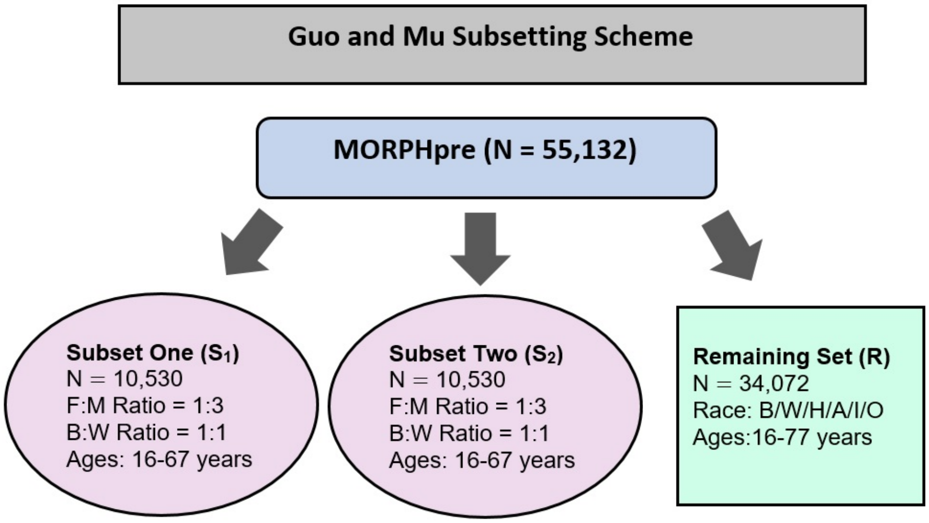


FIGURE 3.1. Flowchart for the evaluation protocol by Guo's and Mu's in (Guo and Mu, 2010a,b, 2011).

half are black. There are three times as many males as females in S. R denotes the set of remaining images, $W \setminus S$. In R, there are approximately 34,000 images including both genders, all 6 races, and ages 16-77 years. These subsets are summarized in Figure 3.1.

Then S is further divided equally into $S1$ and $S2$, with 10,530 images for $S1$ and $S2$, respectively. $S1$ and $S2$ are used alternately for training and testing purposes. First, $S1$ is used for training and $W \setminus S1$ for testing. Then $S2$ is used for training and $W \setminus S2$ for testing. This process is similar to 2-fold cross-validation, but $S1$ and $S2$ are not obtained by a random partition of the images in S; instead, $S1$ and $S2$ are controlled to be as similar as possible. Note that R is not used in the training set at all.

In $S1$ and $S2$, it is aimed to ensure equivalent age distributions within gender in (Guo and Mu, 2010a). For example, the set of white females in S1, $S1_{WF}$, has an identical age distribution to the set of black females in S1, $S1_{BF}$. The equivalence of age distributions is summarized in Figure 3.2. Through this subsetting scheme, Guo and Mu form training sets that are more proportionate in gender and race. Additionally, they guarantee identically distributed ages within each gender class.

However, it can be tedious and time consuming to manually split the full data into folders with all the criteria mentioned above. Therefore, we aim to investigate how to automatically split the MORPH-II into non-overlapping folds according to evaluation protocol proposed by Guo et al.

3.2. Automatic Subsetting Scheme

3.2.1. Development

Following the evaluation protocol proposed by Guo and Mu, we only consider the white and black races for the training sets, since the number of images for other racial groups is too small. In our subsetting scheme, we seek to retain the ratios in (Guo and Mu, 2010a) of white:black and male:female images, while ensuring independence between training and testing sets. We also prioritize randomization, aiming to create many candidate subsets from which to choose. These

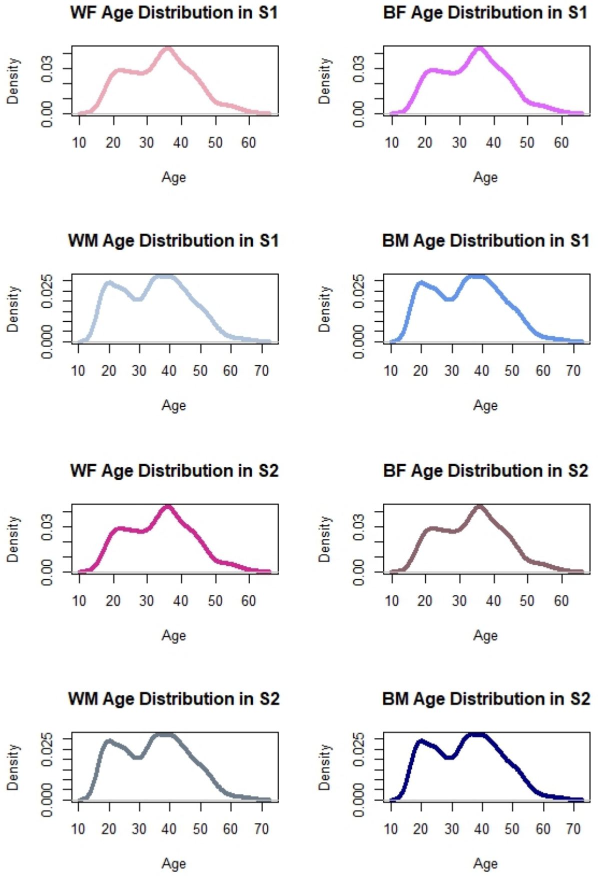


FIGURE 3.2. Age Distributions for subsets $S1$ and $S2$ in S are shown for Guo's and Mu's subsetting scheme (Guo and Mu, 2010a,b, 2011). On the top row of the figure, it is shown that $S1_{WF}$, $S1_{BF}$, $S2_{WF}$, and $S2_{BF}$ have identical age distributions. The bottom row of the figure depicts the identical age distributions for $S1_{WM}$, $S1_{BM}$, $S2_{WM}$, and $S2_{BM}$.

possible subsets can be used for comparative purposes in the future; models could be built and validated on different subsets, and the results could be averaged or compared.

Hereafter, the cleaned version of MORPH-II *morphII_go_for_age* from above is used for the generation of subsets in this study. Recall that there are N=55,134 data entries in the cleaned full dataset of *morphII_cleaned_v2*, while N=54,904 in *morphII_go_for_age*, and N=230 in *morphII_holdout_for_age*. It indicates there are only 230 images with unidentifiable birthdates, with greater than 1 year difference in the inconsistent birthdates. Note that a new variable *age dec* is also created to represent the exact age in decimal of the subject pictured in each image. Hereafter, those age values in decimal will be used in subsetting and future facial demographical study, since they are less biased than integer-valued ages. The decimal age values are also advantageous due to their improved continuity, which is essential as an assumption for the nonparametric tests discussed in a later section.

For ease of comparison, we attempt to be as consistent as possible with the set notation in (Guo and Mu, 2010a,b, 2011). Let W be the Whole cleaned dataset (*morphII_go_for_age*), S be the main training/validation set, and R be the remaining set. We divide S into $S1$ and $S2$, such that $S1$ and $S2$

have the same number of images. We fix the ratios of white:black images to be 1:1 and male:female images to be 3:1.

Because white females are the smallest race-gender combination, we include all 2,570 white females in S. We randomly allocate each white female subject to either $S1$ or $S2$ exclusively, according to the constraint that the total number of white female images in S1 is equal to the number of white female images in S2:

$$|S1_{WF}| = 1285 = |S2_{WF}|.$$

Note that all white females are included in S, hence none are included in R.

For the other race-gender categories (black females, white males, and black males), we include only a portion of their images in S, while the remainder goes in R. For black females, we randomly allocate a subset of subjects to $S1$ and an exclusive subset of subjects to $S2$, such that the total number of black female images in $S1$ is equal to the total number of black female images in S2, as well as equal to the total number of white female images in S_i, $i = 1,2$. The images pertaining to any remaining black female subjects are sent to R.

$$|S1_{BF}| = 1285 = |S2_{BF}| = |Si_{WF}|, i = 1,2.$$

For white males, we randomly allocate some subjects to $S1$ and other distinct white male subjects to $S2$, such that 3 times the number of white female images are in S1. The number of white male images in $S1$ is also set to be equal to the number of white male images in $S2$. Any remaining white males images are sent to R.

$$|S1_{WM}| = 3855 = |S2_{WM}| = 3|Si_{WF}| = 3|Si_{BF}|, i = 1,2.$$

The same process is repeated for black males, so that there are equal numbers of black male images within S1 and S2. The number of black male images is equal to the number of white male images, and other equalities hold too:

$$|S1_{BM}| = 3855 = |S2_{BM}| = |Si_{WM}| = 3|Si_{WF}| = 3|Si_{BF}|, i = 1,2.$$

In this way, we ensure independence between $S1$, $S2$, and R. There is no expected information leakage between the training and testing sets. However, it should be clarified that observations within each set are not independent. For each subject s_j in some set Ω, all of s_j's images are in Ω. Hence, some observations within each set Ω are correlated with each other. Though we are able to guarantee independence between training and testing sets, at this time we have no satisfactory solution to the issue of correlated images within each set. This is an issue inherent to longitudinal data.

3.2.2. Implementation

We implement our subsetting scheme in the statistical software R. We iterate through various random seeds of k ($k = 1,2,...$). Subsets are randomly generated for each value of k, with a random seed of k set anytime randomization is invoked. In this way, numerous candidate subsets are created.

Among the candidate subsets, we seek those with similar age distributions. We obtain the age distributions of images in $S1$ and $S2$. Then for each value of k, we perform both the Anderson-Darling (AD) and Kolmogorov-Smirnov (KS) tests on those distributions. The hypotheses for both tests are as follows:

$$H_o : S1_{age} \text{ has the same distribution as } S2_{age}$$

$$H_a : S1_{age} \text{ does not have the same distribution as } S2_{age}$$

Both the AD (Darling, 1957; Pettitt, 1976) and the KS test (Kolmogorov, 1933) are based on the empirical distribution function (EDF) of data to exam whether two samples come from two identical distributions. For a sample of n observations z_1, z_2, $\cdots$, z_n, the empirical distribution $\hat{F}_n(x)$ can be calculated as follows:

$$\hat{F}_n(x) = \begin{cases} 0, & \text{if } x < z_{(1)}, \\ i/n, & \text{if } z_{(i)} \leq x < z_{(i+1)}, \\ 1, & \text{if } x \geq z_{(n)}, \end{cases}$$

where $z_{(1)} < z_{(2)} < \cdots < z_{(n)}$ are the ordered sample. $\hat{G}_m(x)$ can be defined in a similar manner.

The two-sample AD test is defined by:

$$D_{AD}^2 = \frac{nm}{N} \int_{-\infty}^{\infty} \frac{\{\hat{F}_n(x) - \hat{G}_m(x)\}^2}{H_N(x)\{1 - H_N(x)\}} dH_N(x),$$

where $H_N(x) = (n\hat{F}_n(x) + m\hat{G}_m(x))/N$ is the empirical distribution function of the pooled sample and $N = n + m$. If $H_N(x) = 1$, then the integrand is defined as 0 conventionally.

On the other hand, the Kolmogorov-Smirnov test statistic is the maximum of the absolute values of the difference between two empirical distributions $\hat{F}_n(x)$, and $\hat{G}_m(x)$, where n and m are the sample sizes respectively, and $N = n + m$:

$$D_{KS} = \sqrt{\frac{nm}{N}} \sup_x |\hat{F}_n(x) - \hat{G}_m(x)|.$$

The Kolmogorov-Smirnov test can be carried out as an exact permutation test in order to obtain the p-value. For both AD and KS tests, the null hypothesis that two samples come from two identical distributions is rejected if D_{AD} or D_{KS} is larger than the critical value at a given α. Tables of critical values for different sample sizes for AD test and KS test have been published.

TABLE 3.2. Number of Images in Subsets by Race and Gender

	WF	**BF**	**WM**	**BM**	**dF**	**dM**	**Overall**	**F**	**M**
S1	1,285	1,285	3855	3,855	0	0	10,280	2570	7,710
S2	1,285	1,285	3,855	3,855	0	0	10,280	2,570	7,710
R	0	3,150	220	28,980	144	1,850	34,344	3,294	31,050
Overall	2,570	5,720	7,930	36,690	144	1,850	54,904	8,434	46,470

TABLE 3.3. Number of Distinct Subjects in Subsets by Race and Gender

	WF	**BF**	**WM**	**BM**	**dF**	**dM**	**Overall**	**F**	**M**
S1	311	332	1,005	948	0	0	2,596	643	1,953
S2	313	336	988	943	0	0	2,580	649	1,931
R	0	809	55	6,899	40	568	8,371	849	7,522
Overall	624	1,477	2,048	8,790	40	568	13,547	2,141	11,406

We use the P-Values of both tests to identify the best subsets. High P-Values indicate $S1$ and $S2$ have similar age distributions for a particular seed k. Hence, we use the P-Values for these nonparametric tests as metrics for judging suitable subsets. In this context, the P-Values are not to

TABLE 3.4. Additional Race Groups in Remaining Subset R

	HF	AF	OF	HM	AM	OM	Overall	F	M
Subjects in R	28	4	8	502	47	19	608	40	568
Images in R	99	13	32	1,646	140	64	1,994	144	1,850

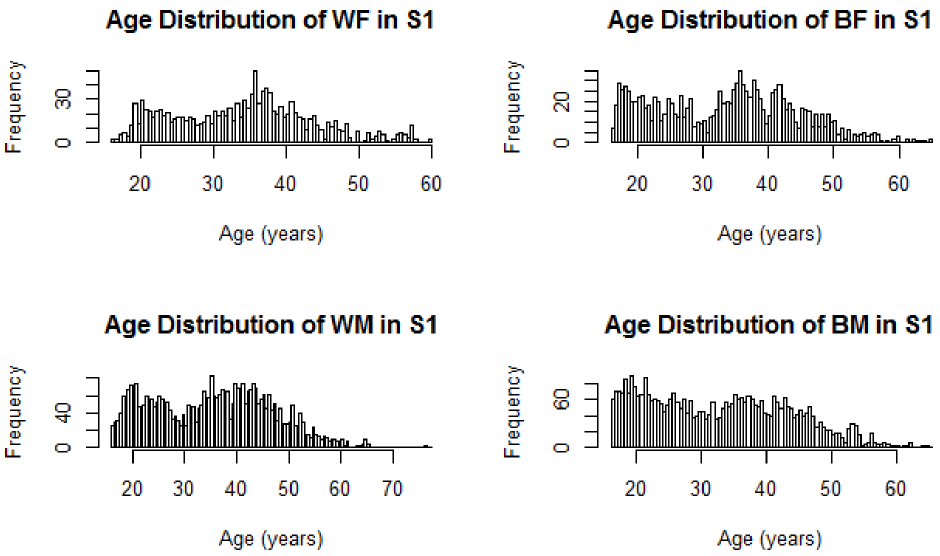


FIGURE 3.3. For random seed 42, the observed age histograms in *S*1 are displayed.

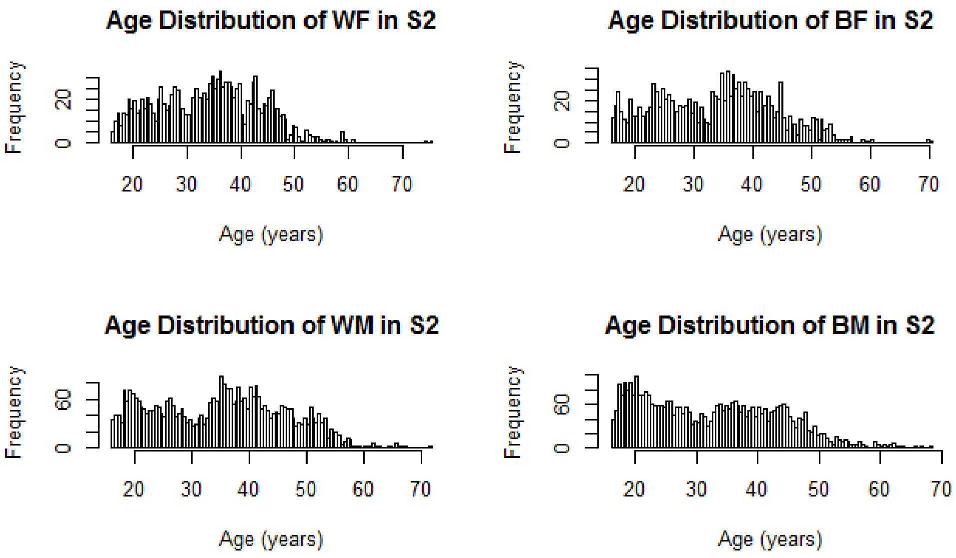


FIGURE 3.4. For random seed 42, the observed age histograms in *S*2 are displayed.

be interpreted as clear probabilities, for the following reasons: not all assumptions for the tests are met (since some observations within each set are dependent) and the significance level cannot be

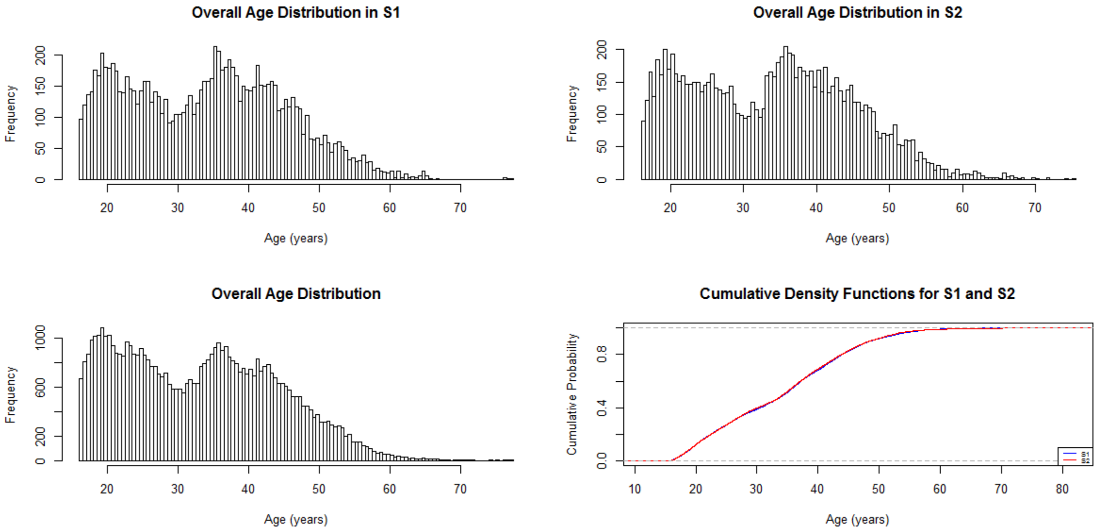


FIGURE 3.5. The age distributions for images in $S1$, $S2$, and the W dataset are depicted.

defined appropriately when an indefinite number of tests are made. We believe our unconventional use of P-Values is valid here, since we are not attempting to make any probability statements based off them.

Using these criteria, we identify random seed number $k = 42$ as one which produces satisfactory subsets. The P-Values for the KS and AD tests are 0.657 and 0.652, respectively. The statistical summaries below further indicate the suitability of these subsets. We do not guarantee that random seed 42 produces the global optimum results, but it is found to be satisfactory for our purposes.

In Tables 3.2, 3.3, and 3.4, we include basic information pertaining to the subsets generated by the random seed 42. We intend the subsets W, S, and R to be used as Guo and Mu did in (Guo and Mu, 2010a,b, 2011): first the model should be trained on $S1$ and tested on $W \setminus S1$, then the model should be trained on $S2$ and tested on $W \setminus S2$. Then two sets of results can be summarized.

Table 3.2 shows the number of images in subsets by race and gender, while Table 3.3 gives the number of distinct subjects in subsets by race and gender. More detailed information on the different races is summarized in Table 3.4, with the number of images and the number of distinct subjects for the additional race groups in the remaining subset R. In Tables 3.2 and 3.3, **d** denotes different race subjects (**H**ispanic, **A**sian, or **O**ther).

Additional graphical and numerical summaries are presented in Figures 3.3, 3.4, 3.5, and Table 3.5. Figure 3.3 displays the observed age histograms in $S1$. It is shown that all the gender-race combinations in $S1$ have similar, right-skewed age distributions. Any differences in distribution here seem minor and unlikely to significantly affect gender or race classification in future experiments. Figure 3.4 presents the observed age histograms in $S2$. Based on the plots, all the gender-race combinations in $S2$ seem to be similarly distributed. Further, we see that the age distributions in $S1$ are not much different than the age distributions in $S2$. We do not expect any deviations in age distribution between sets to negatively impact classification in a significant way.

The age distributions for images in $S1$, $S2$, and the W dataset are depicted in Figure 3.5. It is shown that all three histograms are right-skewed with a roughly bimodal structure, indicating that $S1$ and $S2$ have been chosen successfully; the age distributions of images in the subsets $S1$ and $S2$ are close to the overall age distribution of images in W. The final plot shows the ECDFs of $S1$ and

$S2$. It is difficult to distinguish the densities corresponding to each subset, since departures are so minor. This aligns with our expectations, because the P-Values for the KS and AD tests were quite large (approximately 0.65 for each).

In Table 3.5, the 5-Number Summary, as well as mean and standard deviation, are given for age in $S1$, $S2$, and W. The statistical summaries are nearly identical, further confirming these subsets' balanced age distributions.

TABLE 3.5. Numerical Summary of Age in Sets

	Min.	**Q1**	**Median**	**Q3**	**Max**	**Mean**	**SD**
S1	16.003	24.296	34.495	42.185	77.196	34.041	10.957
S2	16.005	24.370	34.371	42.014	75.421	33.926	10.908
W	16	23.369	33.091	41.422	77.196	33.019	10.950

All numerical and graphical summaries we consider here indicate the suitability of the subsets generated from random seed 42. These subsets are expected to yield good results for a variety of face imaging tasks, including gender and race classification, as well as age regression.

4. Conclusion

Our first preliminary study on data validation and cleaning are critical before any research work is conducted. This not only preserves the accuracy of research results, but also the integrity. Many researchers base their work off of previous results, making it even more important to ensure that one's own work is accurate.

In our second preliminary study, we propose an automatic subsetting scheme of the MORPH-II aging database. Our scheme is inspired by the work of Guo and Mu, but we do make some changes. Most notably, we maintain the racial and gender proportions of Guo and Mu, while ensuring independence between training and testing sets. Our approach is also novel in its generation of various candidate subsets, which are selected based off nonparametric goodness of fit tests KS and AD. We present one suitable choice of subsets for a random seed of 42, but the generation of other subsets from the random seeds k are recommended for comparative purposes in the future. For any models built and tested using the subsetting scheme proposed in this study, it is expected the estimates of test error or accuracy can be less biased. Our automatic subsetting scheme can be used for face imaging tasks involving gender, race, and age. Even though the automatic subsetting scheme is illustrated on the MORPH-II dataset, it can be extended to other experimental designs.

5. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Numbers DMS-1659288. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Antipov, G., Baccouche, M., Berrani, S.-A., and Dugelay, J.-L. (2017). Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, 72:15–26.

Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838.

Fu, S., He, H., and Hou, Z.-G. (2014). Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2483–2509.

Fu, Y., Guo, G., and Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1955–1976.

Guo, G. and Mu, G. (2010a). Human age estimation: What is the influence across race and gender? In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 71–78. IEEE.

Guo, G. and Mu, G. (2010b). A study of large-scale ethnicity estimation with gender and age variations. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 79–86. IEEE.

Guo, G. and Mu, G. (2011). Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pages 657–664. IEEE.

Han, H., Otto, C., Liu, X., and Jain, A. K. (2015). Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1148–1161.

Jana, R. and Basu, A. (2017). Automatic age estimation from face image. In *Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on*, pages 87–90. IEEE.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

Liu, K.-H., Yan, S., and Kuo, C.-C. J. (2015). Age estimation via grouping and decision fusion. *IEEE Transactions on Information Forensics and Security*, 10(11):2408–2423.

MORPH-II. *MORPH Non-Commercial Release Whitepaper*. `http://www.faceaginggroup.com`.

Niu, Z., Zhou, M., Wang, L., Gao, X., and Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928.

Pettitt, A. N. (1976). A two-sample anderson-darling rank statistic. *Biometrika*, 63(1):161–168.

Ricanek, K. and Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE.

(C. Chen) Department of Mathematics and Statistics, The University of North Carolina Wilmington, Wilmington, NC 28403, USA

E-mail address, Corresponding author: `chenc@uncw.edu`