MORPH-II: Inconsistencies and Cleaning

G. Bingham, B. Yip, M. Ferguson, and C. Nansalo
The University of North Carolina at Wilmington

Abstract

This paper gives a detailed summary of the inconsistencies in the non-commercial release of the MORPH-II dataset and covers the steps that were taken to clean it. In addition, we briefly introduce prior research that made use of the uncleaned data and discuss the potential implications for their results.

1 Introduction

The non-commercial release of the MORPH-II dataset is collection of 55,134 mugshots taken between 2003 and 2008 and includes many images of individuals that were arrested multiple times over this 5 year span. This gives the data a longitudinal aspect that has made it very useful in the field of computer vision and pattern recognition. In particular, the MORPH-II dataset is widely utilized in research on gender and race classification, as well as age estimation and synthesis. Our team was preparing to use this dataset for much the same purposes when our preliminary exploration of the data uncovered numerous inconsistencies. As far as we are aware, previous research efforts with MORPH-II have neglected to correct or acknowledge these errors. Accordingly, this paper is intended to provide a thorough explanation of the inconsistencies in MORPH-II and explicitly detail our cleaning methodology.

1.1 The Original Data

The dataset in question is the 2008 MORPH non-commercial release. The below tables were drawn from the whitepaper attached to this release and summarize the original data.

Table 1: Number of Images by Gender and Race

	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**

Table 2: Metadata

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

1.2 Discovering Errors

When we first discovered errors in MORPH-II, we were looking for the number of distinct subjects in the whole dataset (**13,617**), and after subsetting it by male and female. Table 3 (below) summarizes our findings:

Table 3: Number of Distinct Individuals by Gender

	Distinct Individuals
Male	11,459
Female	2159
Total	13,618

Note that the total number of distinct subjects in the dataset is 13,618 when computed as the sum of male and female individuals. This is obviously inconsistent with our original result and suggested to us that further inspection was necessary. Introducing subsetting by race as well as gender gave the following table:

Table 4: Number of Distinct Individuals by Race and Gender

	Black	**W**hite	**A**sian	**H**ispanic	**O**ther	**Total**
Male	8838	2070	49	517	15	11489
Female	1494	634	6	30	5	2169
Total	10332	2704	55	547	20	13658

Again the total number of distinct individuals in Table 4 does not agree with the true count (**13,617**), illustrating the extent of the inconsistencies in the dataset. The next section details the reasons for these discrepancies and introduces the inconsistencies regarding date of birth.

2 Inconsistencies in the Data

Repeat offenders have multiple entries in the MORPH-II dataset. There are some people with more than one gender, race, and/or birthdate. This causes problems when trying to use the images to predict demographics. Below we summarize the inconsistencies we found.

Table 5: MORPH-II Inconsistencies by Attribute

Attribute	Number of People
Gender	1
Race	33
Birthdate	1779

Note that for people with only one entry in the dataset (there are 457 of them) there is no way to check whether their information is consistent or not. We summarize how we cleaned the data in the following section.

3 Cleaning Process

3.1 Gender

There was only one person with inconsistent gender in the database. Since 5 of the 6 images were marked as female, and this person does indeed appear to be a female, we changed picture (b) to female.

(a) Female

(b) Male

(c) Female

(d) Female

(e) Female

(f) Female

3.2 Race

In order to properly correct the inconsistent races in the data set, we wanted to insure our personal bias would have as little influence as possible. To do se we review literature on race classification, and trained our eyes on the correctly classified images in the data set. We reviewed three articles; A Study of large-scale Ethnicity Estimation with gender and age Variation , Human Age Estimation: What is the influence across race and gender , and Learning Race from Face: A Survey. In these paper we learned of the most popular and effective was of identifying race. The literature outlines the significance of eyes and nose and the insignificance of features such as skin tone. We were also made aware of possible bias from the other-race-effect; the tendency to more easily recognize faces of the race that one is most familiar with, which is often ones on race. To reduce our human bias, we trained our eyes. Similar to training a model, we used the correctly classified people in the data set to train our eyes. We focused on the images of Asians, and Hispanics as these made up the majority of the misclassifications. Our training showed us the category of "Asian" was specific to east Asians.

After we trained our eyes on the dataset and reviewed the relevant literature, we 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 in the images below person 1 has 24 images classified as White and 1 image classified as Black, therefore he was classified as white. In the case that the individual was identified as several different races without a clear majority, we attempted to identify the true 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 was unclear or not enough information was available, we identified the individual as the "other" race category. Person 2 below was identified as "other" because she does not clearly exhibit only one race.

(1a) White

(1b) Black

(1c) White

(2a) Asian

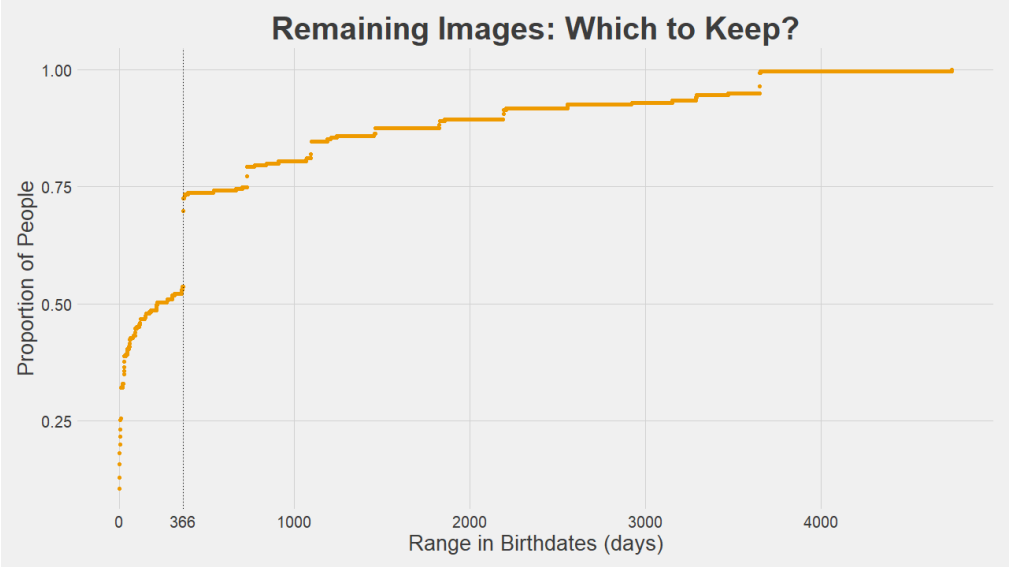
(2b) White

(2c) Black

3.3 Birthdate

Similar to cleaning the race data, we were able to use a simple majority for 1524 of the 1779 people with inconsistent birthdates.

However, the remaining 255 people posed additional problems. For some of them, their birthdates were in a multiway tie. For others, there was no majority, or their birthdates differed by several years. This made it difficult to choose one birthdate over another.



The graph above shows the proportion of peoples' birthdates that are within a given range. For each person whose birthdates differed by no more than one year, we calculated the mean birthdate and assigned this date to all images. The remaining images were set aside as Not For Training. In general, previous researchers have used the floor age instead of the decimal age. For example, this means that someone who is actually 20.6 years old is recorded as simply 20 years old. Thus, we feel that the difference of one year is appropriate. We allowed a difference of 366 days to account for some leap years.

4 New Datasets

After cleaning, three new datasets were formed:

- 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 uncorrectable birthdates were 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 uncorrectable birthdates (greater than 1 year difference).

Each of the above datasets have two new variables, shown below in Table 6.

Table 6: New Variables

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

The corrected column takes a value between 0 and 8 depending on how it was changed. The indicators and their associated meanings are as follows:

0 no change
1 dob - majority
2 dob - averaged
3 dob - uncorrectable
4 race - majority
5 race - perception
6 race - too difficult to tell, assigned to Other
7 more than 1 change
8 gender corrected

5 Dirty Data

5.1 Other Researchers' Work

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 3 (upon further inspection one of these three has inconsistent race), this tells us that they haven't cleaned their data properly.

We do not wish to discredit the important contributions that have been made. However, had these researchers properly cleaned their data, they could have seen increased accuracy and lower error. Below is an example of some papers that were based on an uncleaned version of MORPH-II. It is not an exhaustive list, but rather a sampling for illustrative purposes.

[1] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR 2011*, pages 657–664, June 2011.

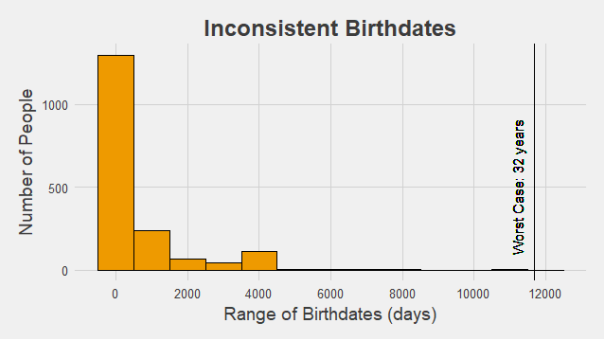
[2] K. H. Liu, S. Yan, and C. C. J. Kuo. Age estimation via grouping and decision fusion. *IEEE Transactions on Information Forensics and Security*, 10(11):2408–2423, Nov 2015.

[3] X. Wang, V. Ly, G. Lu, and C. Kambhamettu. Can we minimize the influence due to gender and race in age estimation? In *2013 12th International Conference on Machine Learning and Applications*, volume 2, pages 309–314, Dec 2013.

[4] D. H. P. Yassin, S. Hoque, and F. Deravi. Age sensitivity of face recognition algorithms. In *2013 Fourth International Conference on Emerging Security Technologies*, pages 12–15, Sept 2013.

5.2 Consequences of Using Uncleaned MORPH-II

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 small.



Age estimation models will see a drop in overall performance manifest in a higher Mean Absolute Error (MAE). For some people in the dataset, their birthdates vary enough that their age decreases with time. This will significantly affect models concerned with age progression.

6 Conclusion

Cleaning the data before doing research is vital. This not only preserves the accuracy of one's 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.

7 Acknowledgements

We would like to thank Dr. Chen, Dr. Wang, and Troy Kling at UNCW, as well as the NSF for funding our REU program.