MORPH-II: Inconsistencies and Cleaning Whitepaper

Participants: G. Bingham, B. Yip, M. Ferguson, and C. Nansalo
Mentors: C. Chen, Y. Wang, and T. Kling
NSF-REU Site at UNC Wilmington, Summer 2017

Abstract

This paper presents a detailed summary of the inconsistencies in the non-commercial release of the MORPH-II dataset and covers 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.

1 Introduction

Since its first release in 2006, the MORPH-II dataset [5] has been cited over 500 times. 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 dataset is a collection of 55,134 mugshots taken between 2003 and late 2007, and 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 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 similar purposes when our preliminary exploration of the data uncovered numerous inconsistencies. 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 being given. As far as we are aware, previous research efforts with MORPH-II have neglected to correct or acknowledge these errors. Accordingly, this paper aims to provide a thorough explanation of the inconsistencies in MORPH-II and to explicitly detail our cleaning methodology.

1.1 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 this release [1].

Table 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**

Table 2: Metadata

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

The gender and race distribution for MORPH-II is listed in Table 1, with metadata details provided in Table 2.

1.2 Discovering Errors

By using the 6-digit subject identifier, we found there to be 13,617 unique individuals in the MORPH-II dataset. However, when we count the number of unique males and unique females, we get the following result.

Table 3: Number of Distinct Individuals by Gender

	Distinct Individuals
Male	11,459
Female	2,159
Total	**13,618**

Table 3 shows that the total number of distinct subjects is now 13,618, suggesting that there may be an individual listed as both male and female. Repeating the same procedure for race produces similar results in Table 4.

Table 4: 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**

Clearly, the total number of distinct individuals in Table 4 (13,658) does not agree with the true count (13,617), illustrating the extent of the inconsistencies in the dataset. In the next section, we discuss the reasons for these discrepancies and investigate similar inconsistencies with date of birth.

2 Inconsistencies in MORPH-II

Many subjects have multiple entries in the MORPH-II dataset. However, some of these subjects have more than one gender, race, and/or birthdate across their database entries. This causes problems when trying to use the MORPH-II images to build facial demographic systems. The inconsistencies among gender, race and birthdate are summarized in Table 5.

Table 5: MORPH-II Inconsistencies by Attribute

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

Note that for the 457 subjects with only one entry in the dataset, there is no way to check whether their information is correct or not. The methods used to resolve the inconsistencies in MORPH-II are detailed in the following section.

3 Cleaning Process

3.1 Gender Inconsistency

There is only one subject with inconsistent gender in the database, as shown in Figure 1. 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.

Figure 1: Gender Inconsistency

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

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

Figure 2: Race Inconsistencies

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

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

3.2 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 [4], [3], and [2]. 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 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 is identified as "Other" because she does not clearly exhibit only one race.

3.3 Birthdate Inconsistencies

There are 1,779 subjects in MORPH-II with inconsistent birthdates. 1,524 of the 1,779 cases were resolved with a simple majority, much like with person 1 in Figure 2. 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 3: MORPH-II Inconsistency by Age

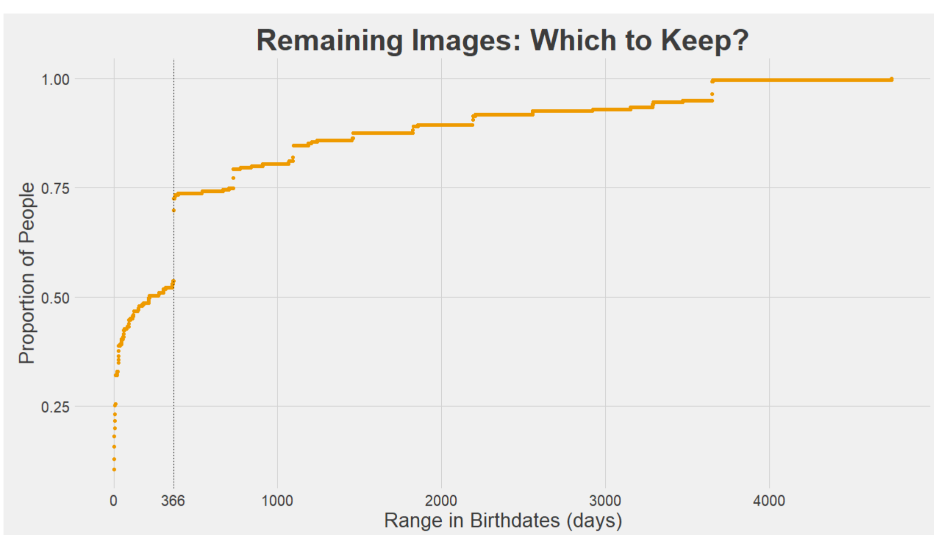

Figure 3 shows the proportion of subjects' birthdates that are within a given range in days. For each subject whose birthdates differed by no more than one year (366 days to include leap year), we calculated the mean birthdate and assigned this date to all images of this subject. This strategy was applied to 185 subjects. The remaining 70 subjects 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.

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 6. 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 7.

Table 6: New Variables Created

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

Table 7: Indicators for new variable *corrected*

Indicator	**Information**
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

Figure 4: Inconsistent Birthdates Summary

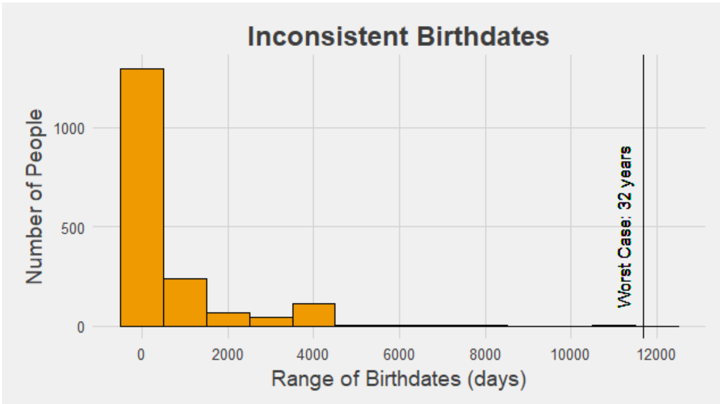

Figure 5: Worst Inconsistent Birthdates


(a) Age=23


(b) Age=55


(c) Age=23


(c) Age=20


(d) Age=21


(e) Age=51

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.

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 4 shows the summary for inconsistent birthdates. The worst case is a subject whose reported birthdates are 32 years apart (see Figure 5). 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.

6 Conclusion

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.

7 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

[1] *MORPH Non-Commercial Release Whitepaper*. `http://www.faceaginggroup.com`.

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

[3] G. Guo and G. Mu. 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, 2010.

[4] G. Guo and G. Mu. 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, 2010.

[5] K. Ricanek and T. Tesafaye. 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, 2006.