The Specs on Face dataset

Mahmoud Afifi

Electrical Engineering and Computer Science Dept., York University, Canada
Faculty of Computers and Information, Assiut University, Egypt
email: mafifi@eecs.yorku.ca - m.afifi@aun.edu.eg

Abstract

The Specs on Face (SoF) dataset [1] is a collection of 42,592 ($2{,}662 \times 16$) images for 112 persons (66 males and 46 females) who wear glasses under different illumination conditions. The dataset is free for reasonable academic fair use. The dataset presents a new challenge regarding gender classification, face detection, eyeglasses detection, emotion recognition, and facial landmark detection. It is devoted to two problems that affect face detection, recognition, and classification, which are harsh illumination environments and face occlusions. The glasses are the common real occlusion in all images of the dataset. However, the glasses are not the sole facial occlusion in the dataset; there are two synthetic occlusions (nose and mouth) added to each image. Moreover, three image filters, that may evade face detectors and facial recognition systems, were applied to each image. All generated images are categorized into three levels of difficulty (easy, medium, and hard). That enlarges the number of images to be 42,592 images (26,112 male images and 16,480 female images). There is metadata for each image that contains many information such as: the subject ID, facial landmarks, face and glasses rectangles, gender and age labels, year that the photo was taken, facial emotion, glasses type, and more.

1. Introduction

In the literature, there are many cases that have shown the bad effects of severe illumination conditions on both face detection and recognition [2, 3]. Although most face detection algorithms normalize the contrast of the input image as a preprocessing step, the face occlusions threatens the performance of many face detectors and facial recognition systems. For instance, Facebook's face detector is considered an invariant against dark images, whereas the accuracy, at least at this time, goes down with people who wear glasses or scarves [3]. As another example, the face detection accuracy of the Snapchat app, which supports augmented reality technology by adding visual effects on faces, at least at this moment, is influenced by wearing glasses and bad illumination conditions, see Figure 1.

Thus, the SoF dataset is considered a challenging dataset that can be used to evaluate face detection, recognition, and classification techniques. The synthetic images, which were generated after adding some image filters, are categorized into three levels: easy, medium, and hard. Each level differs from the others in the way of applying the three image filters. The first level is called easy, because the image filters were applied in a light manner. The second level (medium) is harder than the first one, and so on.

Figure 1: Screenshots of Snapchat app (version 9.39.5.0). The Snapchat's face detector successes in the first two images of each row; however, the challenging conditions in the last image of each row foil it.

The original set of images consists of two parts. The first part contains unconstrained frontal and near-frontal images of persons who wear glasses as an essential occlusion in the images. The second part contains face images that were captured under severe illumination conditions in a controlled environment.

2. General information

The SoF dataset was assembled to support testing and evaluation of face detection, recognition, and classification algorithms using standardized tests and procedures. The first version of the dataset was collected in April 2015 by capturing 242 images for 14 subjects who wear a set of eyeglasses under a controlled environment. This set was updated by capturing 92 images for 15 students from the Egyptian E-Learning University (EELU), Egypt. After that, many volunteers participated in by sharing their photos to build up the first part of the dataset. The images were captured in different countries, such as Egypt, Canada, France, Germany, India, Japan, Kuwait, Malaysia, Taiwan, United Arab Emirates, and USA. The last image in this part was captured on October 2016. The second part of the dataset was filmed in September 2016 in the Multimedia laboratory, Assiut University, Egypt.

3. Technical information

As aforementioned, the original set of images of the SoF comes in two parts. The first part contains 757 frontal and near-frontal (640 x 480 pixels) images for 106 different persons whose head orientation approximately $\pm 35^{\circ}$ in yaw, pitch, and roll. Many subjects participated in the first part by recent photos and old photos that were captured for many years ago. Some images (242 images) in the first part were captured in a systematic way (i.e. same facial expressions in

Figure 2: Samples of the second part of the SoF dataset

the same environment). The rest of the images are unconstrained images that were collected from many volunteers. Both indoor and outdoor lighting conditions are included in this part. The second part is directed to present a challenging set of images (1,905 original images) that were captured under acute lighting conditions. The idea was inspired by the early version of the Light
Stage system [4], where a wheel-whirled lamp was used as the only lighting source in the Multimedia lab, Assiut University. The subjects (12 persons) were filmed under this lamp that was located in arbitrary locations to emit light rays in random directions. The video was converted into a sequence of (640 x 480 pixels) frames which were filtered manually to pick frames that differ from previously existing frames. The differences are in the lighting conditions or the facial expressions, see Figure 2.
For each image of the original set, there are 15 images generated by the synthetic occlusions (6 extra images for each original one) and the image filters (9 extra images for each original one). We categorize the generated image into three levels: easy, medium, and hard. To do that, we adopt the Viola-Jones algorithm [5] as a predominant face detector. We try many values, in an incremental manner, for each synthetic occlusion and image filter; followed by testing the face detector. We assume that the face is successfully detected only if one of the bounding boxes has an Intersection-over-Union (IoU) ratio overlapping with the ground truth annotation above or equal to 50%. Hence, we pick the appropriate values for each category, see Figure 3.

3.1 Synthetic occlusions

Besides occluding the face using natural occlusions (glasses, scarves, head scarves), we have used the handcrafted metadata to generate two extra synthetic occlusions which are nose and mouth occlusions. For each image, we generate 3 versions of nose occlusions and another 3 versions of mouth occlusions by changing the thickness of the obstructive white block, see Figure 4. The choice of the white color is based on the study presented by Michael J. Wilber *et al.* [3] who showed that the white color acts as a strong distractor more than black color on the Facebook's face detector.

3.2 Image filters

For each original image, there are three main distractors added which are: noise, blur, and posterization.

3.2.1 Gaussian noise

For each image in the original set, we add a statistical noise that has Gaussian-distributed values that are given by

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the mean and σ is the standard deviation of the noise. We deal with the values of σ to generate three levels of difficulties (easy $\sigma = 0.01$, medium $\sigma = 0.05$, and hard $\sigma = 0.11$), see Figure 5.

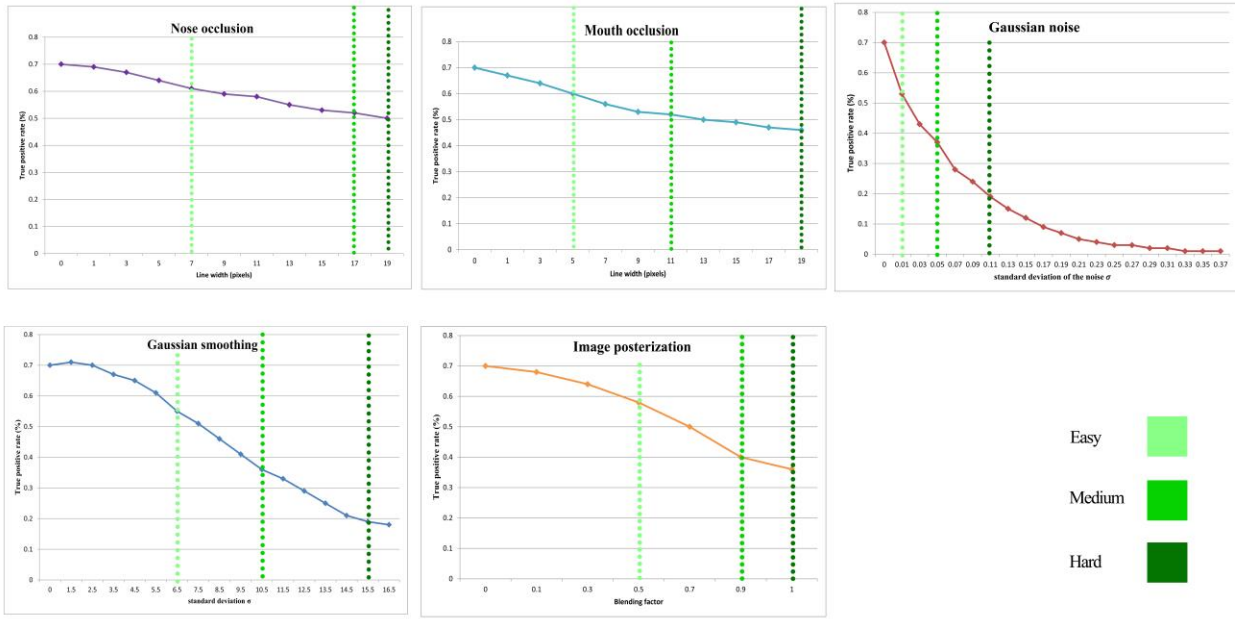

Figure 3: The performance of the Viola-Jones algorithm against the synthetic filters and occlusions. Based on the true positive rate of the Viola-Jones algorithm, the appropriate values of the synthetic effects were picked to generate the three levels of difficulties.

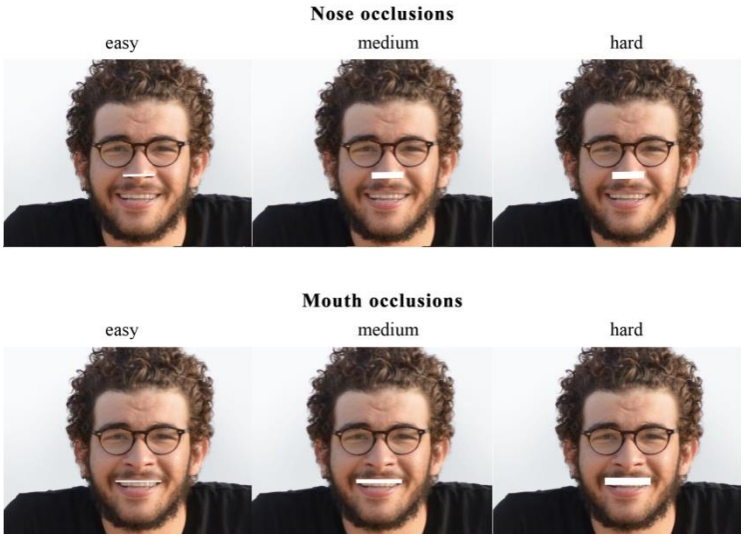

Figure 4: Examples of synthetic occlusions

Figure 5: Samples of Gaussian noise effect

Figure 6: Samples of Gaussian smoothing effect

3.2.2 Gaussian smoothing

The same here, we utilize the Gaussian function that builds the convolution matrix which is applied to each original image. By changing the value of the standard deviation of the Gaussian distribution, we get three level of cruelties (easy $\sigma = 6.5$, medium $\sigma = 10.5$, and hard $\sigma = 15.5$), see Figure 6.

3.2.3 Image posterization using fuzzy logic

We use a very simple and fast image posterization filter to generate a new hindrance that adds a new level of difficulty. For generating a posterized image, two main stages are performed. The first stage is to remove small details of the given image. For this purpose, bilateral filter [6] is used for preserving the edges in the filtered image. After preparing the given image, fuzzy logic is used for classifying each pixel into one of three categories which are: bright, gray, or dark [7]. For each pixel in each channel of the color model of the image, the fuzzification process is performed and the membership of the three categories is calculated. The dark membership function is represented using the following sigma function:

$$v_{dr}(p) = \begin{cases} 1 - \dfrac{a_{dr} - p}{b_{dr}} & if \quad a_{dr} \le p \le a_{dr} + b_{dr} \\ 1 - \dfrac{(p - a_{dr})}{c_{dr}} & if \quad p < a_{dr} \\ 0 & otherwise. \end{cases}$$

Gray membership function is represented using the following triangular function:

$$v_g(p) = \begin{cases} 1 - \dfrac{a_g - p}{b_g} & if \quad a_g - b_g \le p < a_g \\ 1 - \dfrac{(p - a_g)}{b_g} & if \quad a_g \le p \le a_g + b_g \\ 0 & otherwise. \end{cases}$$

Bright membership function is represented using the following sigma function:

$$v_{br}(p) = \begin{cases} 1 - \dfrac{a_{br} - p}{b_{br}} & if \;\; a_{br} - b_{br} \leq p \leq a_{br} \\ 1 - \dfrac{(p - a_{dr})}{c_{dr}} & if \;\; a_{br} < p \\ 0 & otherwise, \end{cases}$$

where a_K, b_K, and c_K are the parameters of the membership functions. $v_K(p)$ represents the membership's degree of the intensity of the current pixel p, $K = \in \{dr, g, br\}$, dr, g, and br denote dark, gray, and bright linguistics, respectively.

Defuzzification process is performed to determine the crisp output of the current pixel. The center of gravity is calculated by:

$$v_O = \frac{\sum_{v=1}^{N} vQ(v)}{\sum_{v=1}^{N} Q(v)},$$

where Q denotes the fuzzy output, v is the output variable, and N is the possible values of Q. In our case, the output of the membership functions is a constant of the pixel, so the previous equation can be represented as:

$$v_O = \frac{v_{dr}(p)v_{dr} + v_g(p)v_g + v_{br}(p)v_{br}}{v_{dr}(p) + v_g(p) + v_{br}(p)},$$

By getting the final value, the final crisp output is given by finding the minimum between v_O and the output variables of the membership functions:

$$min(|v_O - v_{dr}|, |v_O - v_g|, |v_O - v_{br}|).$$

The minimum value refers to the category of the pixel, whether dark, gray, or bright. By applying the following rules, the new intensity of the pixel is found:

IF p is dark, THEN make it v_{dr}

IF p is gray, THEN make it v_g

IF p is bright, THEN make it v_{br}

Eventually, the three categories of difficulties are generated based on the blending factor with the posterized image as shown in the following equations:

$$easy = (0.7 \times I) + (0.5 \times O)$$
$$medium = (0.1 \times I) + (0.9 \times O)$$
$$hard = O$$

Where, I is the input original image and O is the output image of the posterization process, see Figure 7.

You can download the technical report and source code of the “Image posterization Using Fuzzy Logic” from the following links:

Figure 7: Samples of the image posterization effect

Figure 8: The effect of SSR on the badly illuminated faces.

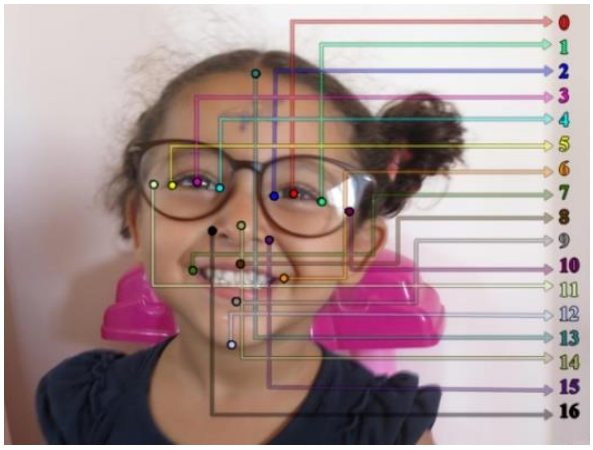

Figure 9: The facial landmarks of each subject

Technical report: here
Source code: here

4. Ground truth information

The metadata describes each subject from different aspects for many research areas, such as face detection, gender classification, and facial feature extraction. For each image, the metadata describes the subject ID (which is the first three letters of his/her first name and the first letter of his/her last name), 17 facial landmarks, face and glasses rectangles, gender and age labels, year that the photo was taken and facial emotion label (4 basic emotions which are normal, happy, sad/angry/disgusted, and surprised/fearful). We support the glasses/face rectangles as a quick way to access the glasses/face regions, so the glasses/face rectangles are upright rectangles, regardless of the head orientation. To get more accurate face rectangles, it is recommended to use the facial landmarks instead. There is a variable, denoted by "glassesType", which refers to the type of glasses (eyeglasses, semi-transparent sunglasses, opaque sunglasses, or others). There is a label called "illuminationQuality" that indicates whether the face is well-illuminated or captured under poor illumination conditions. The poor illumination means that there is at least one facial point, i.e. landmark, which is invisible due to the bad illumination conditions. In other words, the well-illuminated face is the face whose all non-occluded facial features are recognizable by naked-eye. To handle this issue within we were specifying the ground truth of the facial landmarks, emotional status, and face and glasses rectangles, we apply the Single-Scale Retinex algorithm

(SSR) [5], upon request, to images that have badly illuminated faces, see Figure 8. In addition, there is a set of (yes /no) variables which are: cropped, frontal, estimated points, indoor/outdoor lighting, and head scarf.

Figure 9 illustrates the 17 facial landmarks that are included for each subject, which are:

0 = right eye pupil
1 = outer corner of right eye
2 = inner corner of right eye
3 = left eye pupil
4 = inner corner of left eye
5 = outer corner of left eye
6 = right mouth corner
7 = left mouth corner
8 = center point on outer edge of upper lip
9 = center point on outer edge of lower lip
10 = right temple
11 = left temple
12 = tip of chin
13= top of forehead
14 = tip of nose
15 = outer corner of right nostril
16 = outer corner of left nostril

There are some cases should be taken into account; there are some images contain cropped faces. Consequently, the top of forehead or the tip of chin may be invisible. There are a logical variable which is called “cropped”. This variable can be used to determine whether this image contains a cropped face or not. “headScarf” label indicates whether the subject wears head scarf or not. Also, this label is set to true for 11 female images ($+ (11 \times 15)$ generated images) and 8 male images ($+ (11 \times 15)$ generated images), where the subject wear something else, such as sun hats, bucket hats, or skull caps. If the subject is a male, this logical variable refers to whether the subject wears anything on his head (e.g. cap) or not. In all images, the 17 facial landmarks are supported, but some of them may be inaccurate due to the invisibility of some facial features. To be more accurate, we provide a logical variable denoted by “estimatedPoints” that is associated with each facial landmark to indicate whether this point is an estimated or exact one[1]. The facial landmark is deemed an estimated point, if it is invisible due to bad illumination conditions, near-frontal view, cropped image, poor quality of the image, occluded facial point, glasses lens reflection, or opaque lenses.

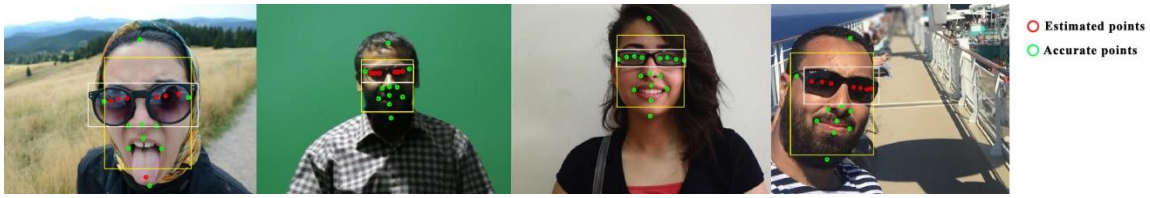

Figure 10: Examples of estimated facial landmarks, face and glasses rectangle in some images

[1]There is an inevitable human error, so the exact points may have some mistakenly shifts.

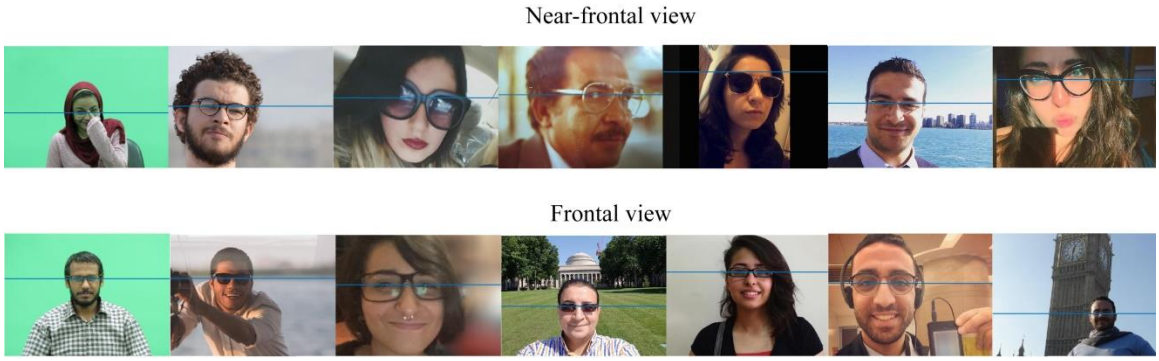

Figure 11: Examples of frontal and near-frontal images

There are 45,254 facial landmarks (36,860 accurate points and 8,394 estimated points) which were specified by two experts who worked individually. The reported landmarks are the average of the specified facial points. The "estimatedPoints" is set to true in the case of at least one expert expressed the point as an estimated point. Figure 10 illustrates examples of the estimated points in some images. Needless to say, many of the facial landmarks for the synthetic versions are estimated points based on the synthetic occlusion type (i.e. mouth occlusion images have mouth facial estimated points, and the same for noise occlusion images).

The head pose (view) was defined relative to the camera. There are two categories of the view, which are: frontal and near-frontal. We accepted pose angle ≈ 0 in the three axes as a frontal view. There is a tolerance in the tilt angle (especially the pitch-angle values). Thereby, we accepted subtle tilts as a frontal view. For all uncertainty cases, we dragged a vertical line between two reference points to estimate the roll angle and use ears visibility to estimate the yaw angle. Figure 11 shows examples of frontal and near-frontal images extracted from the dataset.

The formatting of the metadata is very clear. Table 1 describes more details of the fields of the metadata. Furthermore, the filenames of images can be used to obtain the metadata except some information, such as the facial landmarks and the face/glasses rectangles. The format of the filenames is A_B_C_D_E_F_G_H_I_J_K_L_M.jpg, where:

A= subject ID (xxxx).

B= image sequence number (xxxxx).

C= gender (x); (m) for male images and (f) for female images.

D= age (xx).

E= lighting (x); (i) for indoor lighting and (o) for outdoor lighting.

F= view (xx); (fr) for frontal images and (nf) for near-frontal images.

G= cropped (xx); (cr) or not (nc).

H= facial emotion (xx); (no) for normal, (hp) for happy, (sd) for sad/angry/disgusted, and (sr) for surprised/fearful.

I= year (xxxx).

J= part (x); (1) for part 1 of the dataset and (2) for part 2 of the dataset.

K= occlusion (xx); (e0) for eye occlusion, which is the common in all images, (en) for eye and nose occlusion, and (em) for eye and mouth occlusion.

L= image filters (xx); (nl) for normal images without any filters, (Gn) for images with Gaussian noise, (Gs) for images with Gaussian smooth, and (Ps) for posterized images.

M= level of difficulty (x); (o) for original images, (e) for the easy level, (m) for the medium level, and (h) for the hard level of difficulty.

For example, the subject, whose ID is MahA and his image, which has a sequence number 00986 in the original set, named "MahA_00986_m_25_i_fr_nc_hp_2013_1_e0_nl_o.jpg". This photo was taken in indoor lighting from a frontal view of MahA. This image contains the whole of his face, which shows a happy emotional status. This photo was taken in 2013 when MahA was 25 years old. There is no synthetic occlusion or image filter added in this image, consequently, this is an original image.

5. 5-fold cross validation

As a benchmark dataset for comparison, we suggest reporting performance as 5-fold cross validation.

5.1 Gender classification, face recognition and emotion recognition

Each fold is represented by a txt file that contains the filenames of all images in this fold. Note that, there are two groups of folds, the first one for the original dataset called fold1Original.txt, fold2Original.txt, … etc. The second group contains the folds of the whole dataset called fold1.txt, fold2.txt, … etc. For gender classification problem, all folds contain the same distribution of males and females. For emotion recognition, all folds contain the same distribution of facial emotions.

5.2 Face, eyeglasses and facial landmark detection

For face, eyeglasses and facial landmark detection, we suggest use the whole dataset images to test your model. For training and testing, you can use gender classification 5-fold validation.

6. Statistics

Figure 13 shows some statistical information of the SoF dataset.

7. Acknowledgement

There are 343 images of the dataset were captured by Ali Hussien, Ebram K. William, Mostafa Korashy, so thanks for their effort. Thanks for the administrators of Faculty of Computers and Information who supported this work. Eventually, thanks for all volunteers who trusted us with their photos to accomplish this work.

Table 1 Describtion of the metadata (Matlab)

Field name	Description	Values
userId	Represent the ID of the subject	The first three letters of the subject's first name and the first letter of his/her last name. Example: MahA, where Mah are the first three letters of the subject's first name and A is the first letter of his last name.
imageSequence	Refer to the sequence of the current image in the set of the original images.	Five decimal places number. Example: 0001 is the first image in the original images of the dataset.
gender	Represent the subject's gender	A single character; 'm' for males and 'f' for females.
age	specify the subject's age when the photo was taken	Two decimal places. Example: 03 means the subject was 3 years old when the photo was taken.
lighting	Specify whether the photo was captured in an indoor or outdoor lighting	A single character; 'i' for indoor lighting and 'o' for outdoor lighting.
view	indicate whether the face was captured in the frontal view or in the near-frontal view ($\pm 35°$) in yaw, pitch, and roll	Two characters; 'fr' for frontal view and 'nf' for near-frontal view.
cropped	Indicate whether the face is cropped or not.	Logical value; 0 for no-cropping faces and 1 for cropped faces.
facialEmotion	Refer to the facial emotion of the subject. There are 4 basic facial emotions which are (normal, happy, sad/angry/disgusted, and surprised/fearful).	A single decimal place number which refers to the index of the facial emotion. Example: facialEmotion=2 that means the subject is happy.
year	The year when the photo was taken	Four decimal places numbers
part	There are two main parts of the dataset. This field indicates which part this image belongs to.	A single character; '1' for the first part and '2' for the second part.
glassesType	Refer to the type of the glasses. There are four categories here: eyeglasses, semi-transparent sunglasses, opaque sunglasses, and others for non-glasses occlusions, such as masks.	A single decimal place number that refers to the index of the glassesType. Example: 2 means that the subject wears a semi-transparent sunglasses.
headscarf	Indicate whether the subject wears something on her/his head or not.	A logical variable; 1 for subjects who wears headscarf and 0 for not.
facialPoints	The 17 facial landmarks of each subject.	34-element vector of float numbers that refer to (x,y) of the 17 facial landmarks. Example: [373.75, 153.26, 387.53, 154.18, 360.90, ….] the right eye pupil landmark is (373.75, 153.26) and so on. See Figure 9.
estimatedPoints	The 17 logical (yes/no) values that indicate whether each of the facial landmarks is an estimated or accurate point.	17-element vector of logical variables that associated with the facial landmarks. Example: [0,1,…] that means the right eye pupil point is an accurate point, but the point of the outer corner of right eye is an estimated point, and so on.
faceRect	The ROI of the face	Four-element vector of float numbers of the face rectangle. Example: [300.2991, 129.3953 , 105.5954 , 112.0230] that means that the x= 300.2991, y= 129.3953, width= 105.5954, and height= 112.0230. Where, (x,y) is the left upper corner of the rectangle.
glassesRect	The ROI of the glasses or the masks	The same formatting of the faceRect
illuminationQuality	Refer to whether the face is well-illuminated or badly illuminated.	A logical variable; 1 for well-illuminated faces and 0 for badly illuminated or dark faces.
filename	The first 4 unique parts of the current image	userId_imageSequence_gender_age_*; since each image has several version of synthetic images, the fileName field comes in this format to make easy reach of other versions of the same image that have the same metadata.

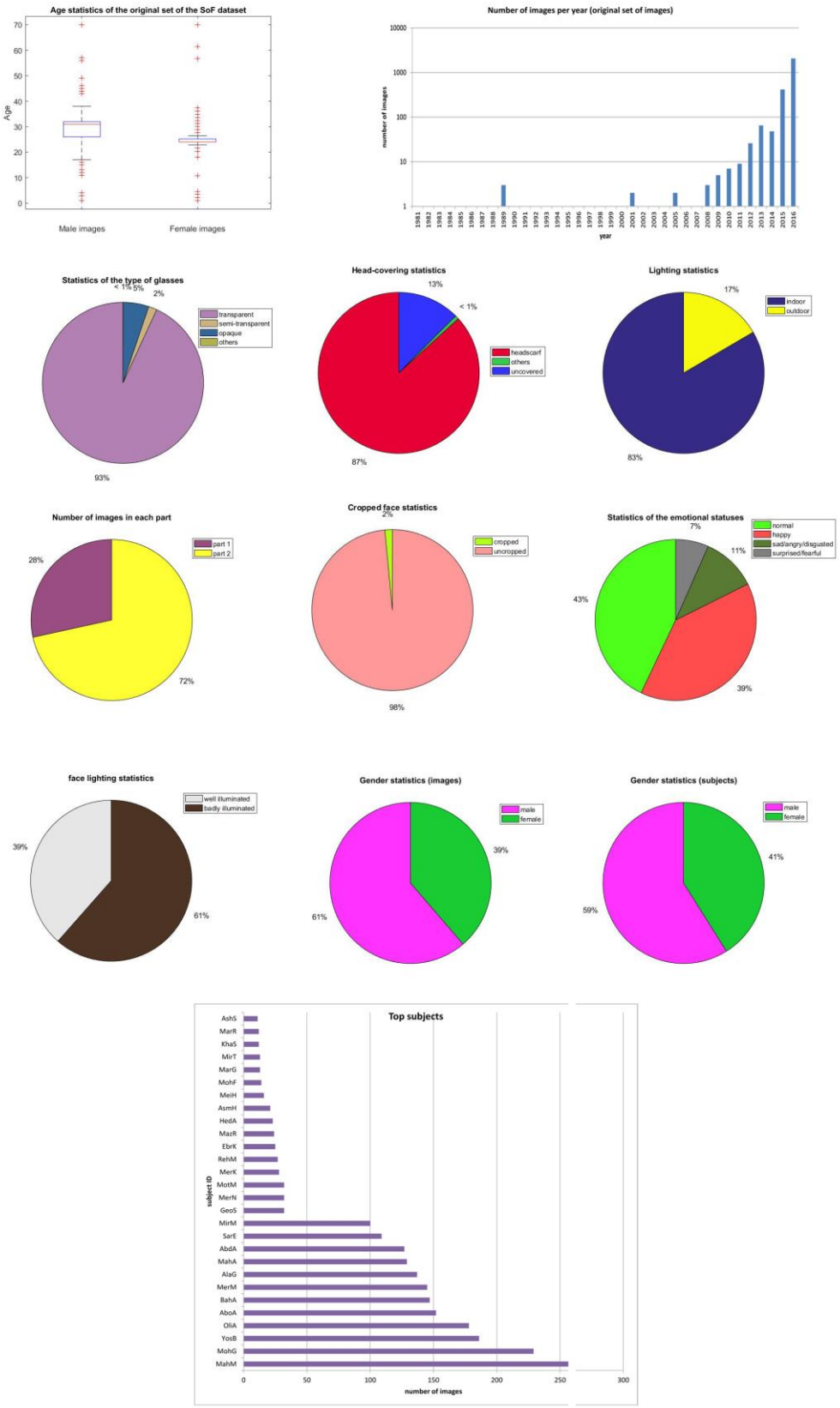

Figure 13: Statistical information of the SoF dataset

8. Bibliography

[1] Afifi, Mahmoud, and Abdelrahman Abdelhamed. "AFIF4: Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces." arXiv preprint arXiv:1706.04277, 2017.

[2] Han, Hu, et al. "A comparative study on illumination preprocessing in face recognition." Pattern Recognition 46.6, 1691-1699, 2013.

[3] Wilber, Michael J., Vitaly Shmatikov, and Serge Belongie. "Can we still avoid automatic face detection?." 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016.

[4] Debevec, Paul, et al. "Acquiring the reflectance field of a human face." Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 2000.

[5] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57.2 (2004): 137-154.

[6] Paris, Sylvain, et al. "A gentle introduction to bilateral filtering and its applications." ACM SIGGRAPH 2007 courses. ACM, 2007.

[7] Gonzalez, Rafael C., and Richard E. Woods. "Digital image processing.", 2008.

[8] D.J. Jobson, Z. Rahman, G.A. Woodell. Properties and performance of a center/surround retinex. IEEE Transactions on Image Processing, Vol. 6, No. 3, str. 451–462, 1997.