Title: Image conditions for machine-based face recognition of juvenile faces

Dr. Ching Yiu Jessica Liu BSc (Hons), MSc, AFHEA, PhD
Liverpool John Moores University, Liverpool School of Art and Design
IC1 Liverpool Science Park, 131 Mount Pleasant, Liverpool, Merseyside, L3 5TF
t: 0151 482 9605 (Lab) e: C.Y.Liu@ljmu.ac.uk

Prof. Caroline Wilkinson BSc, MSc., PhD, FRSE, FRAI
Liverpool John Moores University, Art and Design Academy
2 Duckinfield Street, Liverpool, Merseyside, L3 5RD

Acknowledgements
I would like to thank my supervisory team including Dr. Martin Hanneghan and Dr. Sud Sudirman from the Department of Computer Science of Liverpool John Moores University. Face Lab and the Mathematics and statistic department at the Liverpool John Moores University.

Declaration of interest
This work was supported by the Faculty of Engineering and Technology at Liverpool John Moores University

Key words: Facial identification; juvenile age progression; face recognition

1. Introduction

The Child Victim Identification Program (CVIP), a component of the FBI's Cyber Crimes Program, is designed to help identify child victims in abusive images [1]. Since the program began in 2002, more than 267 million images and videos had been reviewed, and law enforcement has identified more than 15,800 child victims [2].

Most recovery methods of missing children focus on publicising the identity of the missing child in hope for the public to report and contact authorities [3]. With the vast number of children going missing along with the increasing displacements of populations and an overload of media information, human recognition may not be an effective identification method. This research therefore focuses on the ability of machine-based methods for the recognition of children's faces over time.

The National Institute of Standards and Technology (NIST) reported that the false negative and false positive rates for Facial Recognition Systems (FRS) in juvenile were much higher than for adults. With a high false positive rate across all algorithms, they found a progressive trend in the decrease of false identification with increasing age and concluded that it was difficult to discriminate younger children [4]. Ling and colleagues [5] designed a face verification algorithm and tested faces across different ages for children and adults. Their study found that verification was much harder for children in comparison to adult faces and it was extremely difficult to verify the identity of children between 0-8 years of age. This is unsurprising, as the algorithm considers the face to be a universal, distinctive, permanent and collectable biometric [6]. However, as children's faces change rapidly over short periods of time, facial recognition in children cannot be classed as a reliable biometric method, as facial characteristics are invariant. Some researchers consider a child's face as a soft biometric [7] and it was defined as having *"characteristics that provide some information about the individual but lacks the distinctiveness and permanence to identify an individual uniquely and reliably"* [8]. Humans often identify each other with soft biometric traits, for example, height, weight, gender, eye colour, ethnicity etc. [8–10]. Ferguson [11] suggested that the manual facial comparison of juvenile faces is error prone. If we were to consider a child's face as a soft biometric, we would need to consider the human ability to recognise children's faces even when they are years apart. How good are facial recognition systems in the identification of children across time? Perhaps

identification with a focus on stable features and facial markings, such as moles, should be evaluated further [12].

To model and predict the possible changes to an ageing face, age progression methods change the shape, colour and texture of a facial image while retaining the identity of the individual [13]. Age progression is often separated into juvenile and adult [14], and this study focuses on age progression for juvenile faces. Age progression is challenging for individuals younger than 3 years of age, as facial characteristics are underdeveloped at this stage in the growth pattern [14]. Age progression is more accurate with images of older children and accuracy is also affected by the quality of the reference photographs [14].
Current research techniques include manual or machine-based digital image processing and sometimes drawings by artists [14]. Previous literature has described machine-based age progression methods as automated or computerised methods. The level of automation of age progression is still in its infancy and requires a high level of human influence. Different research groups have developed methods to automate age progression [7,15–22].

NCMEC in the USA updates the age-progression image every 2 years before age 18 years, and every 5 years after age 18 years. These images are used to generate further investigative leads [23]. Different approaches to age progression have been attempted, but "currently there is no automatic age progression software that can guarantee any degree of accuracy" [23]. Although the literature has explored the human recognition rate using age progression images [3,24–27], but few has reported the testing of automated recognition rates using the age progression and veridical (target) image.

Computer scientists have explored how image quality can affect facial recognition systems (FRS) [28,29], but when dealing with age progression images, researchers have suggested that inaccurate information can often have a negative effect on human recognition [30–32]. Psychologists have tested how different conditions can affect face recognition, conditions such as colour, illumination, low resolution [33]. Gaussian blurring in face recognition suggested that faces are recognisable even when they are blurred or pixelated [34–36]. External features such as hair can also have an effect on human recognition [24,37,38]. With the conditions above, this study aims to explore how an FRS is able to handle 'resemblance' images at different conditions known to benefit human recognition.

To investigate how different image manipulations can affect the recognition score, this study uses Microsoft face API, Amazon Rekognition and Face++, all Cloud-based AI services. These are commercially available black-box systems to represent machine-based methods, and any conclusions can only inform performance of a black-box system. Even without knowing fully how the algorithm operates, black-box systems have been used in previous research [39–43]. These commercially available software mainly uses deep learning [41] and the accuracy can vary depending on the input image [42,44–49].

The age progression images were subjected to four different conditions using photo-editing software (Adobe Photoshop CS6): black and white, cropped, blur and resolution reduction.

2. Methodology

Based on the 24 FG-NET subjects (14F, 13M), 42 original images (22F, 20M) were selected, with most having 2 progressions to different ages, 80 digital manual age progressions were generated (method not included in this paper), and 83 comparisons were made. The FG-NET database was released in 2004 and it has since been used in research related to face recognition, age estimation and age progression. Each age progression was subjected to four different conditions (black and white, cropped, Gaussian blur and resolution reduction). using photo-editing software (Adobe Photoshop CS6), see Figure 1.

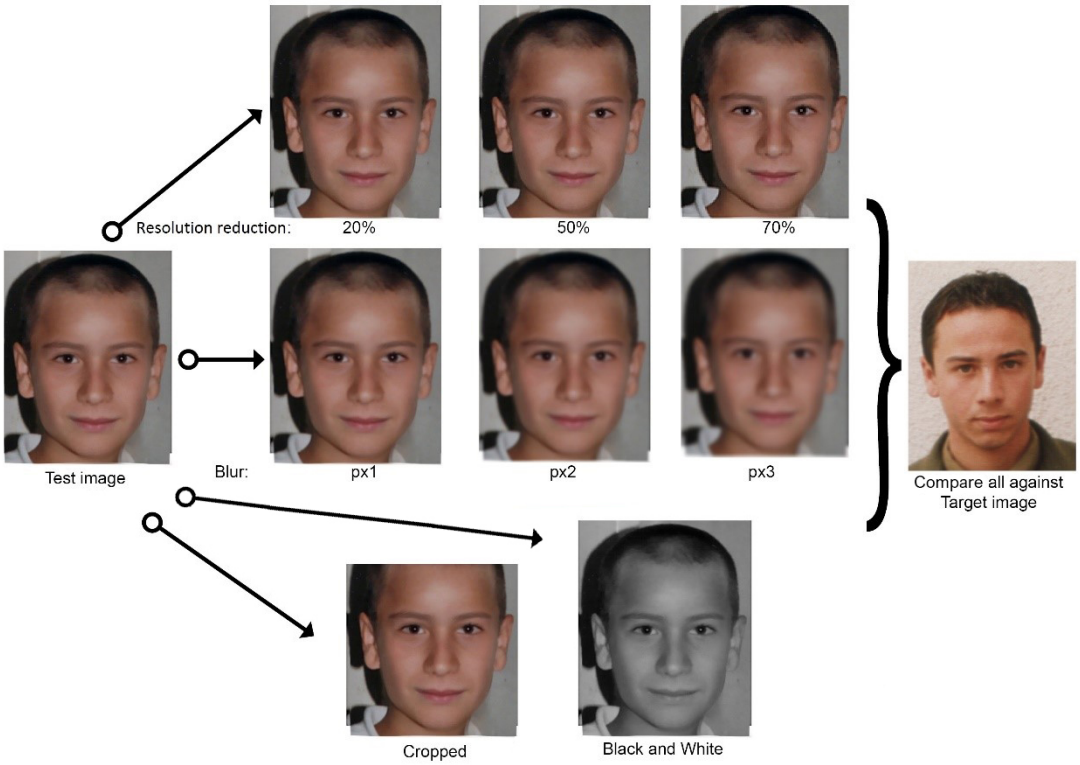


Figure 1: Workflow for image testing (Subject from FG-NET) [Print in colour]

Some original images were black and white, these images were excluded from the black and white comparison (n=8F;4M). Some original images did not show the hairline and were excluded from the cropped comparison (n=1F). The image quality and size varied and therefore the image resolution was not consistent. Rather than standardising all images to the same pixels, the original pixels of the image was documented and be reduced to 20%, 50% and 70%. The setting with a higher confidence score represented the performance of this condition. It should be noted that the resolution was reduced via the image size and not pixelated using the Photoshop filter. Similar to resolution, the quality of the images varied. Therefore, the setting with a higher confidence score represented the performance of this condition. The Gaussian blur filter was applied as a function in Adobe Photoshop CS6. The images were blurred with 1 pixels, 3 pixels and 5 pixels.

The original (out-dated) image and the manual age progression images were compared to the 'veridical' target image (image of the individual at the target age) using three off-the-shelf facial recognition algorithms (Microsoft Face API, Amazon Rekognition, and Face ++).

All three algorithms gave a confidence score between two facial images in percentages: 0% being dissimilar faces and 100% being similar faces. This would provide an objective assessment of the 'likeness' between two images.

The confidence score generated by the three algorithms between the manual age progression and the target image(s) was compared to the score between the original and the target image(s). Age progression was compared with different conditions, and image variability was explored.

3. Results

The original image, age progression image, and progressions with conditions were all compared to the veridical/target image using the three algorithms (Microsoft Face API, Amazon Rekognition, and Face ++).

A between-subjects univariate analysis compared the effect of the software on subject sex and image type. Sex (M,F) showed no significance [$F(1,486) = 0.349$, $p=0.555$] (Figure 2), and Software [$F(2, 486) = 263.744$, $p<0.000$] (Figure 3) showed a significant difference.
The original and age progression image type was statistically different [$F(1, 486) = 24.274$, $P<0.000$], however, when separated into the different softwares (Figure 3), the image types were not statistically different in Microsoft [$t(164)=0.111$, $p=0.912$], but significantly different in Amazon [$t(164)=5.387$, $p<0.000$] and Face++ [$t(164)=3.275$, $p=0.001$].

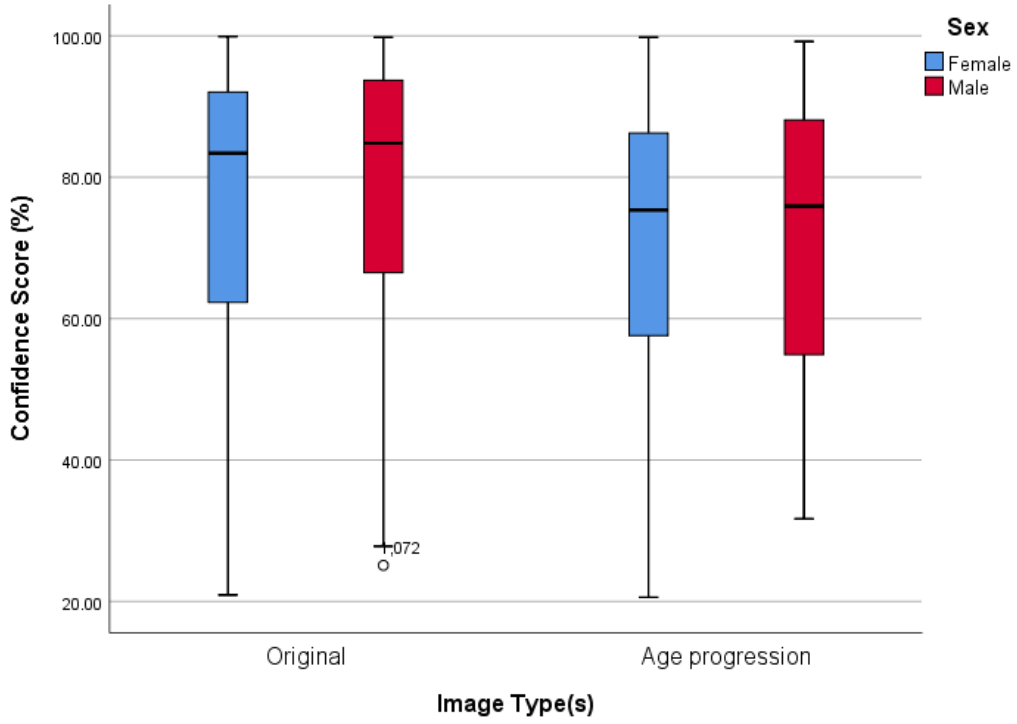


Figure 2: The confidence score (%) of image type(s) between sex when compared to the target images [Print in colour]

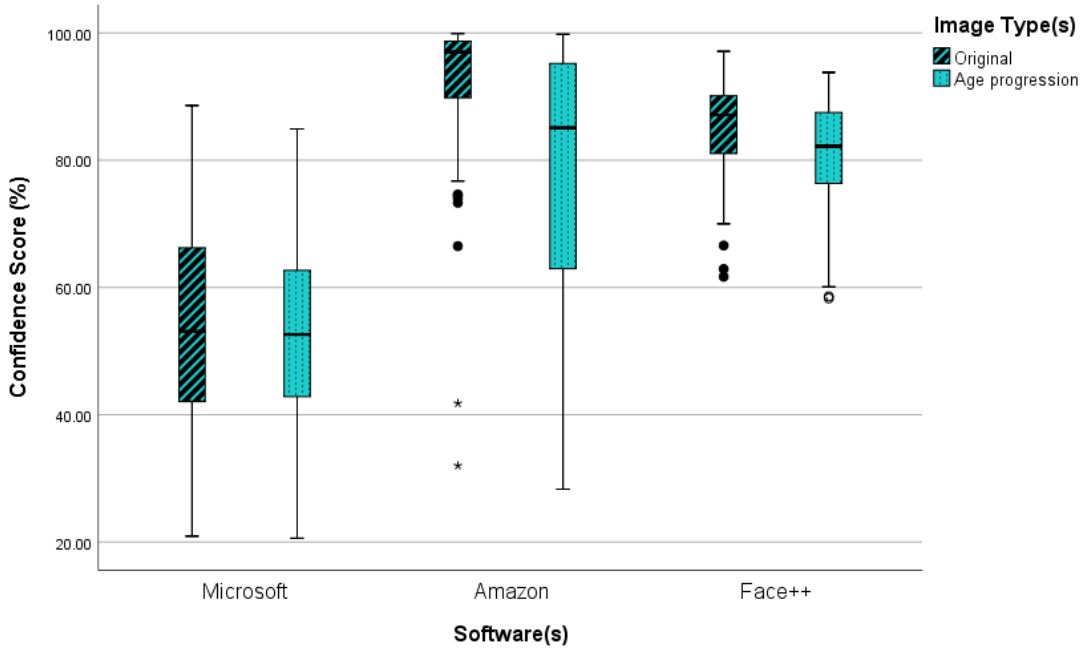


Figure 3: The confidence score (%) of software(s) between the image types when compared to the target images [Print in colour]

When considering the better-preformed image type within each image set, the highest confidence score within each image set was recorded. If the score was the same between the image types, both were also recorded as the 'highest' (Figure 4).

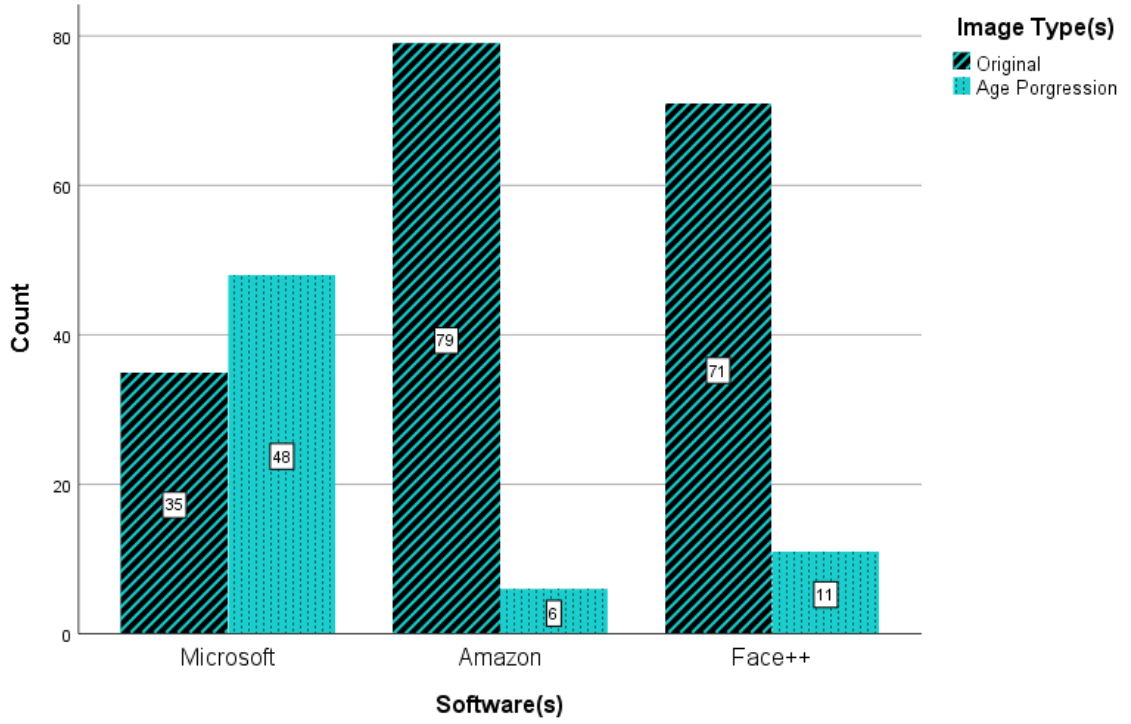


Figure 4: Highest confidence score count for original images and age progression images when compared to the target images [Print in colour]

Overall, 74% (185/250) of the original images achieved a higher confidence score when compared to the age progression images. The chi-square test found a significant difference between the highest confidence score for the original images and the age progressions [$\chi^2(1)=57.60$, $p<0.000$], and no significant difference between the highest confidence scores between the male and female subjects [$\chi^2(1)=0.299$, $p=0.585$].

Figure 4 suggests the original images consistently achieved the highest confidence score with Amazon and Face ++, however, with no significant difference, the age progression images performed slightly better with Microsoft.

Overall, Figure 5 andFigure 6 suggest that as age gap increases, confidence score decreases for all types of softwares and image types. This suggest between the original image and the target image, the face becomes more dissimilar as the age gap increases. Figure 5 suggests the original

images were more similar to the target images when compared to the age progressions. A simple linear regression indicated that the model was a significant predictor [$F(2,495)=51.282$, $p>0.000$] for the difference in image type ($\beta_1=-6.071$, $p>0.000$) and age gap ($\beta_1=-2.416$, $p>0.000$). This suggests the overall difference between the image types was significant with an increasing age gap. However, the effect size was low ($R^2=0.172$).

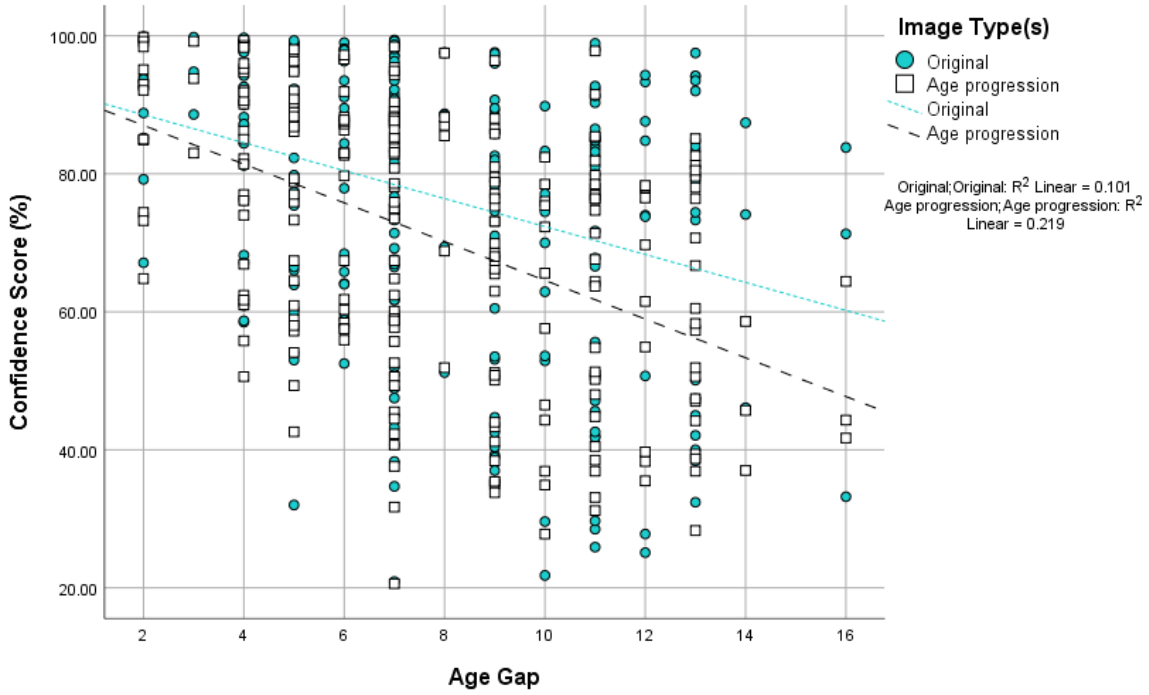


Figure 5: Similarity Score for age gap (Years) related to image types when compared to the target images [Print in colour]

A between-subjects univariate analysis was conducted to compare the effect of the age gap on software, subject sex and image type. Overall, softwares are statistically different to each other [$F(2,414)=147.48$, $p<0.000$]. Post hoc comparisons using the Tukey HSD test indicated that the mean score for Microsoft ($M = 53.63$, $SD = 15.05$) was significantly different to the other softwares. However, Amazon ($M = 84.79$, $SD = 18.03$) was not significantly different to Face ++ ($M = 82.92$, $SD = 8.13$). This indicates age gap affects Microsoft more than Amazon and Face ++. Figure 3 Figure 6 also indicates that Amazon and Face ++ has significant higher confidence scores when compared to Microsoft, and Face++ is more resistant to the effect of age gap when compared to the other two softwares.

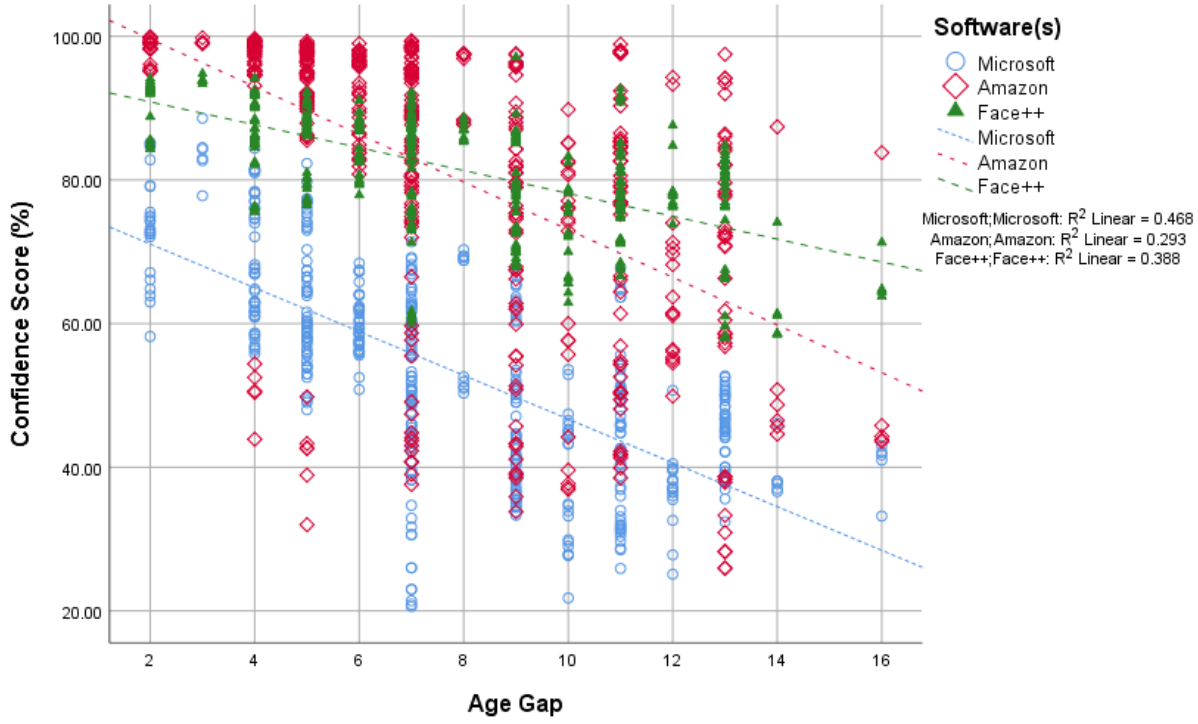


Figure 6: Similarity score for age gap (Years) related to the software(s) when compared to the target images [Print in colour]

A simple linear regression indicated that the model was a significant predictor [$F(2,495)=136.29$, $p>0.000$] for the difference in software ($\beta_1=14.646$, $p>0.000$), but all three softwares together were not able to predict the confidence score with the increasing original age ($\beta_1=14.65$, $p=0.054$), this suggest the influence of original age is weak. However, when looking at Amazon and Face ++ alone [$F(2,329)=4.612$, $p=0.011$], there were no difference in software ($\beta_1=-1.865$, $p=0.221$), but the prediction for original age was significant ($\beta_1=0.996$, $p=0.006$). This suggests the original age had no influence on Microsoft, and the confidence score was higher when the original age of the individual was older on Amazon and Face++, however, the effect size was low ($R^2=0.027$).

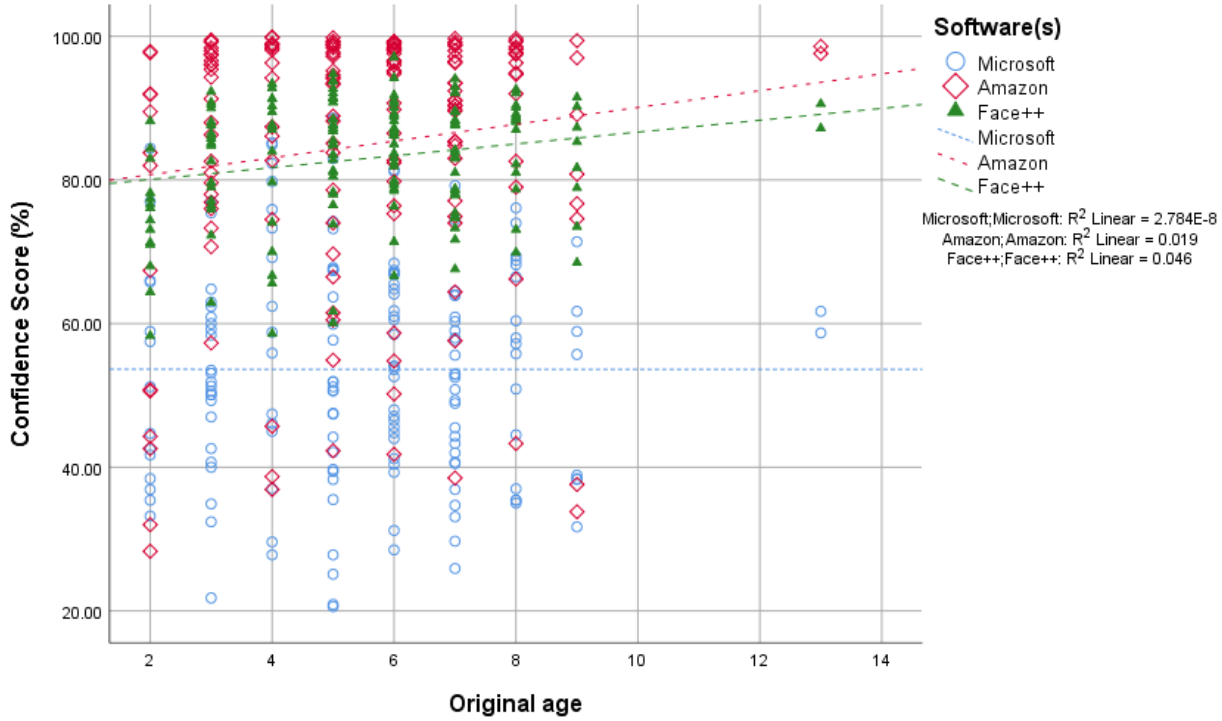


Figure 7: Similarity score for original age (Years) related to the software(s) when compared to the target images [Print in colour]

a. Image conditions

Each age progression was manipulated with conditions black and white, cropped, blur and resolution reduction. Condition(s) with the highest confidence score were recorded; if the score was the same between multiple conditions, all conditions were recorded as the 'highest'. Chi-square tests were performed and found a significant difference between the conditions [$\chi^2(4)=19.65$, $p=0.001$], with no significant difference between the subject sex [$\chi^2(1)=0.299$, $p=0.585$] or softwares [$\chi^2(2)=0.915$, $p=0.633$]. Figure 8 suggests that although the depictions remained the same, recognition can be affected by a change in condition, especially for Microsoft Face API. Amazon and Face++ are more resilient to a change in image condition.

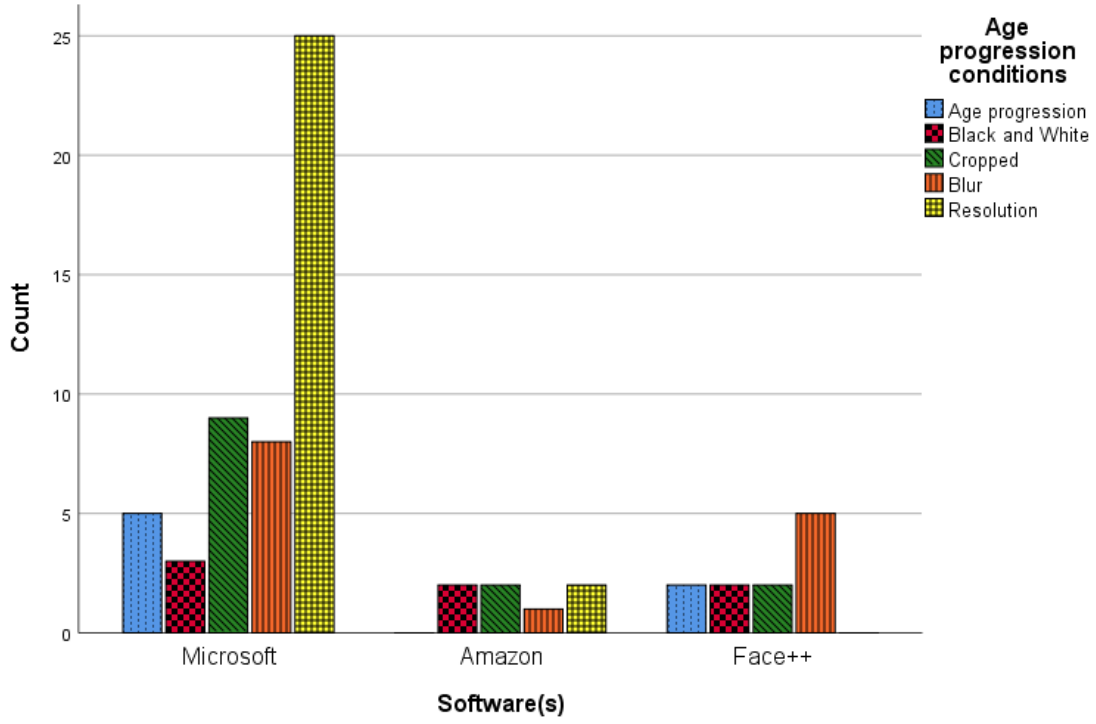


Figure 8: Highest confidence score count between the softwares for the different age progression conditions when compared to the target image [Print in colour]

To explore if a condition reduces the similarity score of an image, for each individual, confidence scores of the different conditions were compared to the original age progression image. Condition(s) with the lowest score were documented. Chi-square tests were performed and found a significant difference between the conditions [$\chi^2(4)=98.867$, $p<0.000$], with no significant difference between the subject sex [$\chi^2(1)=0.299$, $p=0.585$] or softwares [$\chi^2(2)=0.915$, $p=0.633$]. Figure 9 suggests different conditions can have a negative effect on verification, especially the black and white and cropped condition across all three softwares.

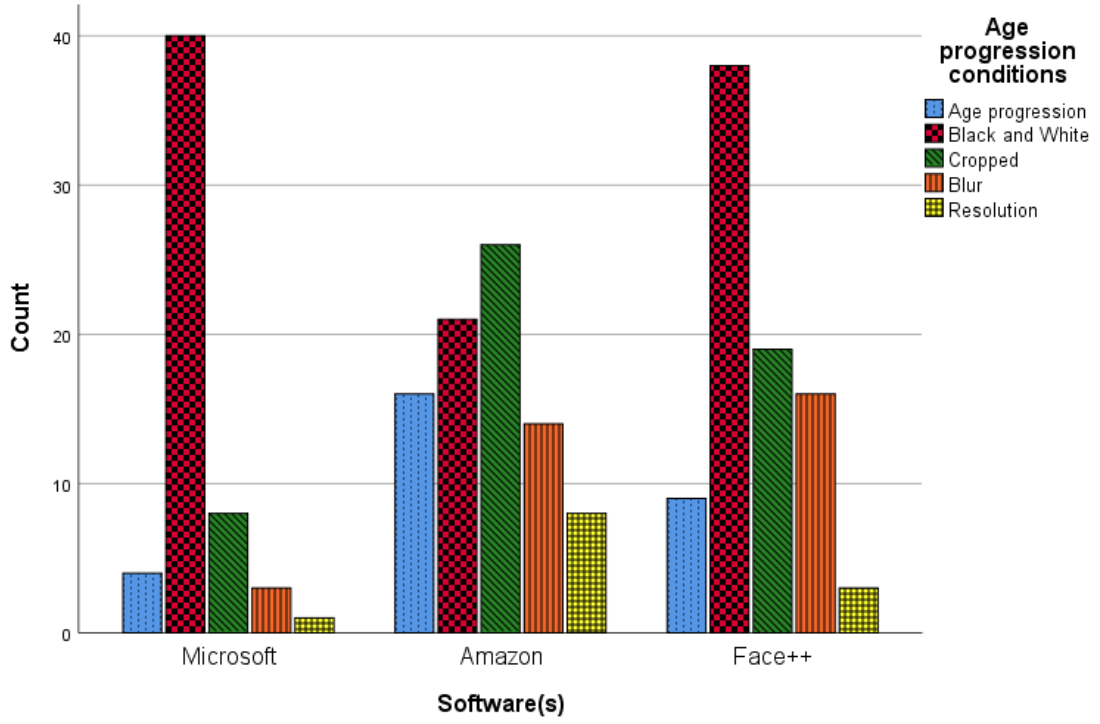


Figure 9: Lowest confidence score count between the softwares for the different age progression conditions when compared to the target image

b. Image variability

Most comparisons only had one target/veridical image for comparison with the exception of three comparisons (Set 1, 2 and 3) from the FG-NET dataset, where two target images at the same age was available.

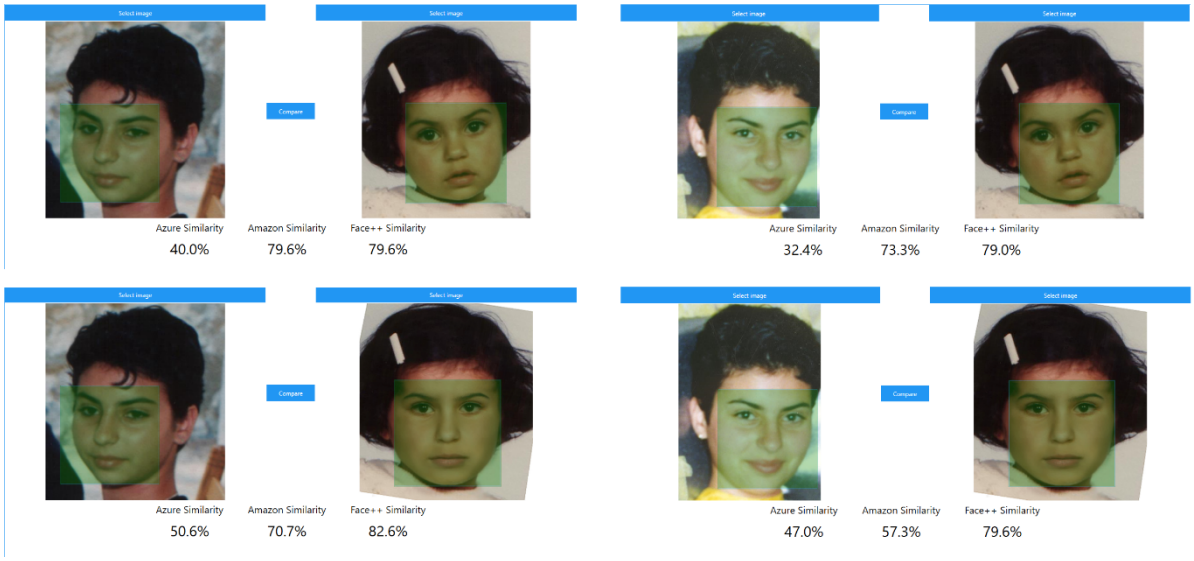


Figure 10: Image variation - Set 1 Age 03 > Target age 16A and 16B

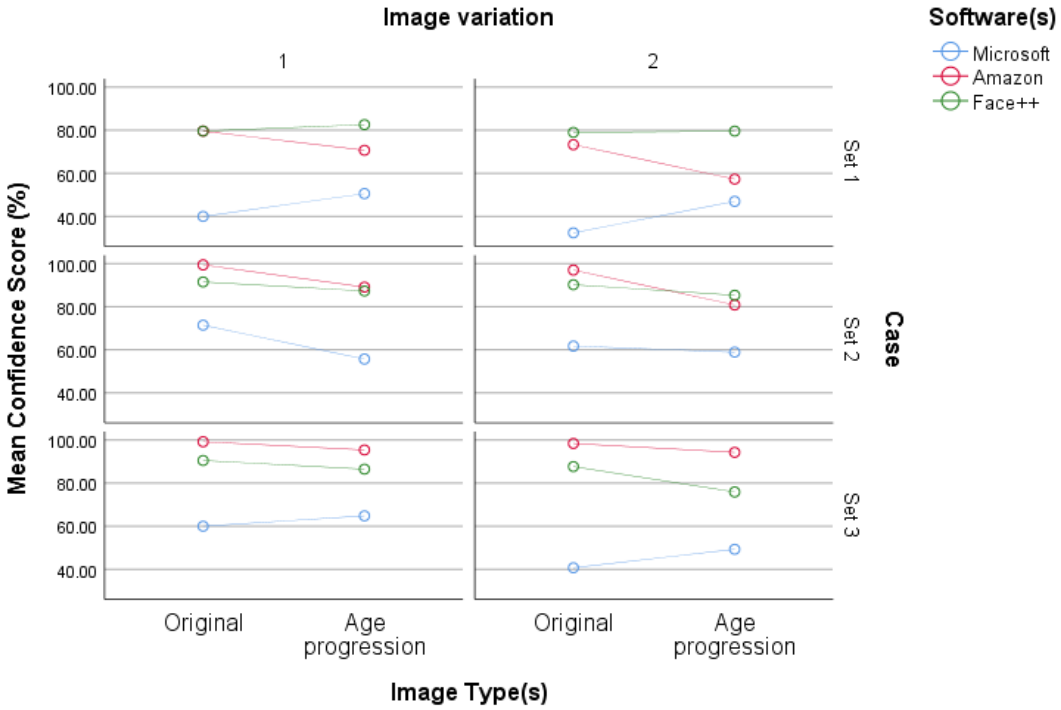


Figure 11: Image variability of different images of the same age (Years) when compared to an original image and an age progression [Print in colour]

Figure 9 and 10 showed the difference in recognition rate when an original image or an age progression was compared to different target images of the same age. This suggests that the difference in image quality can have an effect on the comparisons and could potentially lead to an identification.

To have more than one target image for comparison would be ideal, this information could potentially be available in the process of identification of indecent images of children within a database. However, in a research setting using databases such as FG-NET, the availability of images is limited.

4. Discussion

c. How facial similarities is affected by age gap

80 original images and age progression depictions from 24 subjects were compared to 83 veridical images. The results suggest recognition decreases as age gap increases. This is not suprising as the face of a child is not a permant feature and becomes more disimilar with growth related changes to the face.

This study used three different FRS to compare single images of the same individual at different ages, and facial similarities were given a score. As the age gap increases, the similarity score decrease, this suggests the probability of false positive and false negative recognition will be higher if this was used for identification. The relationship between the age gap and false identification would require further testing.

There was no clear difference between male and female subjects and this could suggest growth related changes are similar for the FRS. Farkas and Hreczko [50] indicated facial growth difference between the sexes, most noticeably in maturation age and in the periods of late accelerated growth. These changes were more significant in males from around age 10 years, including areas such as the width of the mandible, nose height, nasal tip protrusion and lip height [50]; most likely influenced by the hormonal differences during puberty. However, these changes were not reflected in the results, which suggest that even with facial changes both males and females were recognisable at a similar level with the increasing age gap.

d. How different conditions of an age progression can affect machine-based face recognition

Age progression images create a likeness of the individual and it is not possible for these images to be exactly the same as the veridical image. To explore if the similarity scores could be increased by reducing the errors of the depiction, four different conditions were applied to reduce the information in the depictions: i.e. black and white, cropped, blur and resolution reduction. The confidence score of all conditions was compared to the age progressions.

Computer scientists have tested how different conditions can affect Deep Convolutional Neural Network (DCNN) based algorithms. Dodge and Karam [28] tested four image classification algorithms and suggests that all models are sensitive to blur. Similarly, Grm et al. [29] tested face verification performance on four DCNN based FRS; their study indicated that all models were sensitive to noise, blur, missing data and brightness (Missing data, in this case, were similar to partial occlusion of the face). Both studies found that contrast and compression were of low impact [29,51]. These study designs analyse how different conditions decrease the performance of an algorithm and some studies are designed to find the threshold of certain conditions [52–55]. In order to explore if such conditions are able to improve the recognition rate of depictions, the changes in condition in this current study, such as blurring and resolution reduction, were not as extreme. The aim was not to find a threshold, but to analyse if a change in condition can improve recognition.

The performance of a face recognition algorithm can remain similar with images across different resolutions [52–55]. With resolution reduction, the current study suggested the similar score can be increased when compared to other conditions. The improved similarity score suggest conditions such as resolution reduction and blurring decrease the differences between the two images and could make the depictions more recognisable.

Image processing can have a negative effect on recognition [56]. With previous research suggesting that colour information does not have a significant effect on performance [29,51], it was surprising to discover the black and white condition performed the worse. Although concealing what is unknown could be beneficial in some circumstances [24,27,57], the cropped condition had a negative effect on recognition. However, both human and machine-based recognition suggests a positive effect for blurring and resolution reduction.

One suggestion for improvement to FRS is often related to the training database [42]: for example, increase the number of low-quality images [28] or more profile images [58] during training. Perhaps the low recognition rate is related to the lack of certain conditions within the training set; for example, the lack of black and white conditions could lead to a lower recognition rate of these images, and same for the cropped images. However, the effect of cropping (i.e. the percentage of the face shown in an image) could be a factor in how the algorithm recognises the face as a face. Increasing the number of low-quality images in the training dataset for the algorithm may have an effect on the performance of high-quality images

[28], and this could increase the number of false positive and false negative recognitions, which may not be beneficial to an FRS. Practitioners could test what imagery works best with their FRS by creating different conditions of the same image.

e. Performance of the age progression image compared to the original image

Although the difference was significant between the age progression and the original image, the performance differed between the different FRS. Result suggests some FRS (Amazon and Face++) prefers the original images over an age progression image, but not all FRS have the same results i.e. No difference for Microsoft. This suggests that for most FRS, the original out-dated images were more useful, with only a few cases where age progression images achieved a higher confidence score.

Previous research addressing human recognition suggested no significant difference between out-dated images and age progressions [59]. This is perhaps even both image types are different from the target image they remain somewhat similar when compared. The higher recognition rate for an age progression could indicate that the growth prediction was similar for some individuals. The data used to develop this method was population specific, which could have been a contributing factor, especially when the FG-NET database is population unspecific.

The FG-NET database is open-source and often used to test algorithms in age related changes [60]. With these black-box systems, it is unclear if this could introduce bias if the algorithm has been trained on this dataset. The majority of comparisons preferred the original out-dated images, this shows the potential benefits to include the original images when using FRS, but also indicated that the majority of age progressions were not helpful. The common issues with users of commercial black box systems, it is unclear what images the algorithm works best with or what facial features the system is identifying. Systems should be tested before implementation, especially when used for identification.

f. Dataset evaluation

This study faces several limitations: quality of the images differs; it is only available as single source comparison; all age progressions were made by one practitioner, and manual age progression can differ between practitioners; the target age for each progression is dependent on the available ‘veridical’ images within the data for each individual.

Images within the FG-NET dataset are often variable in quality and age gap. Some images were black and white or decolourised due to the age of the photograph, and it is unknown when some of these images were taken. This means some black and white comparisons (12/83; 14.5%) were not available. Some of the images were already cropped, where the outline of the full head differed between images. Therefore, the variability of these images could have affected the difference in conditions.

The experiment for image variability has shown that a different target image at the same age was able to generate different recognition rates. This would no doubt have an effect when images were compared for their similarities. It is a difficult factor to control, especially when considering that indecent images of children will never be standardised. Therefore, multiple images should be used for identity verification or identification.

5. Conclusion

This study suggests softwares can have different performance. With a difference in recognition between the image types, the out-dated images could be more useful than an age progression and should be included when using machine-based recognition, however, this differs between the different FRS. Different conditions can influence machine-based recognition, and result suggests resolution reduction can have a positive effect, where the black and white and cropped conditions showed a negative effect.

In order to enable practitioners to test other FRS under different conditions, the method of testing a black box FRS should be further expanded and standardised. Having the ability to establish which condition works best could potentially improve the probability of a match. However, most FRS are developed using adult data, and a child's face is not a permanent biometric and cannot be treated the same as an adult's face. Developing an FRS with focus on child growth, with the ability to account for the difference in age gap, could potentially improve facial recognition of children.

The acceptance rate in comparing 'resemblance' images such as an age progression depiction should be further discussed. It is near impossible to generate an exact likeness with external

factors such as hairstyle, body modification, makeup, and other forms of alteration to the face. Should these images be treated the same as a normal facial comparison? Perhaps the tolerance of an FRS should be explored, it is unclear whether certain features are more superior for face recognition. By altering different facial features, this could provide a better understanding of the perception of faces by a machine.

If humans are able to recognise an individual based on depiction, perhaps an FRS could be trained to do similar tasks. This could increase false positive and negative identifications, however, if the purpose is to generate an investigative lead rather than an identification tool, the practicality of such tools should be explored.

6. References

[1] FBI. Privacy Impact Assessment (PIA) Child Victim Identification Program (CVIP) Innocent Images National Initiative (IINI). Federal Bureau of Investigation 2003. https://www.fbi.gov/services/information-management/foipa/privacy-impact-assessments/cvip (accessed January 16, 2019).

[2] NCMEC. Key Facts. Key Facts: Exploited Children Statistics 2018. http://www.missingkids.com/footer/media/keyfacts#missingchildrenstatistics (accessed January 16, 2019).

[3] Lampinen J, Arnal JD, Culbertson-Faegre A, Sweeney L. Missing and Abducted Children. In: Lampinen J, Sexton-Radek K, editors. Protecting Children from Violence: Evidence-Based Interventions, East Sussex: Psychology Press; 2010, p. 129–66.

[4] Grother P, Ngan M. Face Recognition Vendor Test (FRVT): Performance of face identification algorithms. Maryland: NIST; 2014.

[5] Ling H, Soatto S, Ramanathan N, Jacobs DW. Face verification across age progression using discriminative methods. IEEE Transactions on Information Forensics and Security 2010;5(1):pp.82–91.

[6] Jain AK, Ross A, Prabhakar S. An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology 2004;14 (1):pp.4–20.

[7] Matthews H, Clement J, Kilpatrick N, Fan Y, Claes P. Estimating age and synthesising growth in children and adolescents using 3D facial prototypes. Forensic Science International 2018;286:pp.61-69. doi:10.1016/j.forsciint.2018.02.024.

[8] Jain AK, Dass SC, Nandakumar K. Soft biometric traits for personal recognition systems. Biometric Authentication, Springer; 2004, p. 731–738.

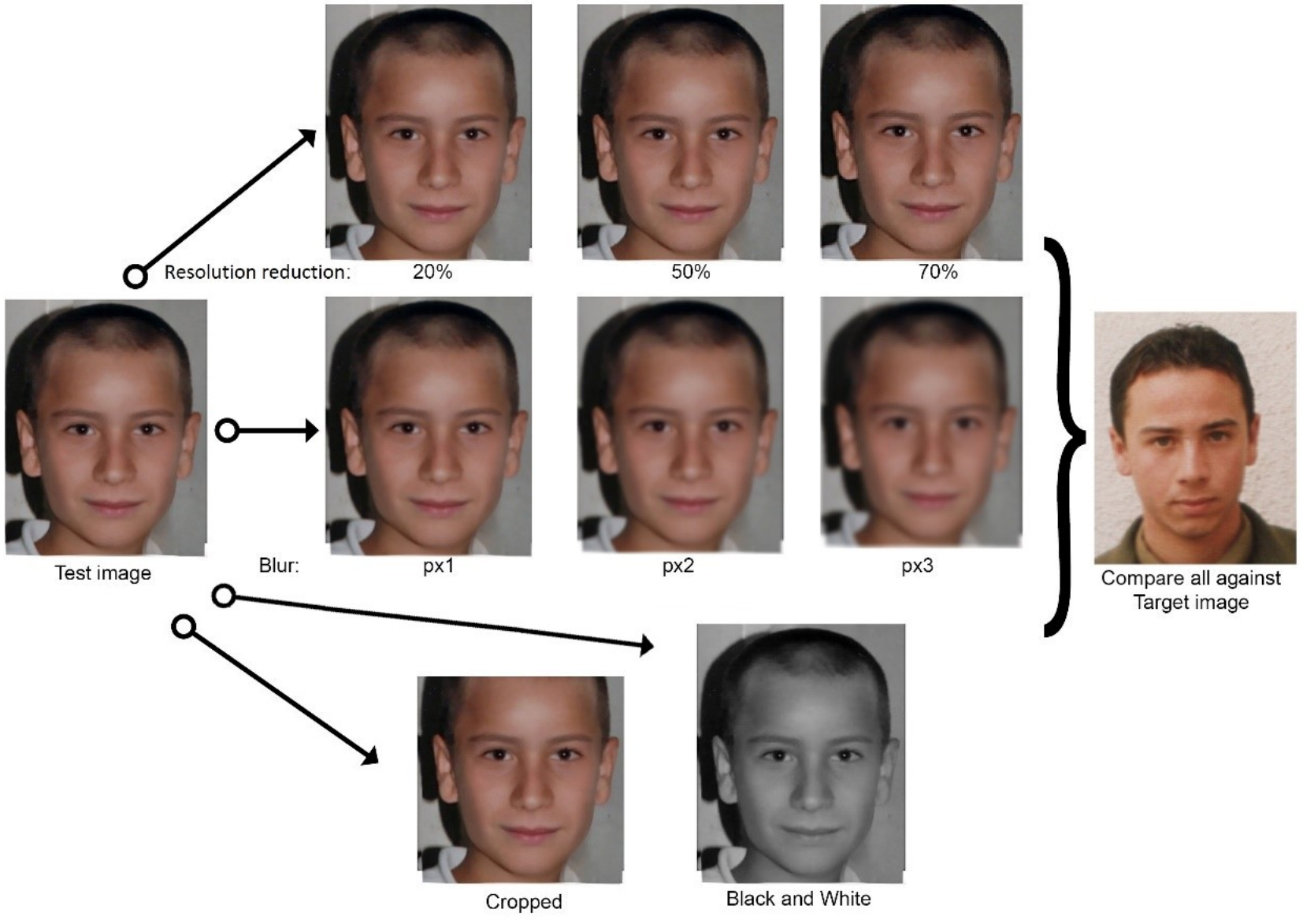
[9] Reid D, Samangooei S, Chen C, Nixon M, Ross A. Soft biometrics for surveillance: an overview. Handbook of Statistics 2013;31:pp.327–352. doi:10.1016/B978-0-444-53859-8.00013-8.

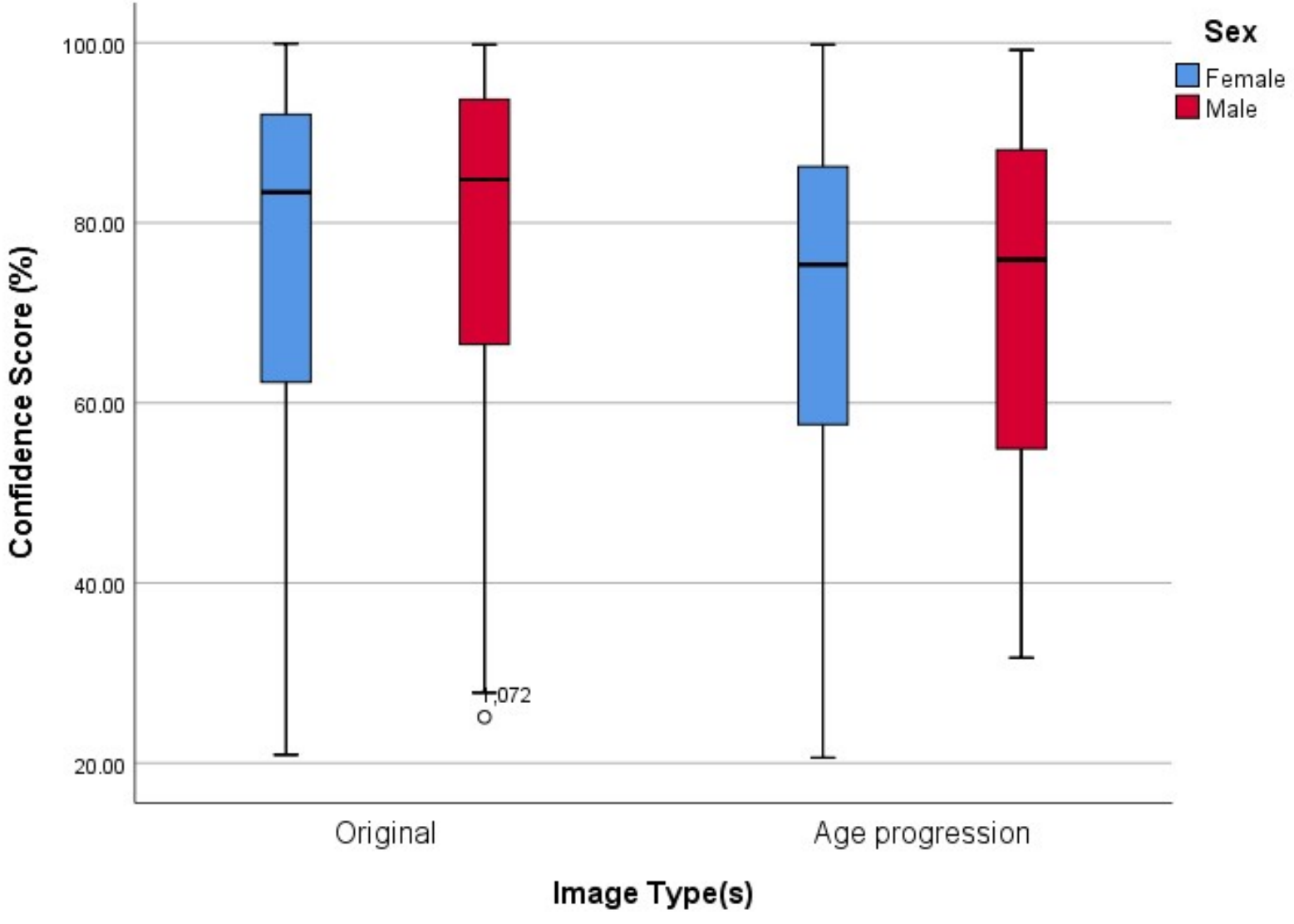
[10] Reid DA, Nixon MS. Using comparative human descriptions for soft biometrics. Biometrics (IJCB), 2011 International Joint Conference on, IEEE; 2011, p. 1–6.
[11] Ferguson E. Facial Identification of Children: A Test of Automated Facial Recognition and Manual Facial Comparison Techniques on Juvenile Face Images. The University of Dundee, 2015.
[12] Caplova Z, Compassi V, Giancola S, Gibelli DM, Obertová Z, Poppa P, et al. Recognition of children on age-different images: Facial morphology and age-stable features. Science & Justice 2017;57(4):pp.250-256. doi:10.1016/j.scijus.2017.03.005.
[13] Hunter D, Tiddeman B, Perrett D. Facial ageing. In: Caroline W, Christopher R, editors. Craniofacial Identification, Cambridge: Cambridge University Press; 2012, p. 57–67.
[14] Mullins J. Age progression and regression. In: Caroline W, Christopher R, editors. Craniofacial Identification, Cambridge: Cambridge University Press; 2012, p. 68–75.
[15] Ramanathan N, Chellappa R. Modeling age progression in young faces. Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1, IEEE; 2006, p. 387–394.
[16] Wu T, Chellappa R. Age Invariant Face Verification with Relative Craniofacial Growth Model. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, editors. Computer Vision – ECCV 2012, Springer Berlin Heidelberg; 2012, p. 58–71.
[17] Kemelmacher-Shlizerman I, Suwajanakorn S, Seitz SM. Illumination-aware age progression. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE; 2014, p. 3334–3341.
[18] Bukar AM, Ugail H, Hussain N. On Facial Age Progression Based on Modified Active Appearance Models with Face Texture. Advances in Computational Intelligence Systems, Springer, Cham; 2017, p. 465–79. doi:10.1007/978-3-319-46562-3_30.
[19] Scherbaum K, Sunkel M, Seidel H-P, Blanz V. Prediction of individual non-linear aging trajectories of faces. Computer Graphics Forum, vol. 26, Wiley Online Library; 2007, p. 285–294.
[20] Shen C-T, Huang F, Lu W-H, Shih S-W, Liao H-YM. 3D Age Progression Prediction in Children's Faces with a Small Exemplar-Image Set. Journal of Information Science and Engineering 2014;30(4):pp.1131–1148.
[21] Koudelová J, Dupej J, Brůžek J, Sedlak P, Velemínská J. Modelling of facial growth in Czech children based on longitudinal data: Age progression from 12 to 15 years using 3D surface models. Forensic Science International 2015;248:pp.33-40. doi:10.1016/j.forsciint.2014.12.005.
[22] Gibson SJ, Scandrett CM, Solomon CJ, Maylin MIS, Wilkinson CM. Computer assisted age progression. Forensic Science, Medicine, and Pathology 2009;5(3):pp.174–181.
[23] NCMEC. Long-Term Missing Child Guide for Law Enforcement 2016. http://www.missingkids.org/publications/longtermmissingguide (accessed March 13, 2017).
[24] Erickson WB, Lampinen JM, Frowd CD, Mahoney G. When age-progressed images are unreliable: The roles of external features and age range. Science & Justice 2016;57(2):pp.136-143. doi:10.1016/j.scijus.2016.11.006.
[25] Charman SD, Carol RN. Age-progressed images may harm recognition of missing children by increasing the number of plausible targets. Journal of Applied Research in Memory and Cognition 2012;1(3):pp.171-178. doi:10.1016/j.jarmac.2012.07.008.
[26] Lampinen J, Arnal JD, Adams J, Courtney K, Hicks JL. Forensic age progression and the search for missing children. Psychology, Crime and Law 2012;18(4):pp.405-415. doi:10.1080/1068316X.2010.499873.

[27] Lampinen J, Erickson WB, Frowd C, Mahoney G. Mighty Morphin'age progression: how artist, age range, and morphing influences the similarity of forensic age progressions to target individuals. Psychology, Crime and Law 2015;21(10):pp.952–967.
[28] Dodge S, Karam L. Understanding how image quality affects deep neural networks. Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, IEEE; 2016, p. 1–6.
[29] Grm K, Štruc V, Artiges A, Caron M, Ekenel HK. Strengths and weaknesses of deep learning models for face recognition against image degradations. IET Biometrics 2017;7(1):pp.81–89.
[30] Claes P, Vandermeulen D, De Greef S, Willems G, Clement JG, Suetens P. Computerized craniofacial reconstruction: conceptual framework and review. Forensic Science International 2010;201(1):pp.138–145.
[31] Claes P, Vandermeulen D, De Greef S, Willems G, Clement JG, Suetens P. Bayesian estimation of optimal craniofacial reconstructions. Forensic Science International 2010;201(1):pp.146–152.
[32] Mahoney G, Wilkinson CM. Computer-generated facial depiction. In: Wilkinson CM, Rynn C, editors. Craniofacial Identification, Cambridge: Cambridge University Press; 2012, p. 222–37. doi:10.1017/CBO9781139049566.
[33] Sinha P, Balas B, Ostrovsky Y, Russell R. Face recognition by humans: Nineteen results all computer vision researchers should know about. Proceedings of the IEEE 2006;94(11):pp.1948–1962.
[34] Bachmann T. Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? European Journal of Cognitive Psychology 1991;3(1):pp.87–103.
[35] Hole GJ, George PA, Eaves K, Rasek A. Effects of geometric distortions on face-recognition performance. Perception 2002;31(10):pp.1221–1240.
[36] Lander K, Bruce V, Hill H. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition 2001;15(1):pp.101–116.
[37] Frowd CD, Skelton F, Atherton C, Pitchford M, Hepton G, Holden L, et al. Recovering faces from memory: The distracting influence of external facial features. Journal of Experimental Psychology: Applied 2012;18(2):pp.224-238. doi:10.1037/a0027393.
[38] Toseeb U, Keeble DRT, Bryant EJ. The Significance of Hair for Face Recognition. PLOS ONE 2012;7(3):e34144. doi:10.1371/journal.pone.0034144.
[39] Grother PJ, Quinn GW, Phillips PJ. Report on the Evaluation of 2D Still-Image Face Recognition Algorithms. Maryland: NIST; 2010.
[40] Leonard KR. Assessment of Facial Recognition System Performance in Realistic Operating Environments. Face Recognition Across the Imaging Spectrum, Springer, Cham; 2016, p. 117–38. doi:10.1007/978-3-319-28501-6_6.
[41] Hoshino A, Saito T, Oka M. Inferencing the best AI service using Deep Neural Networks. Proceedings of the 2019 2nd International Conference on Information Science and Systems, ACM; 2019, p. 204–209.
[42] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on Fairness, Accountability and Transparency, 2018, p. 77–91.
[43] Jung S-G, An J, Kwak H, Salminen J, Jansen BJ. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. Twelfth International AAAI Conference on Web and Social Media, 2018.

[44] Puri R. Mitigating Bias in Artificial Intelligence (AI) Models. IBM Research Blog 2018. https://www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models/ (accessed June 12, 2019).
[45] Vincent J. The tech industry doesn't have a plan for dealing with bias in facial recognition. The Verge 2018. https://www.theverge.com/2018/7/26/17616290/facial-recognition-ai-bias-benchmark-test (accessed June 12, 2019).
[46] Virdee-Chapman B. Comparing Face Recognition: Kairos vs Amazon vs Microsoft vs Google vs FacePlusPlus vs SenseTime. Kairos 2018. https://www.kairos.com/blog/comparing-face-recognition-kairos-vs-amazon-vs-microsoft-vs-google-vs-faceplusplus-vs-sensetime (accessed June 12, 2019).
[47] Venditti L, Fleming J, Kugelmeyer K. Algorithmic Surveillance: A Hidden Danger in Recognizing Faces. Honors Thesis. Colby College, 2019.
[48] Snow J. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. American Civil Liberties Union 2018. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28 (accessed June 12, 2019).
[49] Kemelmacher-Shlizerman I, Seitz SM, Miller D, Brossard E. The megaface benchmark: 1 million faces for recognition at scale. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, p. 4873–4882.
[50] Farkas LG, Hreczko T. Age-related changes in selected linear and angular measurments of the craniofacial complex in healthy North American Caucasians. In: Farkas LG, editor. Anthropometry of the head and face. 2nd ed., New York: Raven Press; 1994, p. 89–102.
[51] Karahan S, Yildirum MK, Kirtac K, Rende FS, Butun G, Ekenel HK. How image degradations affect deep CNN-based face recognition? Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE; 2016, p. 1–5.
[52] Boom BJ, Beumer GM, Spreeuwers LJ, Veldhuis RNJ. The effect of image resolution on the performance of a face recognition system. Proceedings of the Ninth International Conference on Control, Automation, Robotics and Vision (ICARCV), Piscataway, NJ, USA: IEEE; 2006.
[53] Lemieux A, Parizeau M. Experiments on eigenfaces robustness. Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 1, IEEE; 2002, p. 421–424.
[54] Marciniak T, Chmielewska A, Weychan R, Parzych M, Dabrowski A. Influence of low resolution of images on reliability of face detection and recognition. Multimed Tools Appl 2015;74(12):pp.4329-4349. doi:10.1007/s11042-013-1568-8.
[55] Wang J, Zhang C, Shum H-Y. Face image resolution versus face recognition performance based on two global methods. Proceedings of Asia Conference on Computer Vision, vol. 47, 2004, p. 48–49.
[56] FISWG. Image Processing to Improve Automated Facial Recognition Search Performance (Section 4.x) 2016. https://fiswg.org/documents.html (accessed May 23, 2018).
[57] Abudarham N, Yovel G. Reverse engineering the face space: Discovering the critical features for face identification. Journal of Vision 2016;16(3):40:pp.1-18. doi:10.1167/16.3.40.
[58] Mehdipour Ghazi M, Kemal Ekenel H. A comprehensive analysis of deep learning based representation for face recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, p. 34–41.
[59] Lampinen J, Arnal JD, Adams J, Courtney K, Hicks JL. Forensic age progression and the search for missing children. Psychology, Crime & Law 2012;18:405–415.

[60] Panis G, Lanitis A. An Overview of Research Activities in Facial Age Estimation Using the FG-NET Aging Database. In: Agapito L, Bronstein MM, Rother C, editors. Computer Vision - ECCV 2014 Workshops, Springer International Publishing; 2014, p. 737–50.


Resolution reduction:
20%
50%
70%
Test image
Blur:
px1
px2
px3
Compare all against
Target image
Cropped
Black and White


Sex
Female
Male
100.00
80.00
60.00
40.00
20.00
Confidence Score (%)
,072
Original
Age progression
Image Type(s)

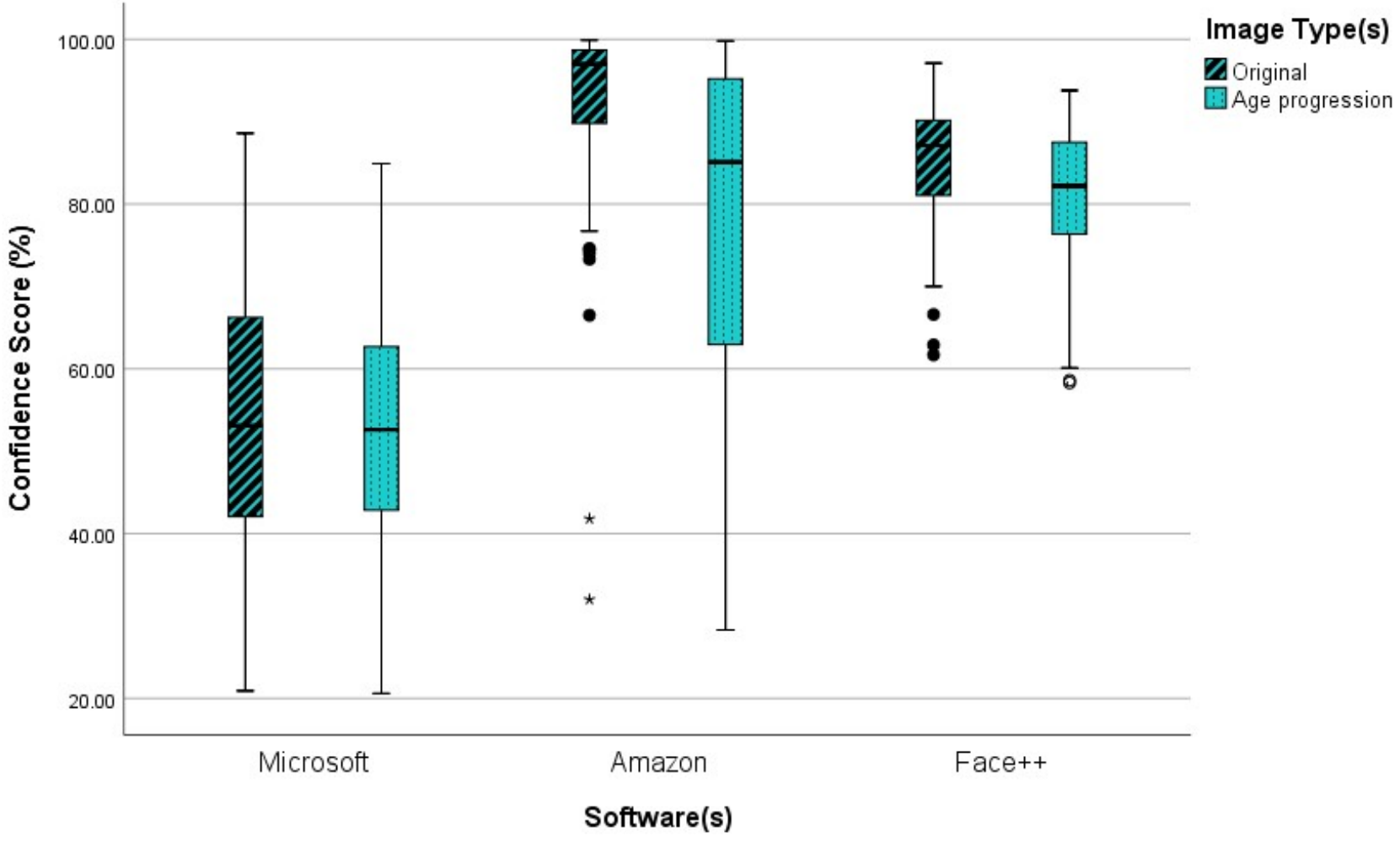


Image Type(s)
Original
Age progression
Confidence Score (%)
100.00
80.00
60.00
40.00
20.00
Microsoft
Amazon
Face++
Software(s)

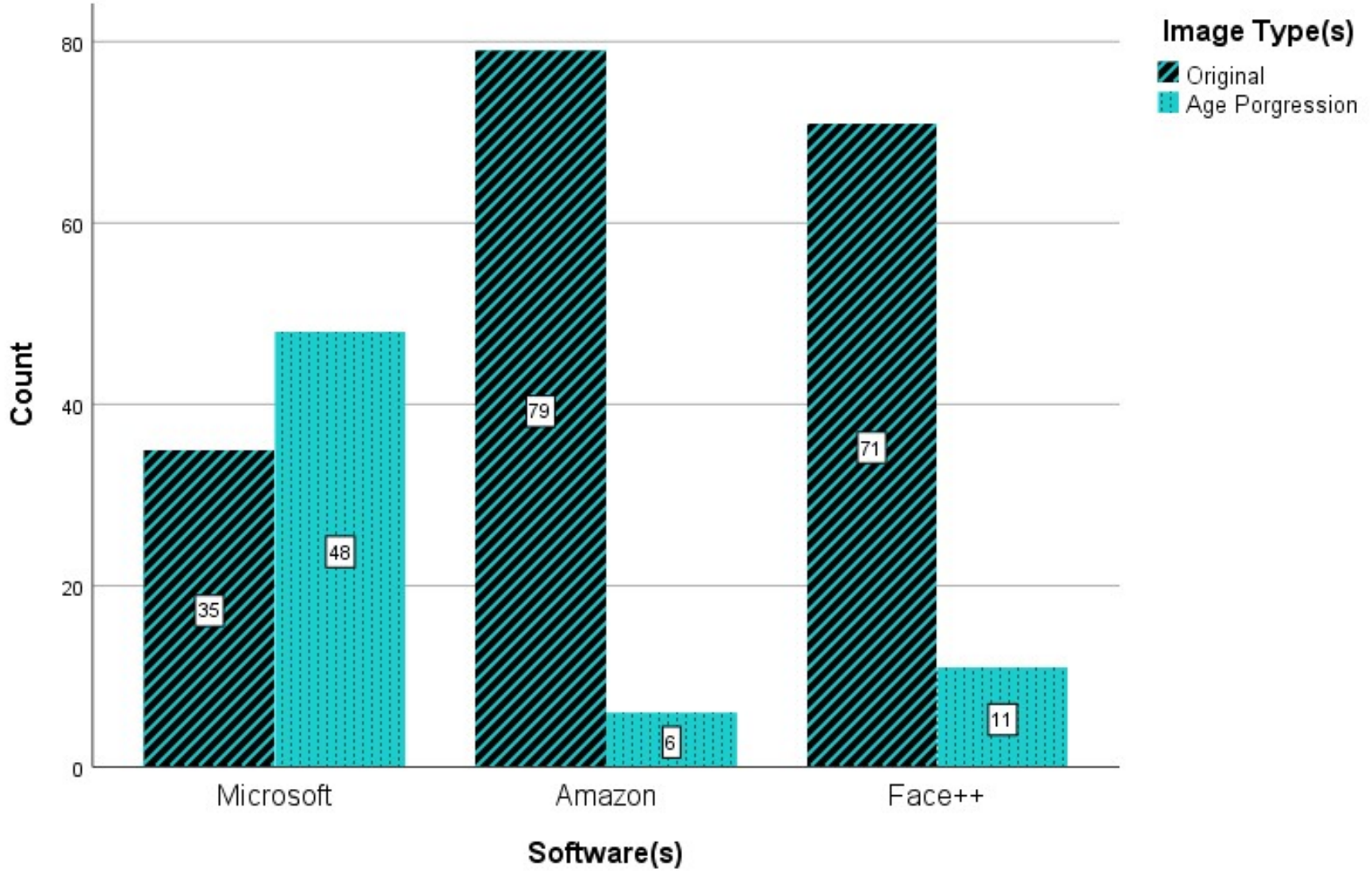
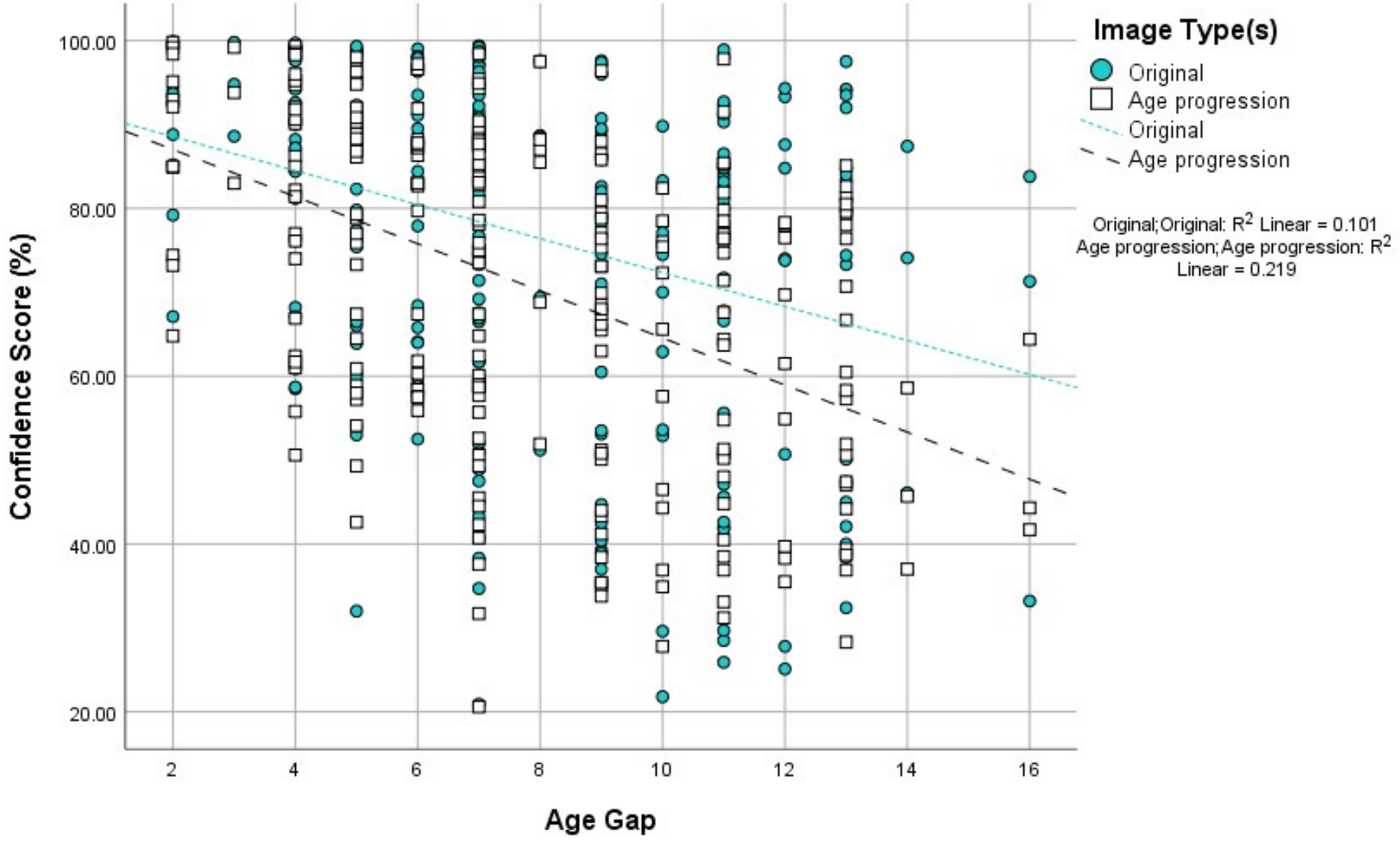
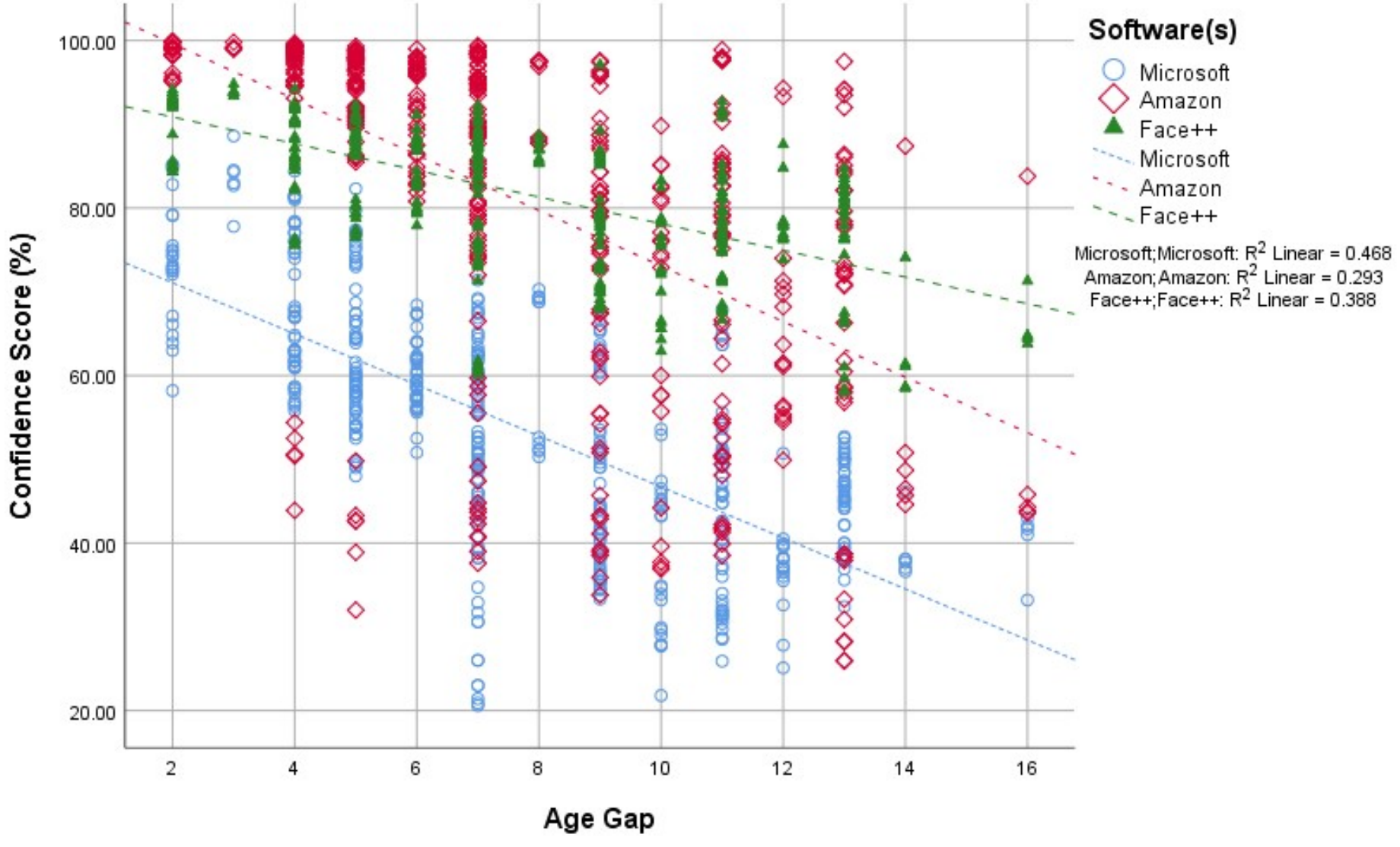
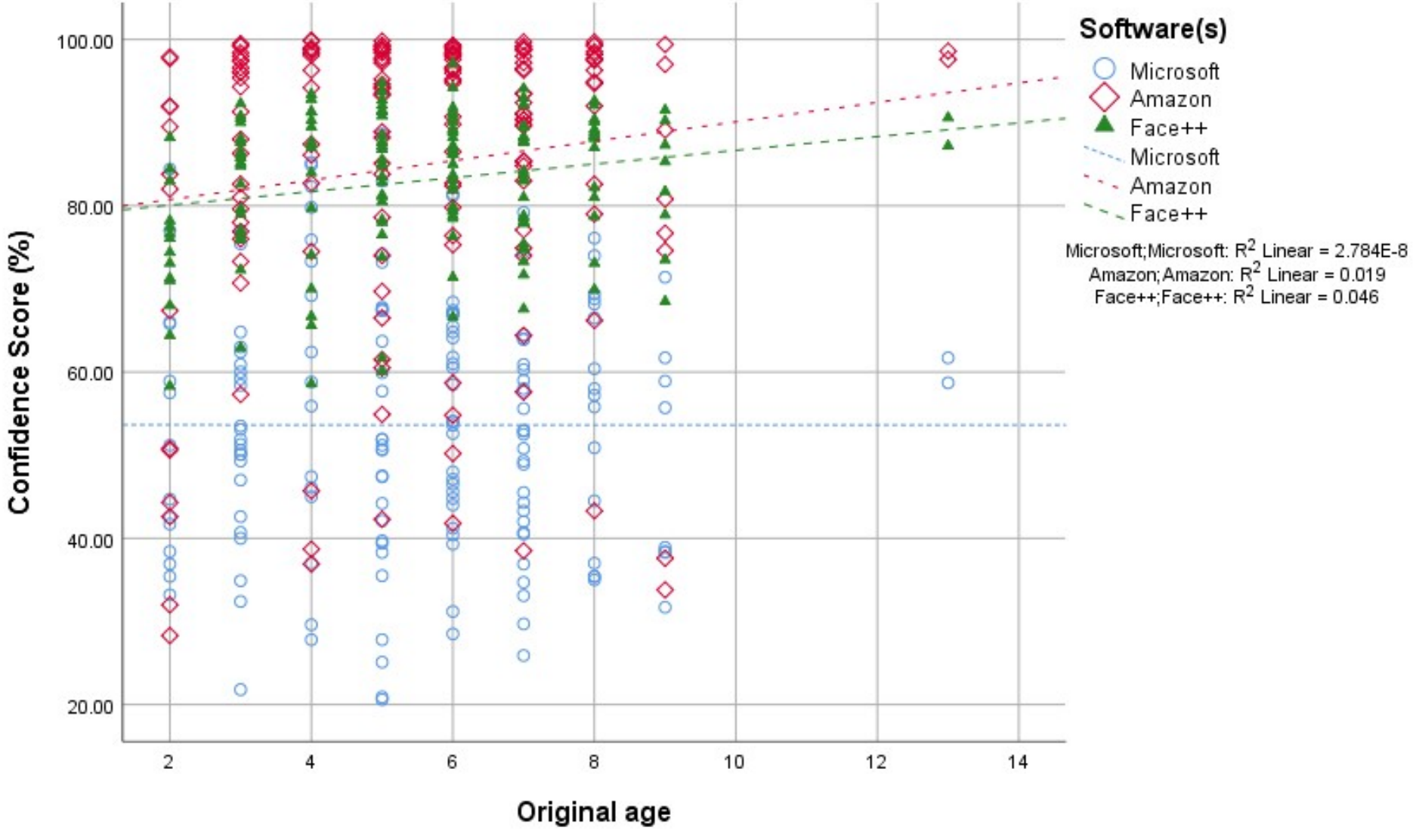
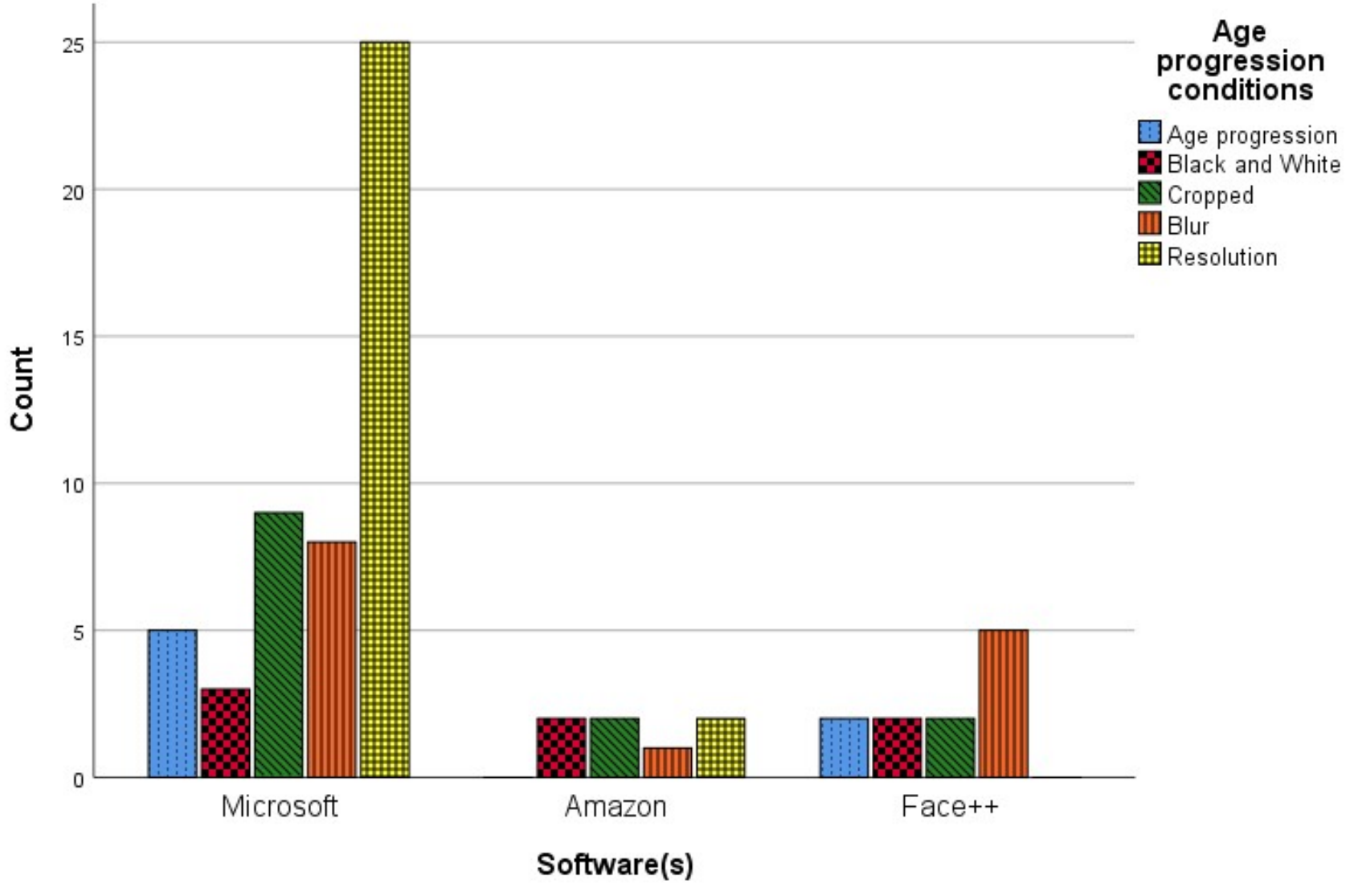


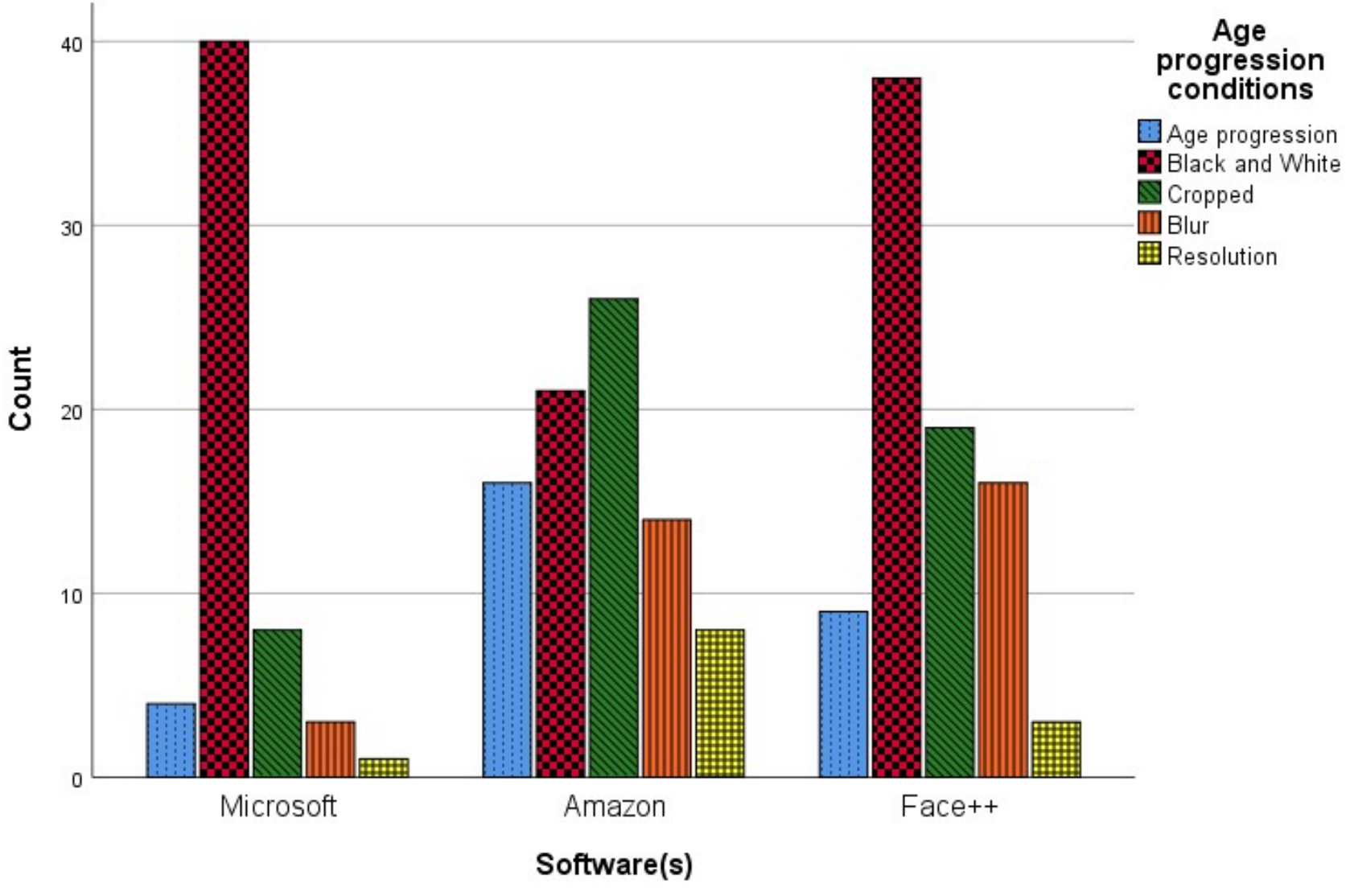
Image Type(s)
Original
Age Porgression
Count
80
60
40
20
0
35
48
79
6
71
11
Microsoft
Amazon
Face++
Software(s)


Confidence Score (%)
100.00
80.00
60.00
40.00
20.00
2
4
6
8
10
12
14
16
Age Gap
Image Type(s)
Original
Age progression
Original
Age progression
Original;Original: R2 Linear = 0.101
Age progression;Age progression: R2 Linear = 0.219

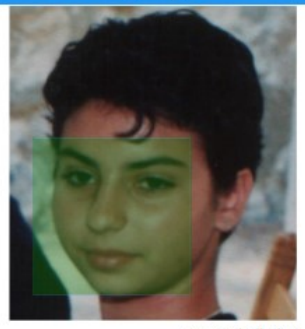

Confidence Score (%)
100.00
80.00
60.00
40.00
20.00
2
4
6
8
10
12
14
16
Age Gap
Software(s)
Microsoft
Amazon
Face++
Microsoft
Amazon
Face++
Microsoft;Microsoft: R2 Linear = 0.468
Amazon;Amazon: R2 Linear = 0.293
Face++;Face++: R2 Linear = 0.388


Confidence Score (%)
100.00
80.00
60.00
40.00
20.00
2
4
6
8
10
12
14
Original age
Software(s)
Microsoft
Amazon
Face++
Microsoft
Amazon
Face++
Microsoft;Microsoft: R2 Linear = 2.784E-8
Amazon;Amazon: R2 Linear = 0.019
Face++;Face++: R2 Linear = 0.046


Age progression conditions
Age progression
Black and White
Cropped
Blur
Resolution
Count
25
20
15
10
5
0
Microsoft
Amazon
Face++
Software(s)


Age progression conditions
Age progression
Black and White
Cropped
Blur
Resolution
Count
40
30
20
10
0
Microsoft
Amazon
Face++
Software(s)

Select image

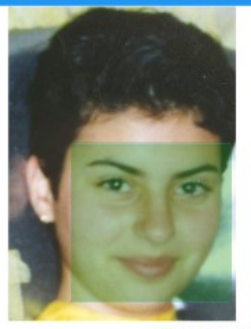


Select image

Compare

Azure Similarity	Amazon Similarity	Face++ Similarity
40.0%	79.6%	79.6%

Select image

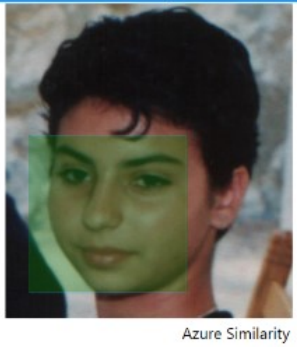


Select image

Compare

Azure Similarity	Amazon Similarity	Face++ Similarity
32.4%	73.3%	79.0%

Select image

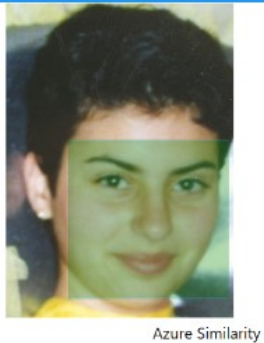


Select image

Compare

Azure Similarity	Amazon Similarity	Face++ Similarity
50.6%	70.7%	82.6%

Select image



Select image

Compare

Azure Similarity	Amazon Similarity	Face++ Similarity
47.0%	57.3%	79.6%

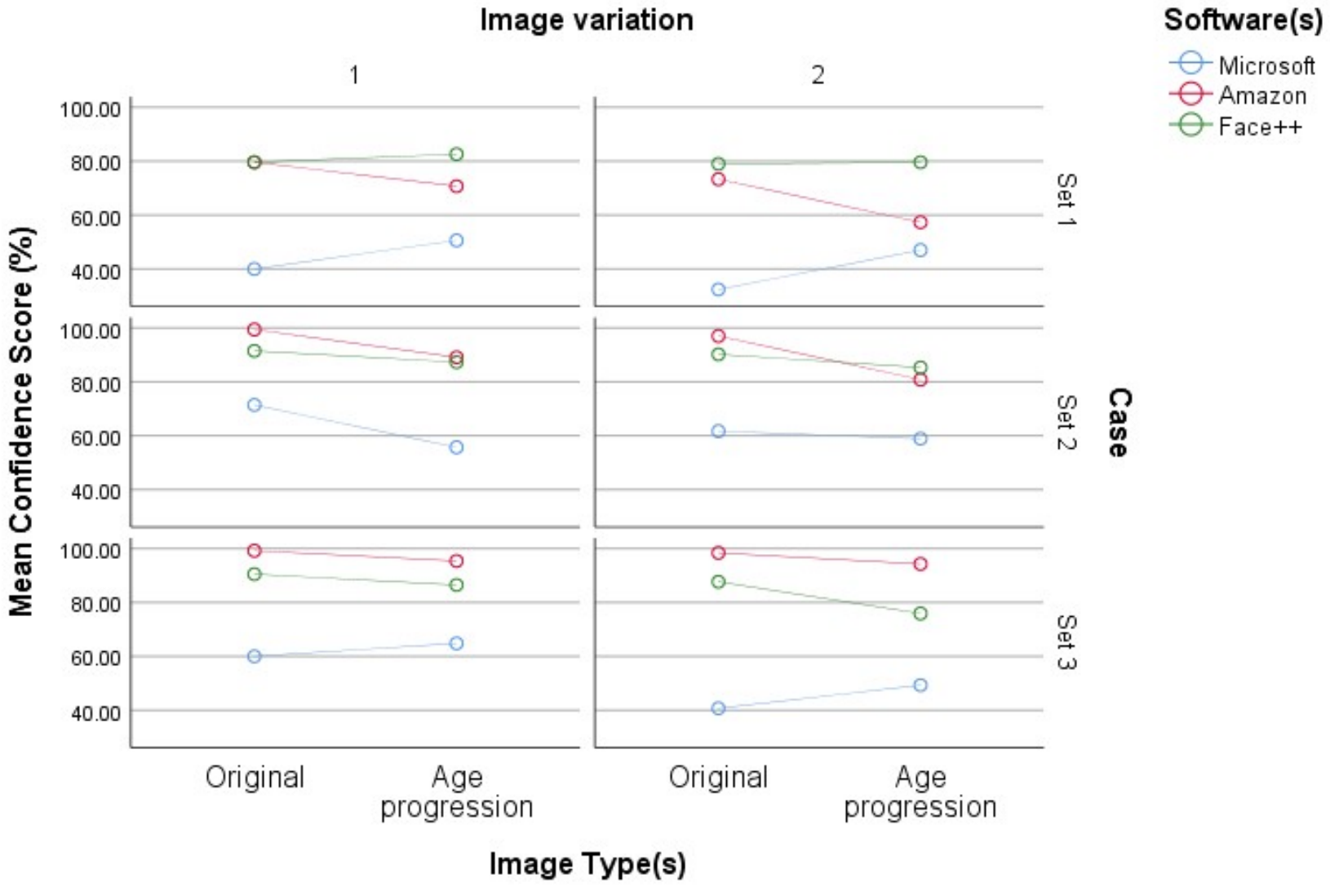


Image variation
1
2
Software(s)
Microsoft
Amazon
Face++
Mean Confidence Score (%)
100.00
80.00
60.00
40.00
Set 1
Set 2
Set 3
Case
Original
Age progression
Original
Age progression
Image Type(s)