Hummingbird Project:

Outcomes of Phase 1:
Human-like Computing in Face Matching

Professor Sarah Stevenage, on behalf of the Hummingbird Team.

Contents Page

Executive Summary	3
Human Factors affecting Face Matching	4
Algorithms and Datasets	5
Baseline Performance	6-7
Simulating Familiarity: Building Robustness at Enrolment	8-12
(i) Composites versus Multiple Instances at Enrolment	8-9
(ii) Variation via Different Poses at Enrolment	10-12
Simulating the Human Weighting of Internal Features	13-15
Simulating the Human Capacity for 3D Rotation	16-17
Conclusions	18
References	18

Executive Summary

The purpose of this report is to summarise the results of Phase 1 of the Hummingbird Project. This sought to incorporate human-like strategies into computer algorithms for face processing as used in verification mode involving GENUINE and IMPOSTER trials.

The goal of this phase of work was to identify the strategies used by human perceivers during face processing and then to find innovative ways of incorporating those strategies into the computer algorithms in the hope of enhancing capability. A broad survey of human face processing revealed four factors of importance: facial distinctiveness, facial familiarity, internal feature weighting, and the use of 3D information for mental rotation.

Facial distinctiveness is a beneficial factor in terms of human face processing and is associated with what is known as a 'distinctiveness advantage'. Surprisingly, however, when this was explored in the context of computer processing, distinctiveness was not associated with performance at all. The lack of correlation may have emerged because human ratings of distinctiveness capture something other than mere stimulus characteristics. For instance, they may capture subjective traits such as likeability or attractiveness. Nevertheless, the lack of a distinctiveness advantage within any of the computer algorithms usefully highlights a difference across human and computer decision makers. Specifically, whilst the computer is unaffected by distinctiveness, the human will be especially good at face recognition when a face is distinct.

Facial familiarity has a straightforward effect on human face processing capabilities - human perceivers perform better with faces they know than with faces they don't. One reason for this familiarity effect may be that the perceiver has a wealth of knowledge about the subtle variation in appearance of a familiar face which assists their decision-making. When building this into the computer algorithm, the benefit was slim. A set of multiple images provided better overall performance than a single composite, possibly because of the introduction of ghosting within the composite generation. A set of images may also offer some resilience against difficulty when a test image was rotated from the ideal full-frontal passport style image especially at larger angles of rotation.

Internal feature weighting represents a very powerful factor guiding human face perception. Attention may be weighted towards internal features (over external features) because their high contrast draws the eye, or because of their social informativeness. Additionally, attention to the internal features represents what is known as a 'hallmark of familiarity' in that attention shifts more to the internal features as a face becomes more familiar. When segments corresponding to internal features were weighted more heavily in an overall match decision conducted by a simple image matcher, performance significantly and substantially improved. This rule may well be implicit within a deep learning neural network designed for face processing. However, our work makes this rule explicit, transparent and explainable, and shows the very clear benefits that result.

Finally, the human perceiver cannot ignore their knowledge of the fact that faces are 3D objects. Their use of 3D information may enable them to mentally rotate a novel instance of a face in order to provide a better point of comparison to a mentally held template. When this thinking was applied to the computer process, by incorporating 3D rotation as a precursor to 2D matching, performance was again improved.

As a take-home message, computer algorithms for face matching will be improved by the use of multiple enrolment images, a match process which favours internal over external features, and 3D rotation prior to a 2D match. In this way, the strengths of the human perceiver can successfully be captured in an automated process, improving transparency and capability without incurring a substantial computational burden.

Human Factors affecting Face Matching

Humans are remarkably adept at recognising faces, even under challenging conditions. The challenges to human face recognition come in two forms: person-related factors and image-related factors. Person-related factors include factors such as expression and age, which reflect changes to the face itself. Image-related, or viewing-related, factors include viewpoint and lighting that depend on the relation between the observer and the face. There is a considerable body of literature exploring the strategies that humans use to overcome these challenges to face recognition (Johnston & Edmonds, 2009). The strategies associated with successful human face recognition include the creation of a more robust representation, a focus on internal features, and the use of 3D facial information.

It is well-known that humans are significantly better at recognising familiar compared to unfamiliar faces. This advantage likely occurs because humans have much more experience with familiar faces including, crucially, how they can vary under different conditions. Due to this advantage for familiar face recognition, several recent studies have explored how faces become familiar and how familiar faces are represented in the human brain. As to the former, it is thought that humans learn new faces through exposure to within-person variability. With exposure to variability, humans learn what features are stable across instances as well as how much a given face can vary (Bruce, 1994). Recent evidence suggests that exposure to greater variability facilitates the learning of new faces (Andrews et al., 2015). Therefore, it is possible that exposing the algorithm to greater variability may improve recognition performance. In terms of how familiar faces are represented in the brain, there are two prominent hypotheses. One possibility is that humans store a prototype of each face, which reflects the average across multiple instances of a face (Burton, Jenkins, Hancock, & White, 2005). Alternatively, humans may store each individual instance of a face. There is currently some evidence to suggest that humans benefit from face averages, which raises the question of whether average, or composite, faces are beneficial to machine performance as well.

Over the years, several studies have attempted to identify the specific properties of the face that are most important for successful face recognition. While there is no single feature that is most diagnostic of identity for all faces, performance is facilitated by attention to distinctive facial features and attention to internal features (e.g. eyes, nose, mouth) over external features (e.g. hairline and face shape). Internal features may be advantageous since these features are less likely to change over time compared to external features, which are more variable (Longmore, Liu, & Young, 2015). In support of this internal feature advantage, eye-tracking studies have shown that humans exhibit a stereotypical t-shaped scanning pattern between the two eyes and the mouth when viewing a face (Yarbus, 1967). Therefore, differentially weighting internal features may improve machine performance.

The human visual system is able to extract some 3-dimensional cues from 2-dimensional images, resulting in what is referred to as a '2½ dimensional' representation. The ability to extract surface and depth cues helps to form a structural representation of the face that is less reliant on image properties. For this reason, 3D face representations may support recognition across different facial transformations, such as different viewpoints (Troje & Bülthoff, 1995) meaning that the human perceiver is fairly tolerant to changes that result from rotation of the head. In terms of machine performance, creating a 3D representation of the face may enable more accurate performance across viewpoints.

In summary, successful human face recognition is supported by robust neural representations that incorporate within-person variability, a focus on internal features of the face, and 3D facial representations. By incorporating these human-like strategies into the computer algorithm, it may be possible to improve transparency and capability.

The Algorithms and Datasets

A VERIFICATION test was used to determine machine capability when matching faces. Here, we summarise the algorithms and datasets used together with a rationale for these choices.

Algorithms

Testing was conducted using 5 machine algorithms. These included:

Neurotechnology VeriLook	an industry-leading neural-network based black box algorithm providing a state of the art benchmark of capability. See https://www.neurotechnology.com/cgi-bin/biometric-components.cgi?ref=vl&component=face-mat
Eigenface	A holistic face matching algorithm with a method of processing that looks at the relative placement of features. This was selected as an approximation in method to the human strategy of configural processing. See Turk & Pentland (1991).
Local Binary Pattern (LBP) Matching	A pixel-based face matching algorithm with a method of processing that looks at the pixel intensities of pixels around each point in turn in order to compute a similarity score. This was selected as a comparison point to the above. It is NOT how humans process faces. See Ahonen *et al.*, (2004).
4SF	A hybrid open source face matching algorithm (OpenBiometrics) which combines the holistic approach of Eigenface with the local approach of LBP. This perhaps represents a 'best of both' approach. See Klare (2011).
Dlib	A face matching algorithm based on a neural network, and thus providing a second neural network approach for comparison to Neurotechnology. See https://github.com/davisking/dlib

The Dataset

The studies conducted within this phase of research used the faces gathered within the EPSRC funded SuperIdentity Stimulus Database (www.southampton.ac.uk/superidentity)[1]. This consisted of a total of 121 Caucasian faces from 18-40 years, captured under systematically varying angles of rotation (-90 to +90 degrees from full-face), and expression. The database contained test images collected on the same day but taken in 2 different venues and captured using 2 different cameras.

For the purposes of this project, the faces of 98 individuals (49 males, 49 females) were used as stimuli as these individuals provided a full set of images across all required conditions.

Testing

Three algorithms (Eigenface, LBP, and 4SF) required multiple images at the enrolment stage from which to abstract an 'average' for comparison to the test image. In contrast, the remaining two algorithms (Neurotechnology, Dlib) could take single images or multiple images at enrolment, for comparison to the test image. Protocols varied accordingly through the testing sequence.

[1] This database is unfortunately NOT available for research purposes given the linked data held for each data participant.

Baseline Performance

The purpose of this phase of study was to determine the baseline level of performance on a face matching task for each of our algorithms of interest. If they performed perfectly, there would be little room for improvement. However, if they performed well but NOT at ceiling, there would be the potential for improvement when human-like capabilities were incorporated.

Each algorithm worked by (i) seeking to detect the face, and then (ii) performing a comparison between the enrolled face and the test face. Consequently, a comparison was only performed if the enrolled and test faces had been detected.

The comparison of enrolled and test faces generated a continuous matching score which was expressed as a 'distance score' between 0 and 1 for Eigenface, LBP, and 4SF, (where 0 indicated high similarity and 1 indicated high difference between enrolled and test faces), and as a 'match score' between 0 and 1 for Dlib, and between 0 and ∞ for Neurotechnology (where 0 indicated a low match, and higher values indicated a high match). These continuous matching scores can be treated as indicating the *extent* of a match or no-match decision, and thus provide a sensitive measure of performance. Continuous scores were presented and analysed for all algorithms in all studies.

Once obtained, the continuous scores are traditionally converted to a categorical match/no-match decision by application of a threshold which seeks to minimise the number of false-positive errors (an imposter being erroneously matched) and the false-negative errors (a genuine user being erroneously rejected). The accuracy of the resultant categorical decision (or classification) may be evaluated using receiver operating characteristic (ROC) curves and equal error rate (EER) values. These two measures are only of value if the number of errors (either failures to find the object, or misclassifications once found and compared) differ across conditions. Within the studies completed here, ROCs and EERs only differed across conditions for one algorithm in one study (the Internal Feature Weighting study) suggesting their general lack of value for our studies. Given the common practice within the Computer Science literature of focussing on match scores, ROCs and EERs are not reported for any of the work presented here.

Baseline performance was evaluated for all 5 algorithms according the following protocol:

Algorithms were provided with one image or a set of images at enrolment prior to a single, non-identical test image. Enrolment images either reflected a standard full-frontal neutral (set of) image(s), a full-face expressive (set of) image(s), or a neutral but rotated (set of) image(s). The test image was always full-face and of neutral expression, thus providing *baseline* conditions for testing.

The data suggested improvement in performance when more images were used at the enrolment stage. There was some difference in performance across the 5 algorithms, with the data suggesting optimal performance using the Dlib algorithm. More importantly, the data suggested that performance was not perfect, and thus could be improved upon.

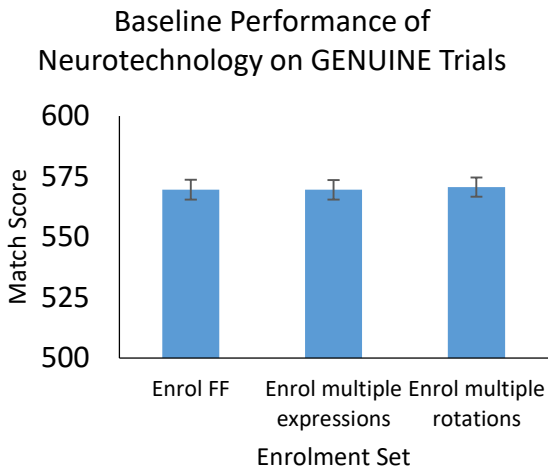
Baseline Performance: Continuous Matching Scores

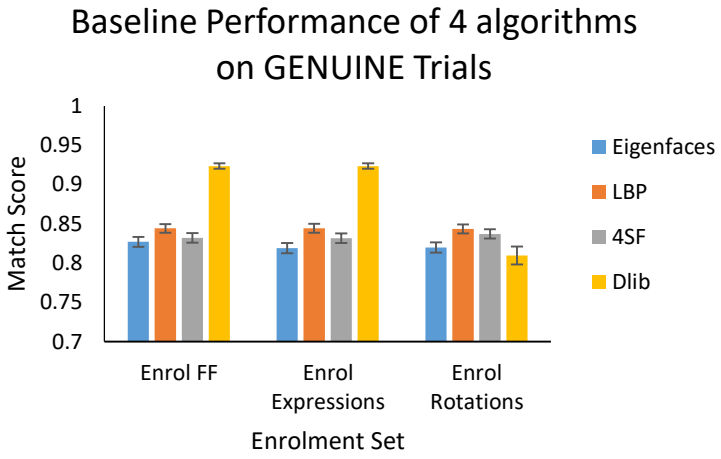
For Neurotechnology and Dlib algorithms, a high score indicated a good match between the target and test. Thus, GENUINE trials should attract high scores, and IMPOSTER trials should attract low scores. Scores for the remaining algorithms reflected a 'distance score' rather than a 'similarity score' but have been reversed (1-distance = similarity) for consistency of graphical representation within this report.

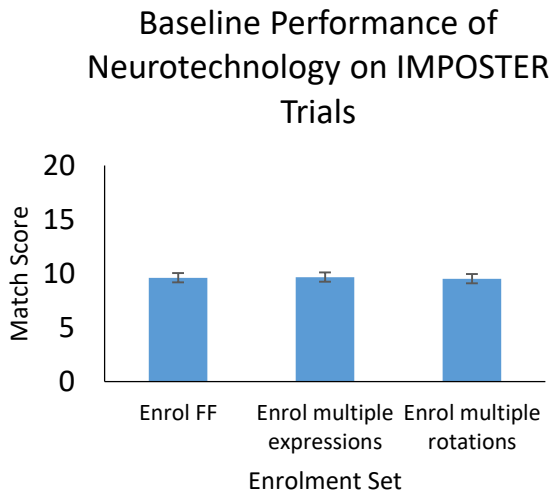
For the Neurotechnology algorithm, there was no difference in performance according to the number of images at enrolment. For the other 4 algorithms, Dlib appeared to be the strongest algorithm in terms of both

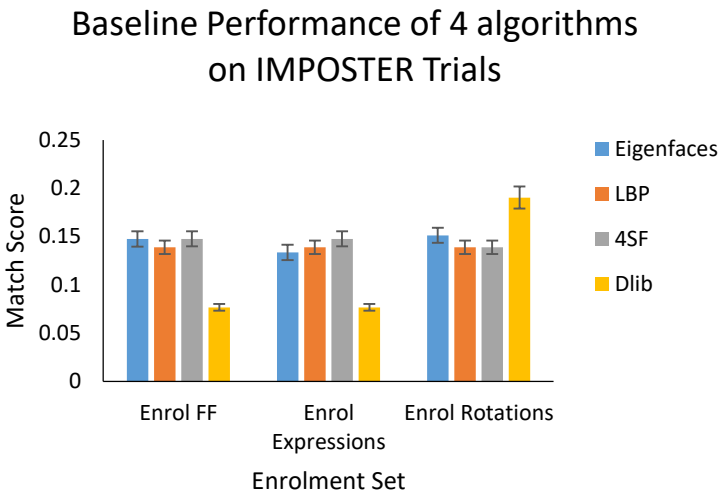
GENUINE trials and IMPOSTER trials as long as rotated images were not used at enrolment. The Eigenface algorithm was the weakest algorithm on GENUINE trials, but there was no difference across Eigenface, LBP and 4SF on IMPOSTER trials.

The data suggested slightly better performance on IMPOSTER trials when more images were used at enrolment because there was more information to contrast with the imposter face at test. More importantly, the data also suggested that performance was not at ceiling either for GENUINE or IMPOSTER trials.









In a Nutshell: Summary of Baseline Analysis

THE CONTINUOUS MATCHING SCORES REVEALED DIFFERENCES IN CAPABILITY ACROSS ALGORITHMS.

OF MOST IMPORTANCE, PERFORMANCE WAS NOT AT CEILING FOR ANY ALGORITHM, SUGGESTING ROOM FOR IMPROVEMENT.

CURIOUSLY, PERFORMANCE DID NOT CORRELATE WITH DISTINCTIVENESS (ALL R VALUES <.139).

Simulating Familiarity: Building in Robustness at Enrolment

(i) Composites versus Multiple Images at Enrolment

One potential weakness of a machine learning algorithm may result from reliance on a single image at enrolment and a single image at test. In contrast, humans store a wealth of knowledge about a known person, and they can quickly capture a range of instances from a brief live interaction with an unfamiliar person.

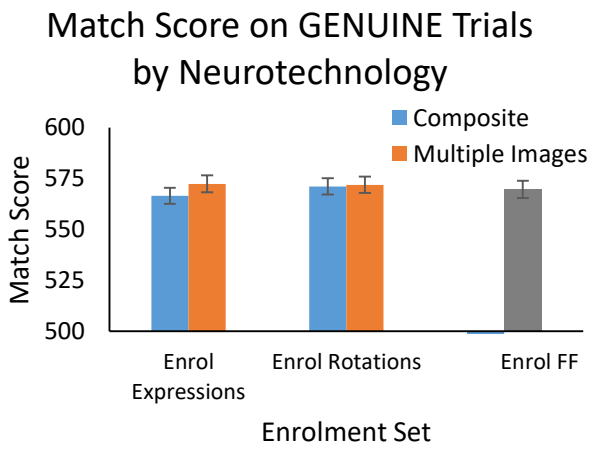
One strategy to address this potential machine weakness is to provide multiple images at enrolment. This may be achieved in two distinct ways, and the literature on human processing is mixed as to what humans do. Multiple images may be combined into a single *composite*. Alternatively, multiple images may be stored as *separate multiple instances*. The purpose of the next test was to determine (i) whether a single composite may be more efficient as a stored template compared to a set of multiple instances, and (ii) whether both the composite and the multiple instances may lead to better matching performance than the baseline condition involving a single enrolled image.

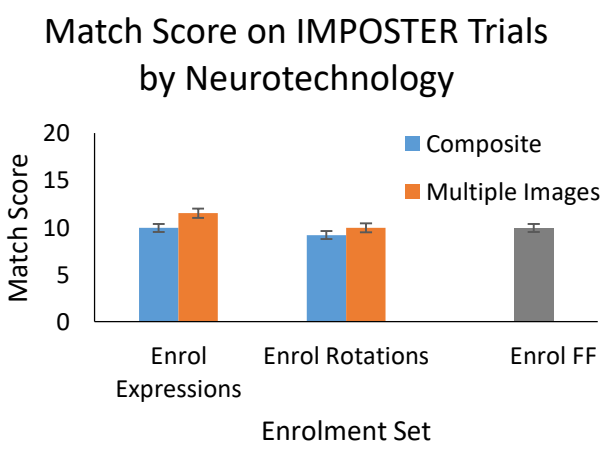
This test could only be conducted with 2 of our 5 algorithms – Neurotechnology and Dlib - because these were the only algorithms where enrolment could reflect either a single composite image, or a set of images. Testing was conducted by providing either a single composite of expressive or rotated images, or a set of expressive or rotated images at enrolment, prior to a single full-face neutral (non-identical) test image. Composites were generated by application of the Interface software created at the University of York (see http://www.facevar.com/downloads-interface).

Performance was evaluated relative to the baseline condition in which a single full-face image was used at enrolment, and a non-identical single full-face neutral image was used at test. When a set of images was used at enrolment, performance was evaluated against the *best match in the set*. Usually, but not always, this best match was the image that corresponded most closely to the pose of the test image.

Composites versus Multiple Images at Enrolment: Continuous Match Scores

In terms of the continuous match scores, the Neurotechnology algorithm showed no measurable benefits of either the composite or the multiple image set over the single enrolled image when GENUINE trials were considered. However, on IMPOSTER trials, there was a small but significant benefit to performance when the composite rather than the multiple images were enrolled. Moreover, there was a general benefit when enrolling a set of images varying in rotation rather than expression, because rotation made the target look less like the imposter. This said, neither the composite nor the set of images gave any benefit over a single enrolment image.

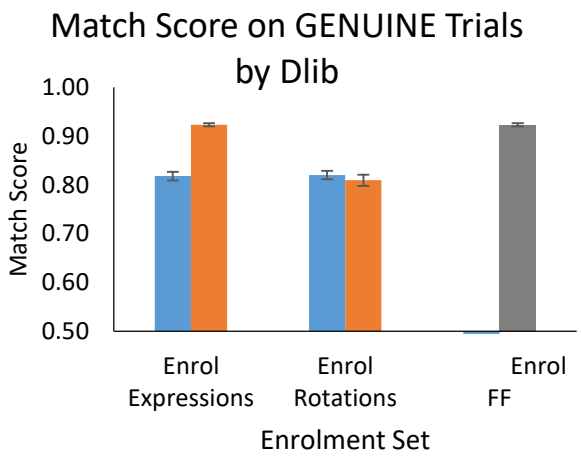


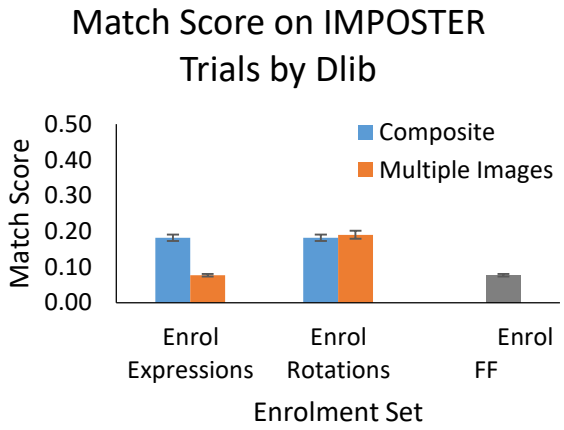


Considering Dlib as a point of comparison, a different overall picture emerged to that for Neurotechnology.

Some differences in performance existed between composite and multiple image enrolment conditions when the set of images differed in expression. This most likely reflected the degree of image ghosting around the mouth region in particular when generating the composite across smiling and non-smiling (expressive) images. In contrast, performance was equivalent when based on the composite and the multiple enrolment images that varied in rotation.

Of most importance, however, the provision of either a composite image or a set of multiple images at enrolment did not improve performance compared to the baseline condition using a single enrolment image. In fact, matching scores were equivalent or worse when composite or multiple inputs were provided at enrolment compared to the baseline single enrolment condition.





In a Nutshell: Composites versus Multiple Images at Enrolment

A benefit may occur when a set of several instances is enrolled

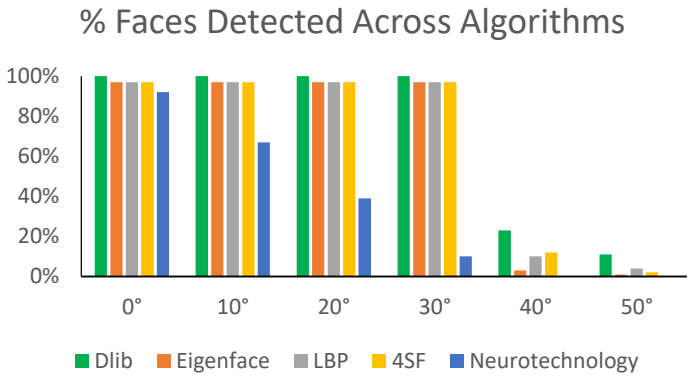
TAKEN AS A WHOLE, THERE WAS NO CLEAR BENEFIT WHEN ENROLLING A SET OF IMAGES, OR A COMPOSITE IMAGE, RELATIVE TO A BASELINE SCENARIO INVOLVING A SINGLE IMAGE.

IF MULTIPLE IMAGES ARE TO BE USED AT ENROLMENT, THEY ARE BEST PRESENTED AS A SET, RATHER THAN AS A COMPOSITE, POTENTIALLY DUE TO GHOSTING WITHIN THE COMPOSITE ITSELF.

THIS APPROACH PROVIDED A LEVEL OF TRANSPARENCY THAT WAS USEFUL EVEN IF PERFORMANCE WAS NOT IMPROVED.

(ii) Variation via Different Poses at Enrolment

A second weakness of the machine algorithm may be a limited capacity to generalise its match ability to anything other than the standard full-face image. In other words, it may do well at matching between a full-face neutral enrolled image and a full-face neutral test image. However, it may perform less well when the test face deviates from this optimal image.



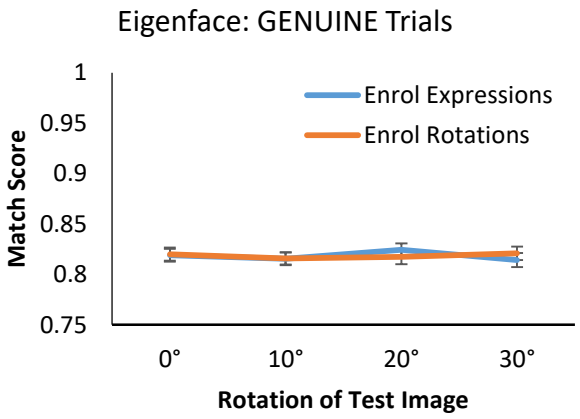
This sort of weakness stems from a marked reduction in ability to locate the face in an image when that face is rotated (see figure). This problem is especially apparent in the Neurotechnology algorithm. In comparison, this impact of rotation is less evident in human perceivers than in machine algorithms as human perceivers can draw on their knowledge of how a face changes as it is rotated.

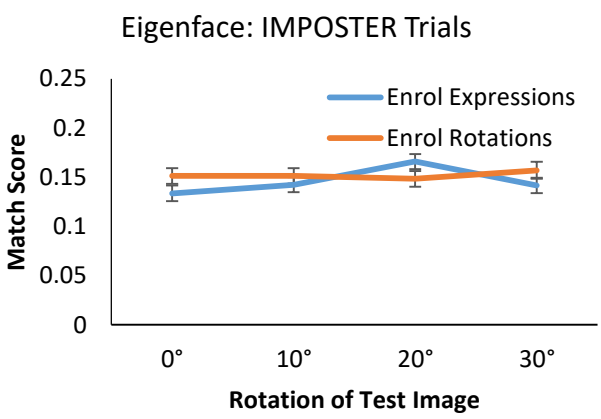
Accordingly, the purpose of the next test was to determine whether it was possible to improve the performance of an algorithm when matching to a rotated test face by using rotated images at enrolment.

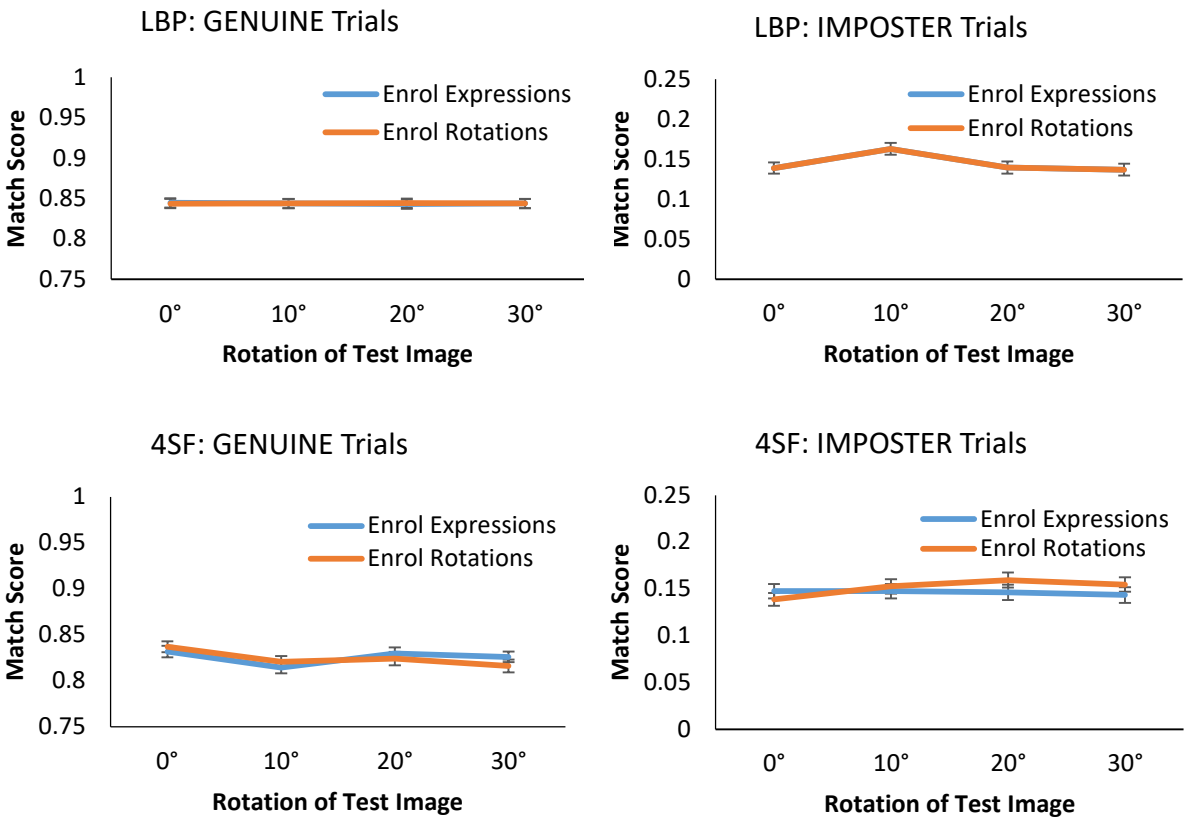
For the purposes of this test, all 5 algorithms were used. They were presented with a set of rotated images at enrolment (0°, -10°, -20°, -30°) prior to a non-identical full-face or angled test image (0°, 10°, 20°, or 30°). In this way the nature of the variation in the enrolment set was RELEVANT to the nature of the variation in the test set. As a point of comparison, performance was evaluated when the enrolment set varied on an IRRELEVANT dimension by enrolling a set of full-face expressive images (happy, sad, neutral, angry). The expectation was that performance would be facilitated following enrolment of a set of images which reflected the nature of the test image.

Training on Different Poses at Enrolment: Continuous Match Scores

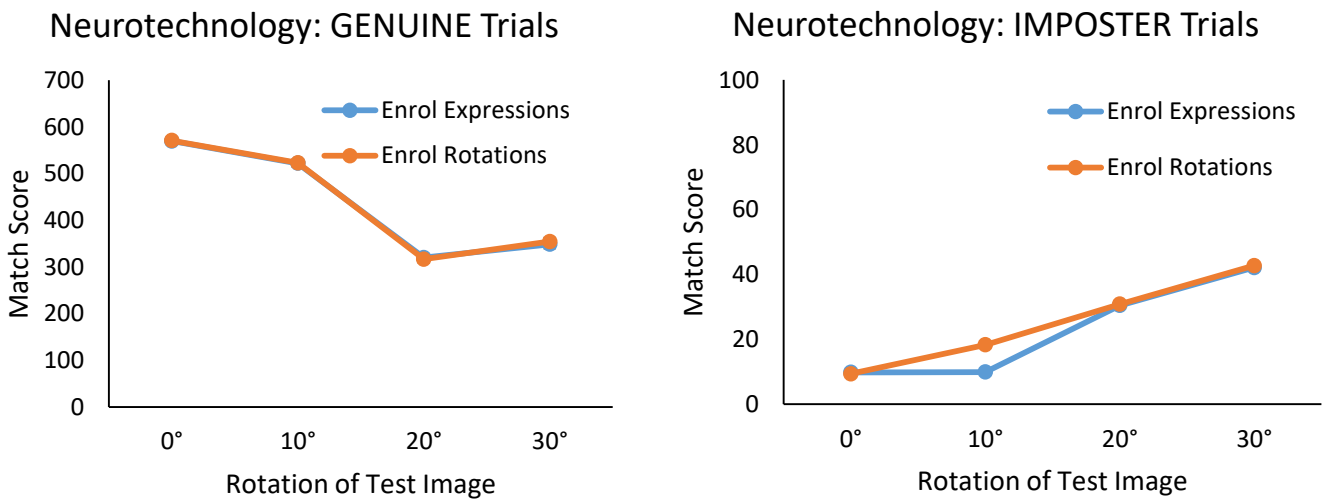
As previously, the continuous match scores provided a sensitive measure which could reveal a benefit in performance. When evaluating matching performance for a rotated test image, it is worth recalling that Eigenface, LBP and 4SF remained able to detect a face up to 30° of rotation. Consequently, matching performance to a rotated test image up to this level of rotation posed little problem. There was no decline in performance as the test image was rotated meaning that there was no deficit to be corrected by relevant training. This pattern held both for GENUINE trials and for IMPOSTER trials.



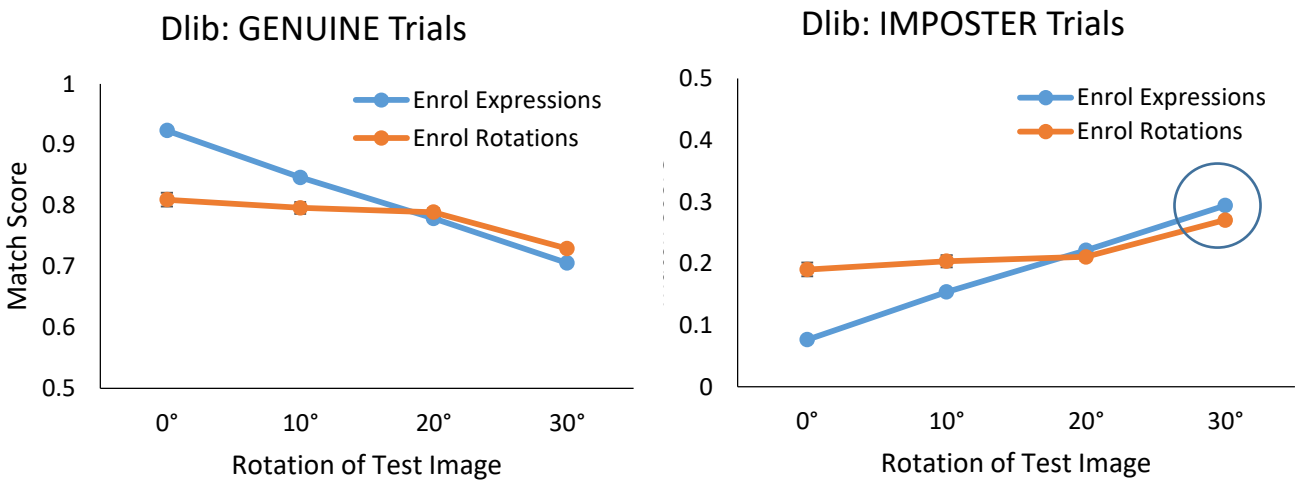




In contrast, when examining the performance of Neurotechnology, the ability to detect a face was affected by angle of rotation. Accordingly, matching performance significantly declined as the test face was rotated and this was shown on both GENUINE trials and IMPOSTER trials. Despite this, we were surprised to see that training on the relevant (rotated) dimension did not offset this difficulty compared to training on the irrelevant (expressive) dimension.



Finally, examination of the final algorithm – Dlib – suggested that performance was similarly affected by angle of rotation. Matching performance significantly declined as the test face was rotated and this was shown on both GENUINE trials and IMPOSTER trials. This time, however, whilst rotations at enrolment impaired performance at minor rotations, a very small benefit was evident when the enrolled images varied on the relevant (rotated) dimension rather than the irrelevant (expressive) dimension, and this may be seen at the most extreme angle of rotation tested here (30°).



In a Nutshell: Variation via Different Poses at Enrolment

Rotation training – improves transparency but slim benefits to performance

ROTATION TRAINING PROVIDED NO OVERALL BENEFIT WHEN CONSIDERING THE CONTINUOUS MATCH SCORES.

THIS LACK OF BENEFIT MAY OCCUR BECAUSE THE HIDDEN LAYERS IMPLICITLY WEIGHT THE DIMENSIONS THAT VARY WITHIN THEIR WORLD VIEW ANYWAY, ROTATION BEING ONE OF THESE.

TRANSPARENCY IS, HOWEVER, GAINED (WITHOUT SYSTEMATICALLY HARMING PERFORMANCE) WHEN WE BUILD HUMAN-LIKE COMPUTING INTO THE ALGORITHMS.

Simulating the Human Weighting of Internal Features

A third human-like strategy that may be of value within the machine algorithm is the tendency of humans to concentrate on the internal features of a face over the external or more changeable features. This human tendency may reflect an innate interest or need to attend to the more communicative features of the face. Equally, it may reflect a reliance on more stable and thus more diagnostic features of the face.

Whilst neural networks may implicitly learn the diagnostic features, transparency is often lost within a neural network or black box system. Accordingly, the purpose of the next test was to determine the impact of an explicit technique designed to differentially weight the internal features over the external features.

This test was completed using a non-specialist image matcher (Picture Matching Function) which is a third-party Matlab function (see https://uk.mathworks.com/matlabcentral/fileexchange/5456-picture-matching-function). The image matcher could be applied to segmented areas of the facial image, and the results combined across the segments in a way that allowed differential weighing of those segments.

Two variables were manipulated. First, three levels of weighting were used (baseline (unweighted), Fade 10, Fade 20). These varied in the extent of internal feature weighting such that Fade 20 placed more emphasis on internal than outer regions (and thus applied more of a fade) compared to Fade 10. The unweighted condition provided no weighting at all and acted as a baseline point of comparison. Second, 7 different grids were applied to the facial image (5x5, 7x7, 9x9, 11x11, 13x13, 15x15, 17x17). The finer the grid, the smaller each segment of the face, enabling the differential weighting of more tightly defined facial regions.

Determining the Weightings

The matrix of weightings was determined as follows: First, the T region corresponding to the eyes and nose (down to the mouth) was identified within a cropped Viola-Jones image. With a grid of n x n squared overlaying the Viola-Jones image, the nose always coincided with the central vertical column of squares.

The T-region of squares was then identified and assigned a weighting of 1. The squares around that T-region formed layer 1, with each square earning an incremental weight according to how many of the T-region squares (horizontal, vertical, or diagonal) it bordered. The squares around layer 1 formed layer 2, with each square again earning an incremental weight according to the summed weights of the layer 1 squares it bordered. Additional layers were identified in the same way until all squares in the grid were weighted.

Finally, the weighting in each square was divided by either 10 (for Fade 10 weightings) or by 20 (for Fade 20 weightings). In this way, the central T-region was always weighted highest, with a reduced weighting radiating out from that T reflecting a lesser (Fade 10) or greater (Fade 20) reduction in weighting.

Example Grid (7x7) with Fade 10 weightings

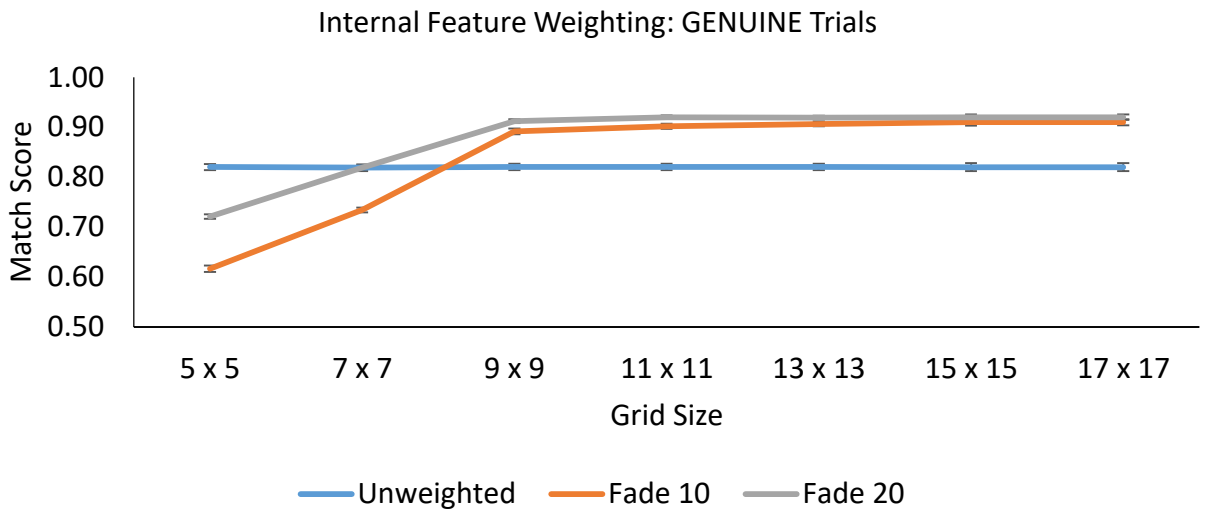
0.1	0.2	0.3	0.3	0.3	0.2	0.1
0.1	1	1	1	1	1	0.1
0.1	0.2	0.5	1	0.5	0.2	0.1
0.03	0.14	0.3	1	0.3	0.13	0.03
0.03	0.08	0.3	1	0.3	0.08	0.03
0.013	0.05	0.2	1	0.2	0.05	0.013
0.008	0.03	0.1	0.1	0.1	0.03	0.008

Internal Feature Weighting: Continuous Match Scores

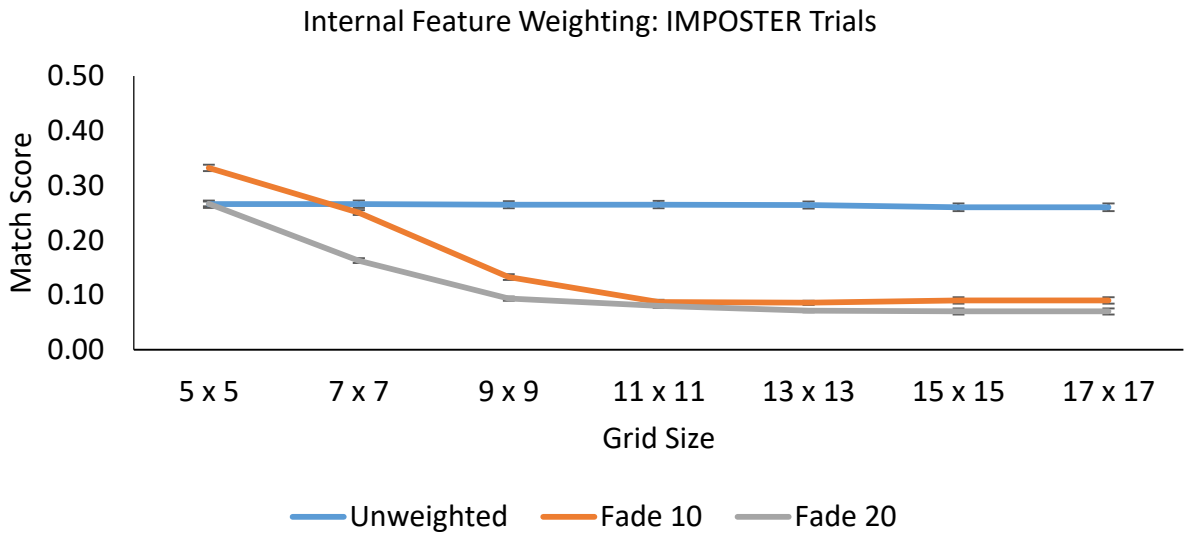
The continuous match scores were generated by calculating a weighted score with internal features weighted according to the baseline, Fade 10 or Fade 20 protocols. These scores varied between 0 and 1, with higher scores reflecting a greater similarity between target and test image. Performance was assessed when the target and test represented the same person (GENUINE trials) and when they represented different people (IMPOSTER trials).

On both GENUINE and IMPOSTER trials, there was a significant improvement with increasing grid size and with increasing level of fade. These two factors interacted, reflecting an improvement with grid size when internal feature weighting was applied, but not when it wasn't.

On GENUINE trials, optimal performance was achieved with a grid size of 11 x 11 and maximal internal feature weighting (Fade 20). Beyond this point, performance did not significantly improve.



On IMPOSTER trials, Optimal performance was achieved with a grid size of 13 x 13 and maximal internal feature weighting (Fade 20). At this point, performance was still better than that associated with smaller grids or less weighting.



The results of the internal feature weighting study were particularly successful. They indicated that an explicit modification of a matching algorithm to mimic the human weighting of internal features substantially improved performance.

Whilst neural networks may implicitly achieve this internal feature weighting to emphasise diagnostic features of the face, the current study provided clarity through transparency. A strong endorsement was given to the human-like strategy of weighting internal facial features.

In human terms, the resultant performance benefit may reflect either the consistency, individuality or communicative value associated with the internal features, and it is difficult to separate out these competing explanations. However, the fact that the machine benefits from attention to the internal features despite its ambivalence to communicative signals suggests that the performance advantage is likely to rest on consistency or individuality rather than communicative value.

In a Nutshell: Summary of Benefit of Internal Feature Weighting

A clear benefit emerged through human-like weighting of internal features

THE RESULTS WERE UNEQUIVOCAL IN SHOWING AN ADVANTAGE WHEN A SIMPLE IMAGE MATCHER WORKED WITH FACIAL IMAGES, AND MEANINGFUL INTERNAL FEATURES WERE EMPHASISED.

EFFECTS WERE STRONGEST WHEN THE WEIGHTING WAS STRONGEST.

BENEFITS PLATEAUED AT A GRID SIZE OF 13 X 13 BEYOND WHICH THE GRANULARITY OF SEGMENTS STARTED TO LOSE CORRESPONDENCE WITH THE INTERNAL FEATURES THEMSELVES.

Simulating the Human Capacity for 3D Rotation

A final strategy that humans use when matching faces is the capacity to use their knowledge of faces as 3D objects in order to mentally rotate a face and imagine it from a different angle. Human observers may achieve this mental rotation by being able to generate what is known as a '2½ dimensional representation' from a flat 2d image, by incorporating knowledge about the volume, depth and surface characteristics associated with faces and their features. This may enable the human perceiver to be more tolerant than a computer algorithm to changes in facial pose or orientation.

Accordingly, the final test conducted here mimicked the approach of building in human-like capability to mentally rotate a target or test face so that their orientation is equivalent prior to matching.

This question was tested using Neurotechnology and Dlib only as it was only these algorithms that were capable of performing a one-to-one match. Two discrete test scenarios were presented. In the first scenario, a 3D head was generated from a full-frontal enrolled image. The resultant 3D head was then rotated to match a 10° angled test image ('rotate full-frontal enrolment'). In the second scenario, a 3D head was generated from the angled test image which was then rotated back to full-frontal to match the standard enrolled image ('rotate angled test').

Performance in these two test scenarios was compared to the baseline matching performance which involved matching a full-face enrolment image to an angled test ('unrotated angled test'). As in all experiments, performance was examined on GENUINE trials and IMPOSTER trials.

Generation of the 3D Head

The 3D head was generated using an Open Source Python algorithm called Face3D (see https://github.com/YadiraF/face3d/blob/master/3D%20Face%20Papers.md). Uniquely, this software was capable of generating a 3D model of a head from a *single* enrolment image making it extremely efficient at a computational level. Indeed, it sits in stark contrast to the growing literature which relies on an intensive set of enrolment images to generate a 3D model of an individual.

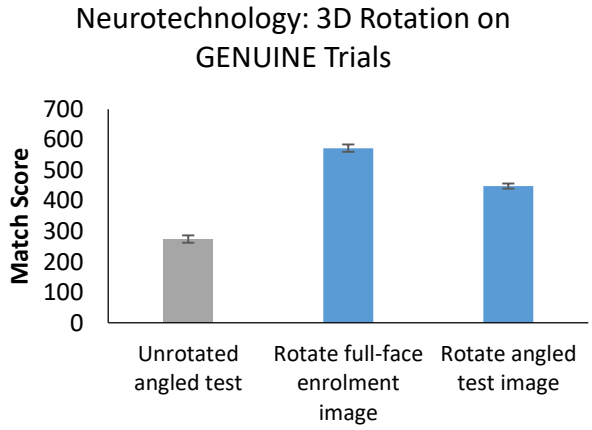
Within the current study, it is worth noting that we do not report on 3D comparison between an enrolled individual and a test individual. To do so would take time and computational power. Instead, we generate a 3D model of either the enrolled image or the test image, rotate the resultant model to match the pose of the comparison face, and then perform a 2D comparison on the aligned images.

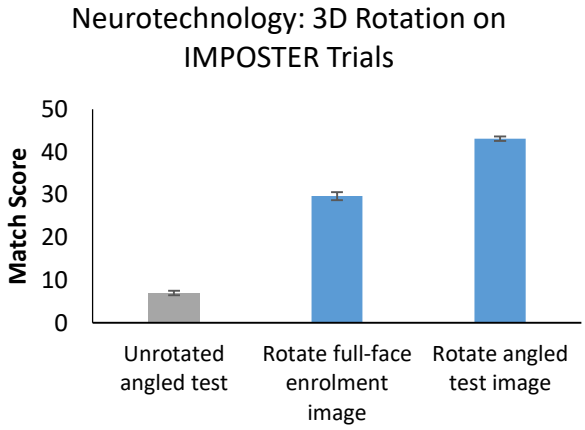
3D Rotation: Continuous Match Scores

Both algorithms showed some benefit of 3D rotation, although there were subtle differences in the pattern of performance across the algorithms.

Taking the Neurotechnology algorithm first, the data suggested a benefit in the form of higher match scores on GENUINE trials when one or other image had been rotated such that the pair was aligned prior to comparison. This benefit was greatest when the 3D model had been generated from the full-face enrolment image.

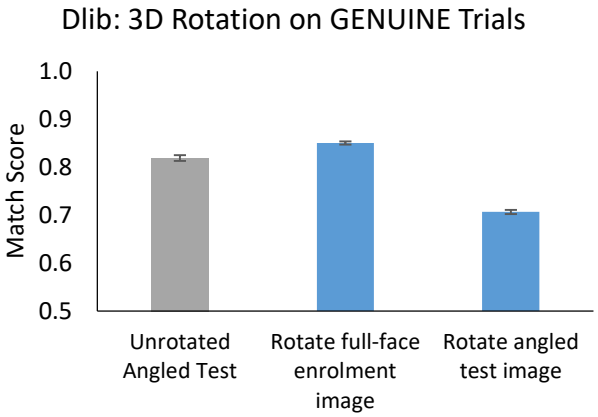
In contrast, rotation did not help on IMPOSTER trials. In fact, both rotated test scenarios produced worse performance (inappropriately higher match scores) when rotation was involved and this was most probably because rotation removed one source of difference between the enrolled image and the test image.

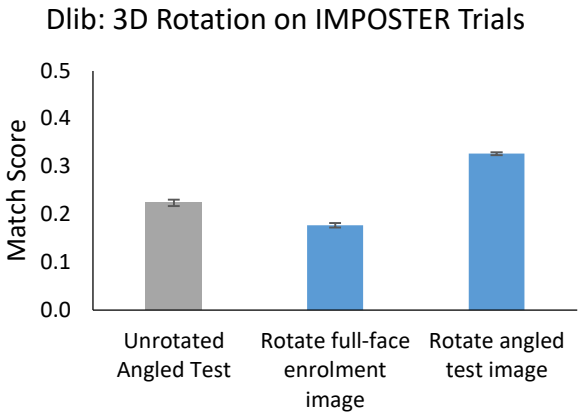




When considering the performance of the Dlib algorithm, the data again suggested a benefit when 3D rotation was used in GENUINE trials, but only when the 3D model had been generated from the more informative full-face enrolment image. For this algorithm, rotation from an angled test image prior to matching did not assist in the matching process, suggesting that the Dlib algorithm may be somewhat more sensitive to angle.

In IMPOSTER trials, rotation helped dismiss the imposter when the 3D model was generated from the full-face enrolled image, but it hindered dismissal of the imposter when the 3D model was generated from the angled test image. Taking these results together, the only scenario in which there was a reliable benefit to performance was that which involved 3D rotation of a full-face enrolled image prior to a 2D matching process.





In a Nutshell: Summary of 3D Head Rotation Prior to Matching

A benefit emerged if a full-face enrolled image could be rotated to match an angled test prior to matching.

ROTATION OF ONE IMAGE TO MATCH THE ANGLE OF ANOTHER IMAGE BENEFITTED MATCHING ON 'GENUINE' TRIALS BUT REMOVED ONE SOURCE OF DIFFERENCE BETWEEN IMAGES AND THUS WEAKENED PERFORMANCE ON 'IMPOSTER' TRIALS.

ROTATION WAS BEST ACHIEVED USING A 3D MODEL DERIVED FROM A FULL-FACE ENROLMENT IMAGE RATHER THAN FROM AN ANGLED TEST IMAGE.

THE BEST APPROACH WOULD BE TO REQUIRE A FULL-FACE ENROLMENT IMAGE FROM WHICH A 3D MODEL CAN BE GENERATED. THIS WOULD ENABLE THE ENROLLED MODEL TO BE ROTATED TO MATCH A LESS CONSTRAINED TEST IMAGE.

Conclusion

In summary, a considerable body of work has been presented which sought to build human-like strategies into computer algorithms for face matching. Our purpose was to determine which, if any, of these human-like strategies may improve transparency or capability within algorithmic performance. The results suggested that performance was optimal when more than one image was provided at enrolment (ideally in the form of a set of images rather than a single composite). Variety at enrolment improved transparency but did not appear to improve performance. However, two factors provided notable benefit to matching capability. These involved (i) the use of differential weighting to focus on informative internal features and (ii) the 3D rotation of a full-face enrolled image to match an angled test image.

One question has been beyond the scope of the present feasibility study is whether these two strategies exert a greater benefit when used in conjunction with one another. Further work on this point is encouraged.

References

Ahonen, T., Hadid, A., and Pietikainen, M. *Face Recognition with Local Binary Patterns.* Computer Vision - ECCV 2004 (2004), 469–481.

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A.M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68(10)*, 2041-2050. DOI: 10.1080/17470218.2014.103949

Bruce, V. (1994). Stability from variation: The case of face recognition the MD Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology*, *47*(1), 5-28.

Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256-284.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577-596.

Klare. B., (2011). Spectrally sampled structural subspace features (4SF). In Michigan State University Technical Report, MSUCSE-11-16.

Longmore, C. A., Liu, C. H., & Young, A. W. (2015). The importance of internal facial features in learning new faces. *The Quarterly Journal of Experimental Psychology*, *68*(2), 249-260.

Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, *36*(12), 1761-1771.

Turk, M., & Pentland. A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 71-86.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171-211). Springer, Boston, MA.