Forensic Science International 286 (2018) 8–11

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Technical Note

Biometric correspondence between reface computerized facial approximations and CT-derived ground truth skin surface models objectively examined using an automated facial recognition system

Connie L. Parks[a], Keith L. Monson[b,*]

[a] Counterterrorism and Forensic Science Research Unit, Visiting Scientist Program, FBI Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135, United States
[b] Counterterrorism and Forensic Science Research Unit, FBI Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135, United States

ARTICLE INFO

Article history:
Received 15 December 2015
Received in revised form 14 June 2016
Accepted 18 February 2018
Available online 27 February 2018

Keywords:
Forensic anthropology
Facial approximation
Facial recognition
Unidentified decedents
Missing persons
Biometrics

ABSTRACT

This study employed an automated facial recognition system as a means of objectively evaluating biometric correspondence between a ReFace facial approximation and the computed tomography (CT) derived ground truth skin surface of the same individual. High rates of biometric correspondence were observed, irrespective of rank class (R_k) or demographic cohort examined. Overall, 48% of the test subjects' ReFace approximation probes (n = 96) were matched to his or her corresponding ground truth skin surface image at R_1, a rank indicating a high degree of biometric correspondence and a potential positive identification. Identification rates improved with each successively broader rank class (R_{10} = 85%, R_{25} = 96%, and R_{50} = 99%), with 100% identification by R_{57}. A sharp increase (39% mean increase) in identification rates was observed between R_1 and R_{10} across most rank classes and demographic cohorts. In contrast, significantly lower ($p < 0.01$) increases in identification rates were observed between R_{10} and R_{25} (8% mean increase) and R_{25} and R_{50} (3% mean increase). No significant ($p > 0.05$) performance differences were observed across demographic cohorts or CT scan protocols. Performance measures observed in this research suggest that ReFace approximations are biometrically similar to the actual faces of the approximated individuals and, therefore, may have potential operational utility in contexts in which computerized approximations are utilized as probes in automated facial recognition systems.

1. Introduction

Computerized facial approximation systems, once tepidly described as 'promising' by early researchers [1–3], have matured significantly beyond these early indications of feasibility [4–8]. A number of recent research studies suggest that computerized facial approximations systems have advanced to a state of potential utility in real-world operational contexts [9–14]. Despite a relatively mature state of development, the accuracy, precision, and utility of computerized facial approximations is still under examination [4,7,8].

Non-operational research examinations intended to ascertain the accuracy of computerized facial approximations fall into two general categories: (i) subjective evaluations involving non-expert human adjudicators and (ii) objective evaluations intended to quantify metric differences between an approximation and the ground truth skin surface of the same individual. Although research studies using non-expert human adjudicators are informative, the highly subjective nature of the evaluations renders many results only marginally useful. Conversely, the results produced from objective analyses provide functional, predictive data useful for indicating the potential performance of computerized facial approximations.

Recent advances in facial recognition algorithms afford researchers a unique opportunity to objectively examine the accuracy, precision, and potential utility of computerized facial approximations [9–12]. Given the results of recent studies indicating a high degree of metric similarity between computerized facial approximations and ground truth skin surface models [6,9,13], and expanding on research previously reported by the authors, this study employs an automated facial recognition system (NeoFace® Reveal, NEC Corporation of America, Irving, TX) as a means of objectively evaluating biometric correspondence between a ReFace approximation and the ground truth skin surface of the same individual. The results of this study will: (i) contribute to the existing literature examining the accuracy of computerized

* Corresponding author.
E-mail address: Keith.Monson@ic.fbi.gov (K.L. Monson).

https://doi.org/10.1016/j.forsciint.2018.02.019
0379-0738/

facial approximations, (ii) further evaluate the utility of facial approximations in operational contexts, (iii) promote development of computerized facial approximation systems, and (iv) address the first and second aims of facial approximations as defined by SWGANTH [15].

2. Materials and methods

The research model employed in this study has been reported by Parks and Monson [10] and is therefore, only briefly discussed here for the purpose of pointing out certain methodological differences. Methodological aspects not expressly discussed were conducted in adherence to the research model reported by Parks and Monson [10]. Given that the baseline performance for the facial recognition system employed in this study has been established in two prior studies [10,11], no internal control was conducted. In contrast to the methodology of Parks and Monson [10]: (i) only a blind search against the entirety of the gallery was conducted for each test subject's approximation probe and (ii) the approximation probe set was comprised of only average weighted approximations without visible eyes (Fig. 1, gallery sizes and approximation types further defined in Ref. [10]).

The complete image gallery (g = 6159) previously constructed by Ref. [10] was used in this study. The gallery also contained one CT-derived ground truth skin surface image (Fig. 2) of each test subject (t = 96), producing a final gallery size of g = 6255. The test subjects' ground truth skin surface images were used as life photo surrogates in the final image gallery. Two scanning protocols (SP) were present in the CT dataset: (SP3.5) 3.5 mm slice thickness, 1.7 mm slice increments, 512 × 519 resolution, and 0.586 pixel size and (SP6.0) 6.0 mm slice thickness, 3.0 mm slice increments, 512 × 512 resolution, and 0.586 pixel size. Following segmentation of the 3D skin surface model from the CT scan, a single Frankfurt aligned 2D image was exported for each test subject and subsequently enrolled into the aforementioned gallery. The test subjects (t = 96), from which both the facial approximation and ground truth skin surface images were generated, were a subset of the original test subject sample (t = 159) compiled by Parks and Monson [10]. The test subjects (t = 96) ranged in age from 18 to 60 years (mean: 31, median: 27) with representation of both sexes (M = 47, F = 49) and three self-identified ancestral groups (African–American, European–American, and Hispanic). The only modification applied to the approximation probes and the ground truth skin surface images was a 'crop' intended to remove excess background and produce a more 'head and shoulders' image composition (see Ref. [10] for detailed explanation of the crop procedure). The protocols for collection and use of images were approved for use by the Institutional Review Board of the FBI. Protection of the anonymity of the living test subjects precludes publication of both life photos and ground truth skin surface images. The skin surface images depicted in Fig. 2 are artist-rendered composites of multiple skin surfaces images and do not reflect the actual visages of living individuals. These models are provided to demonstrate the appearance of the CT-derived skin surface gallery images discussed above.

Following Parks and Monson [10], performance metrics for biometric correspondence are presented in terms of candidate list inclusion rates. The candidate list inclusion rate is expressed as a fraction of total test subjects (t) who were identified within a specific rank class (R_k) and estimates the probability of obtaining a correct match within the top k images in a candidate list. For example, if in a blind search R_{10} = 82 and t = 96, the estimated identification probability is R_{10}/t, or 0.854, for this specific search type (blind) and rank class (R_{10}). Performance metrics are offered for both sexes, three ancestral cohorts, two CT scan protocols, and three age groups (see Table 1 for age group details). Fisher's Exact tests were utilized to analyze performance and cohort relationships. A sequential Holm-Sidak procedure [16] was employed to control for the possible effect of simultaneous comparisons (>2 groups). Statistical significance was assessed at $\alpha = 0.05$. Data were analyzed using SPSS 16.0 (Chicago, IL) and Microsoft Excel 2007 (Redmond, WA).

3. Results and discussion

Performance metrics by sex, ancestry, age group, and CT scan protocol are detailed in Table 1. Overall, 48% of the test subjects' approximation probes (n = 96) were matched to his or her corresponding ground truth skin surface image at R_1, a rank indicating a high degree of biometric correspondence and a potential positive identification. Candidate inclusion rates improved with each successively broader rank class (R_{10} = 85%, R_{25} = 96%, and R_{50} = 99%), with a 100% inclusion rate by R_{57}. In other words, across all test subjects (t = 96) the highest (worst) rank match between an approximation and a corresponding ground truth skin surface image in a gallery of 6255 images was at candidate position 57. With the exception of the underrepresented age group C (n = 8, Table 1), a similar pattern of performance was observed across all rank classes, demographic cohorts, and CT scan protocols. A sharp increase (39% mean increase) in candidate list inclusion rates was observed between R_1 and R_{10} across most rank classes and demographic cohorts (age group C not included in the mean increase calculations). In contrast, significantly lower ($p < 0.01$) increases were observed between R_{10} and R_{25} (8% mean increase) and R_{25} and R_{50} (3% mean increase), even though R_{25} and R_{50} represent broader rank classes. Excluding the underrepresented age group C (n = 8), no significant performance differences were observed across the demographic cohorts or CT scan protocols, or both. Significance analyses results between ancestral cohorts further discriminated by CT scan protocol, however, should be evaluated with circumspection given the disparate ancestral representation between the two protocols (e.g., no Hispanic

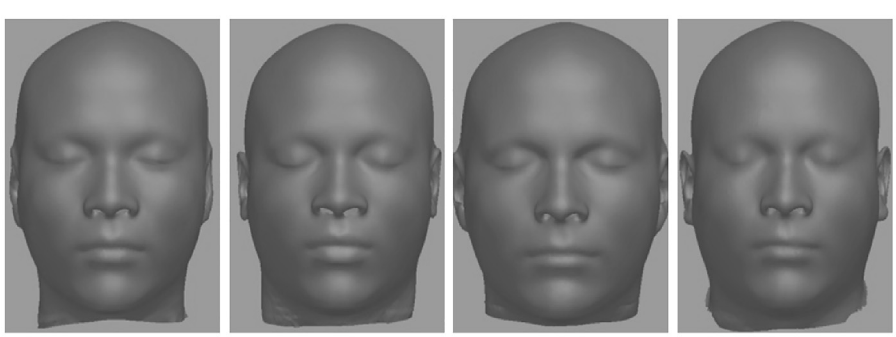

Fig. 1. ReFace approximation probe examples.