Forensic Sci Rev 30:23–34; 2018

Forensic Face Recognition as a Means to Determine Strength of Evidence: A Survey

C. G. Zeinstra[1], D. Meuwly[1,2], A. C. C. Ruifrok[2]
R. N. J. Veldhuis[1], L. J. Spreeuwers[1]

[1] Faculty of Electrical Engineering
Mathematics and Computer Science
University of Twente
Enschede
The Netherlands

[2] Netherlands Forensic Institute
The Hague
The Netherlands

TABLE OF CONTENTS

* Corresponding author: Dr. Chris Zeistra, ????????(Address);
?????????(Phone number) (voice); c.g.zeistra@utwente.nl.

1042-7201/30-1/Jan. 2018/23–34/$18.00 • Copyright © 2018 Central Police University Press

Forensic Face Recognition as a Means to Determine Strength of Evidence: A Survey

REFERENCE: Zeinstra CG, Meuwly D, Ruifrok ACC, Veldhuis RNJ, Spreeuwers LJ: Forensic face recognition as a means to determine strength of evidence: a survey; *Forensic Sci Rev* 30:23; 2018.

ABSTRACT: This paper surveys the literature on forensic face recognition (FFR), with a particular focus on the strength of evidence as used in a court of law. FFR is the use of biometric face recognition for several applications in forensic science. It includes scenarios of ID verification and open-set identification, investigation and intelligence, and evaluation of the strength of evidence. We present FFR from operational, tactical, and strategic perspectives. We discuss criticism of FFR and we provide an overview of research efforts from multiple perspectives that relate to the domain of FFR. Finally, we sketch possible future directions for FFR.

KEYWORDS: Facial biometrics, FISWG, forensic face recognition (FFR) approaches, forensics.

INTRODUCTION

In this survey paper, we present different aspects of forensic face recognition (FFR), with a particular emphasis on strength of evidence. The aim of this paper is to convey the breadth of FFR, with its many aspects and connections to related domains.

FFR is the use of biometric face recognition for several applications in forensic science. Biometric face recognition uses the face modality as a means to discriminate between human beings; forensic science is the application of science and technology to law enforcement.

In general, FFR includes scenarios of ID verification (1:1) and open-set identification (1:N+1), investigation and intelligence (M:N+1), and evaluation of the strength of evidence as described in Meuwly and Veldhuis [42]. There are two image types involved in FFR. The trace image often captures a crime scene and is most often taken under uncontrolled conditions. The reference image is a photograph of a suspect and is taken under controlled conditions. Concrete FFR use cases are given in Zeinstra et al. [73].

A use case in which FFR is frequently employed is to investigate criminal activities that are carried out in places monitored by surveillance cameras, like shops or gas stations. Extracted stills from closed-circuit television (CCTV) recordings that contain the face of the perpetrator are used as trace images. Another example is the withdrawal of money using a stolen debit card. In this case, trace images are recorded by a small camera inside an automated teller machine (ATM) and they typically exhibit perspective image distortion. These use cases are examples of investigation (M:N+1) or, in the case of a concrete suspect, examples in which the strength of evidence against that suspect is evaluated. Another example is when an immigration officer might be convinced that a given identity document is genuine, but that it does not correspond to the person who is presenting it. If the immigration officer forbids the person to enter, the subsequent investigation is an example of evaluation of strength of evidence in which the passport photograph serves as a trace image.

A survey by Jain et al. discusses additional open-set investigation (M:N+1) use cases: (a) mug shot search that is robust to facial aging, (b) matching forensic (composite) sketches to face photograph databases, and (c) retrieval using facial scars and marks [33]. Case (b) is an example in which trace images consists of a representation (sketch) by an image, instead of a captured image.

A final, very noteworthy but rather extreme, example of an FFR use case (M:N+1) is the “super recognizers” [57] at the London (UK) Metropolitan Police. Super recognizers are claimed to be able to identify persons from CCTV footage, based on an exceptional memory for discriminating facial features in previously seen low-quality images. Super recognizers were used for example during the London riots of 2011 [76].

FFR has its modern genesis in the Bertillonage system [4]. Bertillonage systematically uses facial and body features to describe criminal individuals. It features anthropometric measurements, as well as categorizations of facial features; for example, it recognizes 16 different ear shape types. Also, highly discriminating features like facial marks can be described. **Figure 1** depicts some examples.

Bertillon particularly advocates for a mugshot from face and profile, enhancing that the profile contains information that is in the same time more distinctive and less subject to intra variability (ear, upper profile), that has been forgotten in the modern mugshot process. Finn gives a historical account of Bertillonage; in particular the use and acceptance of photography (“the criminal image”) as a means to represent information and evidence [26]. Bertillonage as such has been superseded as a means to individualize persons by fingerprinting (and DNA profiling in the last 25 years) [16].

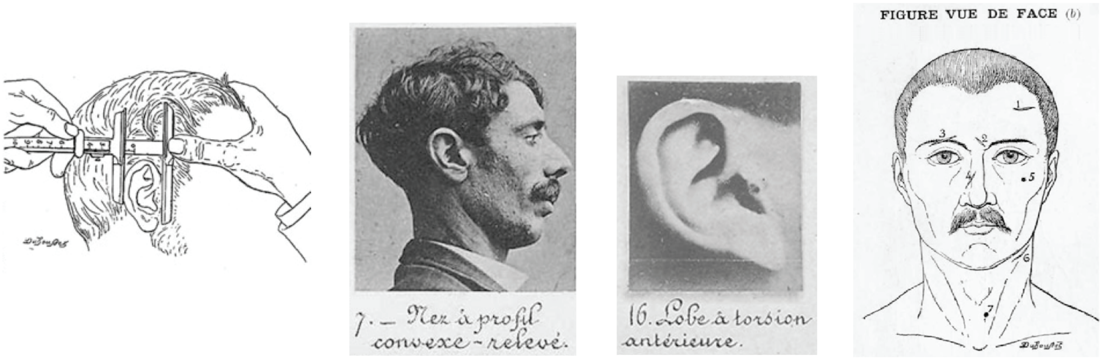
7. — Nez à profil convexe-relevé.
16. Lobe à torsion antérieure.
FIGURE VUE DE FACE (b)
1
2
3
4
5
6
7

Figure 1. Example measurements and categorization of the Bertillonage system. From left to right: ear size measurements, nose and ear profile categories, and the location of scars and marks. Image used with permission from [4].

However, with the proliferation of cameras, either CCTV cameras or digital cameras in mobile phones, potential trace images are omnipresent; it is hardly a surprise that FFR in general is very actively used today.

Despite the advances in automatic face recognition systems, significant parts of FFR are still done manually, notably the evaluation of the strength of evidence. We believe that this can be attributed to several factors that are inherent to FFR and that influence the performance of such automated face recognition systems. Although one can classify these factors into, for example, subject and imaging conditions, there might be some overlap in this classification. Therefore, we use an example to illustrate the cascading effect of these factors.

Assume a perpetrator uses a stolen debit card to withdraw money from an ATM. He has a hurried glance (expression), does not look straight into the camera (pose), and wears a scarf (occlusion). Natural light (illumination) comes from behind the perpetrator. He stands close to the ATM (perspective effect). The scene is captured through a lens (distortion and aberration) by a recording device, either analog (interlacing) or digital (sensor thermal noise/bleeding). The resulting raw material is then possibly converted (analog to digital, resolution, compression artifacts) and extracted (motion blur). The pose, illumination, and expression (PIE) of the person are a common problem for face recognition systems [33].

In the remainder of this paper we will focus mainly on the strength-of-evidence scenario. The paper is structured as follows. In Section I we describe the operational level of FFR structured in terms of the Analysis, Comparison, Evaluation, and Verification (ACE-V) protocol. This protocol is commonly used for source-level inference in forensic science (Langenburg [35] and Finding 22 of Prince [52]). In Section II we present FFR at a tactical and strategic level. During the last decade, criticisms of forensic science in general and FFR in particular became more visible at the public level, but they are much older than the last decade within the profession (see [40] for a discussion). In Section III we address some of these criticisms, and in Section IV we explore research efforts pertaining to FFR. Finally, we present our conclusion and sketch future directions for FFR.

As a final note, in this paper we consistently use the general term *FFR-examiner,* or *examiner* for short, to refer to any individual that undertakes FFR activities, disregarding their level of proficiency; we use *nonexaminer* to refer to individuals that do not undertake FFR activities.

I. OPERATIONAL LEVEL OF FORENSIC FACE RECOGNITION

Prince gives an overview of the use of facial recognition systems and facial image comparison procedures at several forensic institutions in Israel, the Netherlands, the United Kingdom, the United States, and Canada [52]. Although differences exist, the institutions have a significant collective approach. Spaun specifically describes the approach taken at the US Federal Bureau of Investigation [62] and is in line with Prince.

A. ANALYSIS AND COMPARISON

During the Analysis phase, the trace is investigated for its usability and its evidential content. During the Comparison phase the trace and reference images are compared. In this process the examiner takes similarities and dissimilarities between trace and reference material into account. According to Spaun, the examiner distinguishes between class and individual characteristics [62]. However, one could argue there are characteristics that are distinctive, contain information to differentiate people and characteristics who are not, and the degree of distinctiveness varies; hence the class/individual distinction might be too strict.

The Facial Identification Scientific Working Group (FISWG) is an organization centered on scientific knowledge and casework experience [19]. According to FISWG guidelines [22], mainly four FFR methods can be used during the Analysis and Comparison phase:

- Holistic Comparison. Assessment in which all facial features are considered at once. This is a pure perceptual approach without real analysis.
- Morphological Analysis. Assessment of correspondence of the shape, appearance, presence, and/or location of facial features.
- Photo-anthropometry. Assessment of correspondence of dimensions and angles of landmarks and other facial features.
- Superimposition. Assessment in which images are aligned and analyzed using image transitions.

FISWG recommends morphological analysis by trained examiners as the primary method of comparison; superimposition is known to be inefficient and should only be used in conjunction with morphological analysis and is to be confined to rotations, scaling, and translation.

By definition, in holistic comparison and superimposition, trace and reference images are considered simultaneously, so the Analysis and Comparison phase partly overlap. This might lead to bias and is further discussed in Section IV.A. In contrast, when applying morphological analysis or photo-anthropometry, trace and reference images do not have to be considered simultaneously.

FISWG has also published the Facial Image Comparison Feature List for Morphological Analysis [20]. This feature list describes for each facial part a number of characteristic descriptors in considerable detail that can be used during forensic casework. As such, it is not a standard but rather representative of considered facial features of different institutes. Feature lists used at the NFI (Netherlands) and NFC (Sweden) are similar but differ on the type and number of considered details [17,18]. They include:

- Face shape;
- Eyebrows;
- Cheek area;
- Mouth and mouth area;
- Chin; and
- Forehead;
- Eyes;
- Nose;
- Jawline;
- Scars, marks, tattoos

B. Evaluation and Verification

One notable variation between forensic institutions is which opinion scale they use (Finding 31, Prince [52]). In an ideal situation this opinion scale is the strength of evidence that is determined during the Evaluation phase. The strength of evidence is constructed from multiple comparisons between features found in trace and reference images. Essentially one could adopt either a numerical or a verbal formulation of strength of evidence. An example of the former is "the trace and reference specimens are 1 million times more likely under the hypothesis that the suspect is the donor of both, than under the hypothesis that the suspect is not the donor of both." An example of the latter is "there is strong support for the hypothesis that the suspect is the donor of both the trace and reference specimens against the hypothesis the suspect is not the donor of both the trace and reference specimens".

The strength of evidence should be expressed in terms of the (magnitude of the) *likelihood ratio*. Loosely speaking, the likelihood ratio measures the probability of observing *(dis)similarities* between a particular feature found in trace and reference image, scaled by the probability of the *typicality* or *rarity* of observing this feature in trace and reference images in general. The problem with the verbal description is that it can be understood as a verbal scale of posterior probabilities, and it is often the case because of lack of knowledge from both the examiners and the requesters.

An element to take into account is how to *combine* the likelihood ratios of several facial features into *one* likelihood ratio that is reported as the overall strength of evidence. One obvious approach is to assume statistical independence between facial features. In that case, the corresponding overall likelihood ratio reverts to the product of the underlying likelihood ratios. However, the independence assumption is most often wrong and will be immediately pointed at in court. Approaches that capture or model the dependency structure between features are copula or Bayesian Belief Networks (BBN) and can be used for FFR. A copula describes the relationship between the multivariate distribution of (in our case facial) features and their marginal distributions; Sklar proves that this relationship holds for any multivariate distribution [60]. Another popular approach involves BBNs (see Barber [3]) in which a domain expert models dependency structures. The major advantage of BNN is the significant reduction of the dimensionality of the feature space, making inferences feasible with the limited quantity of data available in a forensic case.

In the verbal formulation the strength of evidence is expressed in a standardized verbal description. The advantage is that a court of law can understand the outcome of the evaluation in principle; however, it must be clear that this evaluation is not a posterior odds and as such it can be easily misunderstood. The verbal approach is either a result mapped from the numerical strength of evidence or a protocol that uses criteria to determine the verbal strength of evidence directly, such as "Guidance for Evaluating Levels of Support” [5] or “Guideline for Evaluative Reporting in Forensic Science” [14].

During the Verification phase, one or more of the ACE steps are repeated independently in order to reduce the human factor (see Section IV.A for a further discussion on this topic). According to a private communication with an FFR-examiner, the ACE steps can be performed independently by three examiners, after which the final evaluation is determined by a consensus model [18].

The final, verified, evaluation outcome is reported to a court of law. In some cases, an FFR-examiner will witness during a court session when additional information or clarification is needed.

II. TACTICAL AND STRATEGIC LEVELS OF FORENSIC FACE RECOGNITION

A. Recommendations and Working Groups

European institutes are organized in the European Network of Forensic Science Institutes (ENFSI) [13]. This organization has published a general guideline regarding evaluative reporting [14]. Forensic facial expertise is organized in the Digital Image Working Group (ENFSI-DIWG).

As mentioned previously, FISWG is an organization in which the FBI and several other forensic institutes from the United States and other countries participate [19]. They have published recommendations on facial comparison [22] and on which features should be considered during casework [20]. Recent additions (in draft status) are image-processing steps for the improvement of automated facial recognition searches [23] and the physical stability of adult facial features [24].

B. Levels of Expertise, Training, and Proficiency Tests

Most agencies considered in Prince [52] recognize three proficiency levels. The *foundation level* means that the examiner has had basic training and can only do verification (ACE-**V**). The *advanced level* means that the examiner has had more training and experience to do the full ACE-V process. The *expert level* means that the examiner operates at an advanced level and may give testimony in court. It is remarkable that, at least with respect to the Dutch situation, FFR-examiners are *not* yet included in a national register of forensic experts [46] that does include, for example, forensic psychiatrists.

Training differs per institute but according to [21,35] may involve

- Knowledge of relevant recommendations;
- Competence in quality assessment (interlacing, codecs, compression, lens distortion, etc.) of trace material;
- Competence in extraction of facial images from CCTV;
- Competence in Adobe Photoshop or similar software;
- Knowledge of and competence in image processing techniques like image enhancement;
- Competence in facial comparison, notably anatomical knowledge and considered facial features in morphological analysis;
- Knowledge of standardized evaluation and reporting;
- Awareness of possible bias and other human errors;
- Competence in statistical concepts, notably Bayesian statistics; and
- Basic knowledge of legal aspects and competence in expert testimony.

Apart from initial training to obtain the competences to practice, examiners should participate in proficiency tests on a regular basis [25]. In a recent ENFSI-DIWG Facial Image Comparison Proficiency Test, FFR-examiners had to compare CCTV footage with 10 reference images. For 17 comparisons a single conclusion had to be reported, whereas in one case a full report using the ENFSI evaluative reporting guideline [14] had to be handed in; results have been discussed with peers within the ENFSI-DIWG.

III. CRITICISM OF FORENSIC FACE RECOGNITION

FFR has been criticized for its lack of scientific rigor. According to [16], little research is done on the validation of FFR (“there is no reported error rate (...)” both for human FFR and in assessing the claimed ability of “super recognizers”); as a field it is mostly not scientifically founded yet. This is reiterated by the FISWG guidelines [22] in which morphological analysis is coined as the primary comparison method, but “only limited studies have been done on accuracy or reproducibility”. Only in recent years, some validation studies have appeared and indicate that examiners are better than nonexaminers; see Section IV.B for some examples. Formally, human-based methods are not validated/accredited on basis of performance but of competence and proficiency; this provides some safeguards but less than a method validated/accredited on basis of performance. We refer to Meuwly et al. [41] for a full discussion of this topic and to Section IV.B for a discussion of the FFR-examiner as an expert.

Humans are subjective and it is partly mitigated by the verification step in the ACE-V protocol. However, notably the protocol for assigning strength of evidence is subjective (Mallett and Evison [38]) and the strength of evidence does not necessarily represent a likelihood ratio; that is, the strength of evidence.

Edmond et al. contains a complete and very critical review of examiner identification evidence based on trace images [12]. Their study presents several examples of FFR-examiner testimony that illustrate the nonscientific approach and the examiner as the single bearer of absolute truth. One poignant example is "(...) used photo-anthropometry, morphology and photo superimposition to make a positive identification (...). (...) unwilling to disclose her techniques, particularly the points she relied upon (...), because of concerns about her intellectual property rights".

Another example is "during cross-examination (...) rejected the suggestion that there was a degree of subjectivity in her assessment, i.e. morphological comparison".

These examples exactly show the lack of fundamental understanding of what inference in forensic science is. According to Saks and Koehler, "In normal science, (...) students receive four (…) years of doctoral training where much of the socialization into the culture of science takes place. This culture emphasizes methodological rigor, openness, and cautious interpretation of data. In forensic science, 96% of positions are held by persons with bachelor's degrees (or less), 3% master's degrees, and 1% Ph.D.'s" [58].

The criticism can be placed in the context of the elaborate and critical report [45] on the current state of forensic science in the United States by the National Research Council of the National Academies. One of their recommendations states that "research is needed to address issues of accuracy, reliability, and validity in the forensic science disciplines. (...)".

IV. FORENSIC FACE RECOGNITION RESEARCH DIRECTIONS

In this section we describe several research directions related to FFR. First, we discuss human and expert aspects of FFR-examiners. Another branch of FFR research is concerned with the use of anthropometry. Some FFR datasets are available for research purposes. Finally, several studies have considered using more or less distinctive features.

A. Human Aspect of the Forensic Face Recognition Examiner

The examiner has a pivotal role in FFR. In O'Toole et al. [49] and Spaun [61], the human aspect in FFR is described as being underestimated.

Recent experiments by Papesh and Goldinfer on face matching indicate that under realistic viewing conditions (for example, at an airport) infrequently occurring identity mismatches go undetected [50]. Results that relate to trace images taken under uncontrolled conditions are summarized by Sinha et al. as "people can recognize familiar faces in very low-resolution images" and "the ability to tolerate degradations increases with familiarity" [59]. In particular, Burton et al. reported that even under severely distorted CCTV footage, familiar faces can be recognized, but that does not hold for unfamiliar faces [8]. This might explain why super recognizers have such high success rates, since they "just" recognize a familiar face that they have seen before in other CCTV recordings. In Bruce et al., it is shown that recognition of unfamiliar faces is very error-prone, but this can be claimed of any perceptive intelligence [6,7]. Megreya and Burton reported that there are large individual differences on unfamiliar face matching [39]. Also, Gold et al. have stated that familiarity has a quantitative rather than a qualitative effect on the efficiency with which information is extracted from individual features [27].

Another well-studied negative effect in psychology and forensic science is that of confirmation bias and contextual information. A proper implementation of the ACE-V protocol, with the shield of the examiner from the unnecessary information during the A and C phase, helps to limit this effect. An overview by Pronin describes that people can recognize and estimate the operation of bias in human judgment of other persons, except when it is their own bias [53]. Dror et al. show the risks of contextual information and bias with respect to fingerprint examination, which could easily be extended to any other forensic modality, in particular FFR [11].

B. Expert Aspect of the Forensic Face Recognition Examiner

Norell et al. reported that on a set of image pairs, examiners reached their conclusions with a significantly lower number of errors than nonexaminers [48]. Also, if the quality of the trace was lowered, it led to more careful conclusions by examiners, but not for nonexaminers. We believe that both findings stem from the fact that the proper methodology is used.

Work with similar findings is White et al.: They administered several challenging face-matching tests to examiners and nonexaminers and concluded that examiners not only outperformed untrained participants, but also computer algorithms, thereby providing the evidence that these examiners are experts at this task [70].

Zeinstra et al. described an on-line experiment in which examiners and nonexaminers participate. Their task is to compare isolated eyebrow pairs using either a "best-effort" approach versus an approach that uses FISWG characteristic descriptors of the eyebrow. It was found that there are no significant differences in accuracy; however, the group of examiners performed significantly better than the nonexaminers when they used FISWG [74].

These results indicate that experts (a) are more aware of fallacies in their judgment and (b) have better judgment than untrained participants.

C. Anthropometry

Anthropometry is the science of measuring body or facial dimensions, notably distances and angles. Anthropometry is a key ingredient of the Bertillonage system.

In the dissertation of Kleinberg, a series of experiments using locations of anatomically defined facial landmarks is conducted and it is concluded that "using high resolution images to compare video images with photographic images, (...) anthropometry (...) does not generate the results necessary for use as evidence in a court of law" [34].

In a large-scale study by Evison and Vorder Bruegge concerning landmark-based analysis of 3D landmarks of more than 3,000 persons, it was found that "the 3D distribution of anthropometric landmarks (...) is unlikely to be sufficient to allow for identification of individuals (...)" [15].

Davis presented a software-assisted photo-anthropometric facial landmark identification system that uses 37 distance measures and 25 angular measures [9]. Based on a set of 70 subjects adhering to a similar description, "Identification verification was found to be unreliable unless multiple distance and angular measurements from both profile and full-face images were included in an analysis." Here verification refers to ID verification rather than strength of evidence.

Two other studies on statistics of anthropometric measures (one on South African males [56] and one on three European populations [55]) found that although differences might exist between populations, mostly "Matching these rare features on facial photographs will be useful during cases of disputed identification".

We conclude that anthropometry either in 2D or 3D, and either photographs or in vivo, yields in general limited evidential value, unless a rare or extremely valued feature is observed.

D. Automatic Face Recognition Systems

The last 25 years have seen the development of automated face recognition systems into a mature and active area of research, with some use in FFR [52]. Although some initial work predates it, the Eigenfaces paper can be regarded as the work that successfully sparked a whole new research area [67]. Eigenfaces is an example of a global appearance model. Later methods either use a hybrid (global and local appearances) or a local appearance approach to facial features. The underlying concept is that faces reside in a highly nonlinear manifold of the linear space of images [33], so a linear approach should be locally confined. Local appearance methods can use general feature descriptors like Scale Invariant Feature Transform (SIFT) [37], Local Binary Patterns (LBP) [1], and Histogram of Oriented Gradients (HOG) [10]. By combining multiple regions represented by these features types, a compact representation of the face can be constructed and used.

A recent development in face recognition — and more broadly in artificial intelligence and computer vision — is deep learning, also referred to as deep neural networks or convolutional neural networks [28]. An archetypal example in which deep learning has shown impressive results is the DeepFace system [64] developed by Facebook, but it is questionable whether the images used are representative of those found during forensic casework. Neural networks are computational structures that contain adaptable parameters. Neural networks as such are not new; their topology was already known and used 30 years ago. Their resurgence is mainly enabled by the availability of (a) massive amounts of training data and (b) sheer parallel computing power provided by graphical processor units (GPUs), making feasible the training of the parameters of a deep neural network with many layers. A key difference between these neural networks and other local appearance methods is that they *train* which features are used instead of using features *designed* by a human.

As described in the Introduction, automatic face recognition systems are applied in the FFR domain, but mainly for investigation and intelligence purposes. Additional reasons to rely on human FFR-examiners are the liability and repercussion issues rendered by a misjudgment, irrespective of whether it is in favor of or against a suspect. According to Prince, "Facial recognition systems presently lack good integration into forensic facial comparison procedures" [52]. Also, automated face recognition systems produce a score that (a) is based on abstract features and (b) is a relative measure and does represent the strength of evidence. However, "score calibration" methods convert scores determined by a biometric system into what can be interpreted as strength of evidence; see, for example, Ali for an overview of several of such methods [2].

E. Forensic Face Recognition Datasets

We believe that one of the factors that hampers FFR research is the low number of publicly available forensically relevant datasets, especially in relation to what is available for face biometrics (either controlled or "in-the-wild"). Also, particularly datasets that contain images from surveillance cameras are limited in the number of subjects.

The curious situation is that CCTV is primary designed to monitor activities of people. But when these activities are recognized as criminal, it becomes immediately clear that the technology is often not able to capture the relevant features for the source-level inference. This situation has existed at least for a decade. Finally, all except one dataset (ForenFace) lack an elaborate set of forensically relevant annotation.

The SCFace dataset has been used in numerous publications on low-resolution face recognition [29]. It contains only frontal surveillance camera images of 130 subjects. The ChokePoint dataset is designed for "person identification/verification under real-world surveillance conditions" and contains 29 subjects [71]. Since it does not contain reference images, it is not suitable enough for research within a forensic context. The NIST (1,573 subjects) and Morph (13,618 subjects) datasets contain mugshots, and are very well suited for longitudinal research [44,54]. The ATVS Forensic DB only contains high-resolution mugshots of 50 subjects [69].

Two recent additions are the Quis-Campi and ForenFace (97 subjects) datasets [47,68]. The former uses stills from a PTZ camera, showing subjects possibly nonfrontal, partly occluded, blurred, or overexposed. The images are representative of modern CCTV cameras, notably having higher resolution. A subset of Quis-Campi was used in the ICB 2016 Challenge on Biometric Recognition in the Wild [31]. The unique property of ForenFace is the availability of manual annotation from which a large subset of the FISWG characteristic descriptors can be extracted.

F. Computational Forensic Approaches

Face recognition systems use a constellation of abstract features and as such, the outcome of a facial comparison is difficult to understand outside the broader technical domain of computer vision. There also exist approaches in which either more emphasis is laid on the forensic relevance while still using general feature descriptors, or features are used that have a clear forensic semantic meaning.

Examples of the first approach are Tome et al., in which the biometric performance of linear SVM classifiers on 15 forensic facial regions is investigated [65]. This study uses the SCFace and a subset of Morph. They conclude that "(...) depending on the acquisition distance, the discriminative power of regions changes, having in some cases better performance than the full face". Other examples are facial marks. They are interesting from a forensic perspective as they can be very discriminating. They have been the subject of several studies, notably Park et al. [51] and recently Srinivas et al. [63]. Related work is that of Lee et al. [36] that uses SIFT descriptors for the description of tattoos for search purposes in mugshot databases.

Examples of the second approach include another work by Tome et al. [66]. Here the performance of continuous and discrete soft biometric features is evaluated on the Morph and ATVS Forensic DB datasets. Experimental results show high discrimination power and good recognition performance for some specific cases. However, these cases correspond to relatively good-quality images. In some studies, all features are extracted manually. Two small studies concerning the eyebrow [75] and the periocular region [72] both show that FISWG characteristic descriptors are comparable to their nonforensic counterparts under good image quality. A much larger study by Zeinstra et al. [73] extends this work and investigates discriminating power of many FISWG characteristic descriptors in four representative FFR use cases in [68]. According to [30] and a forensic guideline [41] currently used as a basis for the part 8 of 19795 ISO standard, "Methodology and tools for the validation of biometric methods for forensic evaluation and identification application" under development, discriminating power is one of six aspects that should be taken into account during the validation of a forensic evaluation method. They train and evaluate biometric classifiers specialized on single and combined characteristic descriptors. They found that in all but one use case, commercial systems clearly outperform single and combined characteristic descriptors. In the use case with the lowest-quality trace images (11px interpupillary distance, severe image compression) they found that (a) the combination of visibility features and (b) the hairline perform better than a commercial system.

Finally, there is another development that can be mentioned. Landmark detection is important for the automatic detection and extraction of certain facial features, especially the shapelike ones. Recent work by [43] shows that it is now possible to locate to a certain extent landmarks in even uncontrolled scenarios. These results can be used to extract forensically relevant facial features in an automatic manner.

CONCLUSION AND FUTURE DIRECTIONS

In this survey, we have presented several aspects of FFR: the historical context, use cases, the three operational levels of FFR, criticism of FFR, and research efforts pertaining to FFR.

We observe several positive developments. Some recent validation studies indicate that the FFR-examiner is "doing better", in particular with respect to the nonexaminers. Although anthropometry is closely tied to FFR, especially in the minds of members of the general public, multiple studies reinforce the conclusion that it is limited in its ability to produce meaningful strength of evidence. We recognize the potential of automated face recognition systems as an instrument to help the examiners to assess the strength of evidence and complement the human-based approach. Furthermore, recent advances in fast and accurate automatic detection of facial features could aid the work of the FFR-examiner. Examiners can assess features that are difficult to describe statistically but can only be validated mostly on basis of competence and proficiency, and not performance. Automatic approaches use a reduced set of features that can be described statistically but can be validated empirically and extensively, and can be improved.

Despite the recent progress, challenges remain. At a higher, general forensic level, we think the community should (a) better understand the goal of being able to assign/compute the strength of evidence, (b) be able to validate analysis, comparison, and interpretation methods, and (c) be able to combine the human and computer-based approaches to generate the most correct strength of evidence. All these goals are not easy to reach.

At the level of FFR, there are other problems that need to be addressed. Since large, publicly available, forensically relevant datasets are lacking, descriptive statistics of facial features extracted from images representative of forensic use cases are not available. This is important as it would have helped to determine strength of evidence in a more scientific manner. The use of automatic detection of facial features can aid this process. Moreover, current datasets lack the broad variation and use cases needed to systematically investigate the influence of multiple factors found in real forensic casework.

We therefore advocate the collection of a large-scale dataset of images grounded in clear forensic use cases, employing forensically relevant parameters. An alternative approach is the development of a large synthetic dataset for the study of the effect of those forensically relevant parameters.

REFERENCES

1. Ahonen T, Hadid A, Pietikainen M: Face description with local binary patterns: Application to face recognition; *IEEE Trans Pattern Anal Mach Intell* 28:2037; 2006.
2. Ali T: *Biometric Score Calibration for Forensic Face Recognition*, PhD Thesis; University of Twente: Enschede, The Netherlands; 2014.
3. Barber D: *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK; 2012.
4. Bertillon A: *Identification Anthropométrique: Instructions Signalétiques*; Imprimerie: Paris, France; 1893.
5. Bromby MC, Plews S (Forensic Imagery Analysis Group): Guidance for Evaluating Levels of Support; *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550752* (Accessed November 21, 2017).
6. Bruce V, Henderson Z, Greenwood K, Hancock PJB, Burton AM, Miller P: Verification of face identities from images captured on video; *J Exp Psychol-Appl* 5:339; 1999.
7. Bruce V, Henderson Z, Newman C, Burton AM: Matching identities of familiar and unfamiliar faces caught on CCTV images; *J Exp Psychol-Appl* 7:207; 2001.
8. Burton AM, Wilson S, Cowan M, Bruce V: Face recognition in poor-quality video: Evidence from security surveillance; *Psychol Sci* 10:243; 1999.
9. Davis JP, Valentine T, Davis RE: Computer assisted photo-anthropometric analyses of full-face and profile facial images; *Forensic Sci Int* 200:165; 2010.
10. Déniz O, Bueno G, Salido J, De la Torre F: Face recognition using histograms of oriented gradients; *Pattern Recogn Lett* 32:1598; 2011.
11. Dror IE, Charlton D, Péron AE: Contextual information renders experts vulnerable to making erroneous identifications; *Forensic Sci Int* 156:74; 2006.
12. Edmond G, Biber K, Kemp RI, Porter G. Law's looking glass: Expert identification evidence derived from photographic and video images; *Curr Iss Crim Justice* 20:3; 2009.
13. European Network of Forensic Science Institutes; *http://enfsi.eu/* (Accessed November 21, 2017).
14. European Network of Forensic Science Institutes: Guideline for Evaluative Reporting in Forensic Science; *https://www.forensicinstitute.nl/news/news/2015/05/19/european-guideline-for-evaluative-reporting-in-forensic-science* (Accessed November 21, 2017).
15. Evison M, Vorder Bruegge R: *Computer-Aided Forensic Facial Comparison*; Taylor and Francis Group: Boca Raton, FL; 2010
16. Evison MP: Forensic facial analysis; In Bruinsma G, Weisburd D (Eds): *Encyclopedia of Criminology and Criminal Justice*; Springer: New York, NY; p 1713; 2014.
17. Facial Comparison List used at NFC Sweden; Private communication with forensic facial examiner at NFC Sweden; E-mail on September 6, 2015.
18. Facial Comparison List used at NFI Netherlands; Private communication with forensic facial examiner at NFI The Netherlands; E-mail on December 9, 2015.
19. Facial Identification Scientific Working Group; *https://fiswg.org* (Accessed November 21, 2017).
20. Facial Identification Scientific Working Group: Facial Image Comparison Feature List for Morphological

Analysis; *https://fiswg.org/FISWG_1to1_Checklist_v1.0_2013_11_22.pdf* (Accessed November 21, 2017).

21. Facial Identification Scientific Working Group: Guidelines and Recommendations for Facial Comparison Training to Competency; *https://fiswg.org/FISWG_Training_Guidelines_Recommendations_v1.1_2010_11_18.pdf* (Accessed November 21, 2017).
22. Facial Identification Scientific Working Group: Guidelines for Facial Comparison Methods; *https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf* (Accessed November 21, 2017).
23. Facial Identification Scientific Working Group: Image Processing to Improve Automated Facial Recognition Search Performance; *https://fiswg.org/DRAFT_FISWG_ImageProcessingtoImproveFRSearchPerf_v1.0_2016_07_26.pdf* (Accessed November 21, 2017).
24. Facial Identification Scientific Working Group: Physical Stability of Facial Features of Adults; *https://fiswg.org/DRAFT_FISWG_Physical_Stability_of_Facial_Components_v1.0_20160202.pdf* (Accessed November 21, 2017).
25. Facial Identification Scientific Working Group: Standard for Facial Identification and Facial Recognition Proficiency Testing Programs; *https://fiswg.org/DRAFT_FISWG_Proficiency_Test_Program_Standard20140509.pdf* (Accessed November 21, 2017).
26. Finn J: *Capturing the Criminal Image: From Mug Shot to Surveillance Society, New edition*; University of Minnesota Press: Minneapolis, MN; 2009.
27. Gold JM, Barker JD, Barr S, Bittner JL, Bratch A, Bromfield WD, Goode RA, Jones M, Lee D, Srinath A: The perception of a familiar face is no more than the sum of its parts; *Psychon Bull Rev* 21:1465; 2014.
28. Goodfellow I, Bengio Y, Courville A: *Deep Learning*; MIT Press: Cambridge, MA; 2016.
29. Grgic M, Delac K, Grgic S: SCFace — Surveillance cameras face database; *Multimedia Tools and Applications* 51:863; 2011.
30. Haraksim R, Ramos D, Meuwly D, Berger CEH: Measuring coherence of computer-assisted likelihood ratio methods; *Forensic Sci Int* 249:123; 2015.
31. International Challenge on Biometric Recognition in the Wild; http://icbrw.di.ubi.pt/ (Accessed November 21, 2017).
32. Jain AK, Flynn P, Ross AA: *Handbook of Biometrics*; Springer-Verlag New York: Secaucus, NJ; 2007.
33. Jain AK, Klare B, Park U: Face matching and retrieval in forensics applications; *IEEE MultiMedia* 19:20; 2012.
34. Kleinberg KF: *Facial Anthropometry as an Evidential Tool in Forensic Image Comparison*; Ph.D. Thesis; University of Glasgow: Glasgow, UK; 2008.
35. Langenburg GM: *A Critical Analysis and Study of ACE-V Process*; Ph.D. thesis; Université de Lausanne: Lausanne, Switzerland; 2008.
36. Lee JE, Jain AK, Jin R: Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification; *Proceedings — Biometrics Symposium*; Tampa, FL; p 1; September 2008.
37. Lowe DG: Distinctive image features from scale-invariant keypoints; *Int J Computer Vision* 60:91; 2004.
38. Mallett X, Evison MP: Forensic facial comparison: issues of admissibility in the development of novel analytical technique; *J Forensic Sci* 58:859; 2013.
39. Megreya AM, Burton AM: Unfamiliar faces are not faces: Evidence from a matching task; *Mem Cognit* 34:865; 2006.
40. Meuwly D: Le Mythe de l'Empreinte Vocale (I et II). *Revue Internationale de Criminologie et de Police Technique et Scientifique* 56:219; 2003.
41. Meuwly D, Ramos D, Haraksim R: A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation; *Forensic Sci Int* 276:142; 2017.
42. Meuwly D, Veldhuis R: Forensic biometrics: From two communities to one discipline; *Proceedings — 11th International Conference of the Biometrics Special Interest Group (BIOSIG);* Fraunhofer Institute: Darmstadt, Germany; p 1; Sep 2012.
43. Milborrow S, Nicolls F: Active shape models with SIFT Descriptors and MARS; *Proceedings —International Conference on Computer Vision Theory and Applications*; Lisbon, Portugal; p 380; January 2014.
44. National Institute of Standards and Technology: NIST Mugshot Identification Database; *http://www.nist.gov/srd/nistsd18.cfm* (Accessed November 21, 2017).
45. National Research Council: *Strengthening Forensic Science in the United States: A Path Forward*; The National Academies Press: Washington, DC; 2009.
46. Nederlands Register Gerechtelijk Deskundigen; *https://www.nrgd.nl/* (Accessed November 21, 2017).
47. Neves JC, Santos G, Filipe S, Grancho E, Barra S, Narducci F, Proença H: Quis-Campi: Extending in the wild biometric recognition to surveillance environments; *Proceedings — International Conference on Image Analysis and Processing*; ICIAP: Cham, Switzerland; p 59; September 2009.
48. Norell K, Läthén KB, Bergström P, Rice A, Natu V, O'Toole A: The effect of image quality and forensic expertise in facial image comparisons; *J Forensic Sci* 60:331; 2015.
49. O'Toole AJ, Jiang F, Roark D, Abdi H: Predicting human performance for face recognition; In Chellappa R, Zhao W (Eds): *Face Processing: Advanced Methods and Models*; Elsevier/Academic Press: Amsterdam, The Netherlands; 2006.
50. Papesh MH, Goldinger SD: Infrequent identity mismatches are frequently undetected; *Atten Percept Psychophys* 76:1335; 2014.
51. Park U, Jain AK: Face matching and retrieval using soft biometrics; *IEEE Trans IFS* 5:406; 2010.
52. Prince JP: *To Examine Emerging Police Use of Facial Recognition Systems and Facial Image Comparison Procedures*; *http://www.churchilltrust.com.au/media/fellows/2012_Prince_Jason.pdf* (Accessed November 21, 2017).
53. Pronin E: How we see ourselves and how we see others; *Science* 320:1177; 2008.
54. Ricanek Jr K, Tesafaye T: MORPH: A longitudinal image database of normal adult age-progression; *Proceedings — 7th International Conference on Automatic Face and Gesture Recognition*; IEEE Computer Society: Los Alamitos, CA; p 341; April 2006.
55. Ritz-Timme S, Gabriel P, Tutkuviene J, Poppa P, Obertova Z, Gibelli D, De Angelis D, Ratnayake M, Rizgeliene R, Barkus A, Cattaneo C: Metric and morphological assessment of facial features: A study on three European populations; *Forensic Sci Int* 207:239; 2011.
56. Roelofse MM, Steyn M, Becker PJ: Photo identification: Facial metrical and morphological features in South African

males; *Forensic Sci Int* 177:168; 2008.

57. Russell R, Duchaine B, Nakayama K: Super-recognizers: People with extraordinary face recognition ability; *Psychon Bull Rev* 16:252; 2009.
58. Saks MJ, Koehler JJ: The coming paradigm shift in forensic identification science; *Science* 309:892; 2005.
59. Sinha P, Balas B, Ostrovsky Y, Russell R: Face recognition by humans: Nineteen results all computer vision researchers should know about; *Proc IEEE* 94:1948; 2006.
60. Sklar A: Fonctions de répartition à n dimensions et leurs marges; *Publications de l'Institut de Statistique de l'Université de Paris*; 1959.
61. Spaun NA: Facial comparisons by subject matter experts: Their role in biometrics and their training. *Proceedings — 3rd International Conference on Biometrics*; Alghero, Italy; p 161; June 2009.
62. Spaun NA: Forensic biometrics from images and video at the Federal Bureau of Investigation; *Proceedings — First IEEE International Conference on Biometrics: Theory, Applications, and Systems*; Washington, DC; p 1; September 2007.
63. Srinivas N, Flynn PJ, Vorder Bruegge RW: Human identification using automatic and semi-automatically detected facial marks; *J Forensic Sci* 61:117; 2016.
64. Taigman Y, Yang M, Ranzato M, Wolf L: DeepFace: Closing the gap to human-level performance in face verification; *Proceedings — 2014 IEEE Conference on Computer Vision and Pattern Recognition*; Columbus, OH; p 1701; June 2014.
65. Tome P, Fierrez J, Vera-Rodriguez R, Ramos D: Identification using face regions: Application and assessment in forensic scenarios; *Forensic Sci Int* 233:75; 2013.
66. Tome P, Vera-Rodriguez R, Fierrez J, Ortega-Garcia J: Facial soft biometric features for forensic face recognition; *Forensic Sci Int* 257:271; 2015.
67. Turk M, Pentland A: Eigenfaces for recognition; *J Cogn Neurosci* 3:71; 1991.
68. University of Twente: ForenFace; *http://scs.ewi.utwente.nl/downloads/show,ForenFace/* (Accessed November 21, 2017).
69. Vera-Rodriguez R, Tome P, Fierrez J, Expsito N, Vega FJ: Analysis of the variability of facial landmarks in a forensic scenario; *Proceedings — 1st International Workshop on Biometrics and Forensics*; Lisbon, Portugal; p 1; April 2013.
70. White D, Phillips PJ, Hahn CA, Hill M, O'Toole AJ: Perceptual expertise in forensic facial image comparison; *Proc R Soc B* 282:20151292; 2015; *https://www.cogsci.msu.edu/DSS/2015-2016/OToole/OToole_2015_Perceptual.pdf* (Accessed December 14, 2017).
71. Wong Y, Chen S, Mau S, Sanderson C, Lovell BC: Patchbased probabilistic image quality assessment for face selection and improved video-based face recognition; *Proceedings — IEEE Biometrics Workshop, Computer Vision and Pattern Recognition Workshops*; University of Colorado: Boulder, CO; p 81; June 2011.
72. Zeinstra CG, Veldhuis RNJ, Spreeuwers LJ: Beyond the eye of the beholder: On a forensic descriptor of the eye region; *Proceedings — 23rd European Signal Processing Conference (EUSIPSO)*; EURASIP: Nice, France; p 779; September 2015.
73. Zeinstra CG, Veldhuis RNJ, Spreeuwers LJ: Discriminating power of FISWG characteristic descriptors under different forensic use cases; *Proceedings — 15th International Conference of the Biometrics Special Interest Group;* Fraunhofer Institute: Darmstadt, Germany; p 171; September 2016.
74. Zeinstra CG, Veldhuis RNJ, Spreeuwers LJ: Examining the examiners: An online eyebrow verification experiment inspired by FISWG; *Proceedings — 3rd International Workshop on Biometrics and Forensics*; University of Gjøvik: Gjøvik, Norway; p 1; March 2015.
75. Zeinstra CG, Veldhuis RNJ, Spreeuwers LJ: Towards the automation of forensic facial individualisation: Comparing forensic to non-forensic eyebrow features; *Proceedings — 35th WIC Symposium on Information Theory in the Benelux*; Eindhoven, Netherlands; p 73; May 2014.
76. Zolfagharifard E: Are you a super recognizer? *http://www.dailymail.co.uk/sciencetech/article-3125173* (Accessed November 21, 2017).

ABOUT THE AUTHORS

C. Zeinstra; D. Meuwly; A. Ruifrok
R. Veldhuis; L. Spreeuwers

Chris Zeinstra studied mathematics (M.Sc.) at the University of Groningen (Groningen, The Netherlands), and took a special one-year course in computer science (Hons.) at the University of Leiden) Leiden, The Netherlands), between 1990 and 1998. He started his career as a software engineer at Royal Dutch Mail and as a technical project manager at Ericsson ETM, The Netherlands. He joined the Hanze University of Applied Sciences (Groningen, The Netherlands) as a lecturer in computer science in 2002. In 2017 he obtained his Ph.D. for his work on forensic face recognition, a joint project between the University of Twente and the Netherlands Forensic Institute. In this project he applied biometric techniques on features that have a forensic meaning; in particular FISWG characteristic descriptors. His Ph.D. supervisors were Raymond Veldhuis and Luuk Spreeuwers.

Didier Meuwly received a classical education (Latin/Philosophy) and he was educated as a criminalist and criminologist (1993). He obtained his Ph.D. (2000) at the School of Forensic Science (IPS) of the University of Lausanne, Switzerland. Currently he shares his time between the Forensic Institute of the Ministry of Security and Justice of the Netherlands (NFI), where he is a principal scientist, and the University of Twente, where he holds the chair of the Forensic Biometrics Department.

Dr. Meuwly specializes in the automation and validation of the probabilistic evaluation of forensic evidence, and more particularly of biometric traces. He was previously the leader of a project about the probabilistic evaluation of fingermark evidence and of the fingerprint section within the NFI. From 2002 to 2004, he worked as a senior forensic scientist within the R&D department of the Forensic Science Service (UK-FSS), at the time an executive agency of the British Home Office. From 1999 to 2002 he was a leader of the biometric research group of the IPS. He is a founding member of two working groups of the European Network of Forensic Science Institutes (ENFSI): the Forensic Speech and Audio Analysis Working Group (FSAAWG) in 1997 and the European Fingerprint Working Group (EFPWG) in 2000. He is still active within the EFWPG. Dr. Meuwly is also a member of the editorial board and a guest editor of *Forensic Science International* (FSI).

Arnout Ruifrok received his M.Sc. in biology from the University of Groningen (Groningen, The Netherlands), in 1982. In 1987 he received his Ph.D. from the same university for his work on the evaluation of the effects of heat-treatment (hyperthermia) on the cellular plasma membrane. In 2002 he joined the Netherlands Forensic Institute, where he is a team leader in the area of image analysis, biometrics speech, and audio. His main responsibilities are research on the possibilities of use of biometric features in forensic identification, forensic applications of biometric systems, and facial comparison.

From 1987 to 1992 Dr. Ruifrok was involved with research projects concerning the effects of heat and radiation on tumor and normal tissues in animal model systems at the Daniel den Hoed Hospital (Rotterdam, The Netherlands) and the Radboud University (Nijmegen, The Netherlands). In 1992 he moved to the US to work on mathematical models of radiation response of tissues at the M. D. Anderson Cancer Center (Houston, TX, US), Department of Biomathematics, where he became involved in image analysis and pattern recognition. From 1999 to 2002 he joined the M. D. Anderson Department of Pathology, working on automated recognition and classification of cancer cells and in situ hybridization signals. Dr. Ruifrok is active in the Facial Identification Scientific Working Group.

Raymond Veldhuis graduated from the University of Twente (Twente, The Netherlands) in 1981. In 1988 he received a Ph.D. degree from Nijmegen University (Nijmegen, The Netherlands), with a thesis entitled "Adaptive Restoration of Lost Samples in Discrete-Time Signals and Digital Images". Dr. Veldhuis is now a full professor in biometric pattern recognition at the University of Twente, where he is leading a research team in this field. The main research topics are face recognition (2D and 3D), fingerprint recognition, vascular pattern recognition, multibiometric fusion, and biometric template protection. The research is both applied and fundamental.

Dr. Veldhuis was a researcher at Philips Research Laboratories (Eindhoven, The Netherlands) working in various areas of digital signal processing (1982–1992) and in the field of speech processing (1992–2001).

Luuk Spreeuwers studied electrical engineering at the University of Twente (Twente, The Netherlands), from 1982 to 1988. In 1992 he received his Ph.D. from the University of Twente with a thesis entitled "Image Filtering with Neural Networks — Applications and Performance". Subsequently he worked as a postdoc at the University of Twente and at the Hungarian Academy of Sciences (Budapest, Hungary). Currently Dr. Spreeuwers investigates 2D and 3D face recognition in Dr. Veldhuis's group at the University of Twente. Dr. Spreeuwers's current main interests are model-based image processing, pattern recognition, and biometrics.

Dr. Spreeuwers was senior researcher (1997–1999) in the area of image processing in Mindmaker Ltd. (Budapest, Hungary). During 1999–2005, Dr. Spreeuwers worked on 3-D modeling and segmentation at the University Medical Center (Utrecht, The Netherlands).

Dr. Spreeuwers's main interests are model-based image processing, pattern recognition, and biometrics. He holds world records in 3D face recognition on the standard verification and identification benchmarks using FRGCv2 data (99.3% @FAR=0.1% and 99.4% rank 1).