A Framework for Forensic Face Recognition based on Recognition Performance Calibrated for the Quality of Image Pairs

Abhishek Dutta Raymond Veldhuis Luuk Spreeuwers Didier Meuwly

Signals and Systems group, University of Twente Netherlands Forensic Institute

{a.dutta,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl d.meuwly@nfi.minjus.nl

Abstract

Recently, it has been shown that performance of a face recognition system depends on the quality of both face images participating in the recognition process: the reference and the test image. In the context of forensic face recognition, this observation has two implications: a) the quality of the trace (extracted from CCTV footage) constrains the performance achievable using a particular face recognition system; b) the quality of the suspect reference set (to which the trace is matched against) can be judiciously chosen to approach optimal recognition performance under such a constraint. Motivated by these recent findings, we propose a framework for forensic face recognition that is based on calibrating the recognition performance for the quality of pairs of images. The application of this framework to several mock-up forensic cases, created entirely from the MultiPIE dataset, shows that optimal recognition performance, under such a constraint, can be achieved by matching the quality (pose, illumination, and, imaging device) of the reference set to that of the trace. This improvement in recognition performance helps reduce the rate of misleading interpretation of the evidence.

1. Introduction

Forensic investigators now have access to video recording of many crime scenes – thanks to the omnipresent CCTV cameras. In a forensic face recognition case, a trace is the facial image extracted from CCTV footage of a crime scene and a suspect reference set refers to individuals who are being investigated for involvement in that particular crime. Traces are often of very low quality (in terms of pose, illumination, resolution, etc.) and therefore, even the experts trained in manual forensic face recognition have difficulty in comparing and interpreting the image.

Dramatic improvement in accuracy of automatic face recognition systems in the past two decades has encouraged application of the current state-of-the-art in face recognition systems in forensic casework. However, the following limitations of automatic face recognition have impeded its growth as a reliable forensic technology:

1. Even current state-of-the-art face recognition system are known to have very poor recognition performance on facial images captured in an uncontrolled environment. [8]

2. It is difficult to assess the extent of recognition performance degradation if a particular face recognition system is subject to low quality CCTV face images. Therefore, results from such unvalidated recognition systems have very low evidential value in a forensic case.

In this paper, we propose a framework for forensic face recognition designed specifically to address these two limitations.

Recently, in [2] it has been shown that the performance of a face recognition system is related to the quality of both images participating in the recognition process. In other words, if we define quality as any measurable property of an image that is predictive of face recognition performance, then quality is the property of an image pair and not of an individual image. In the context of forensic face recognition and the above two limitations, this implies that performance of a particular face recognition system cannot be solely attributed to the low quality trace. The quality of facial images in the suspect reference set also determine the recognition performance achievable by a face recognition system. Moreover, given the quality of the trace, the quality of the suspect reference set can be judiciously chosen to approach optimal recognition performance under such a constraint.

Motivated by these recent findings of [2], we propose a framework for forensic face recognition in which we search the quality space of the suspect reference set in order to determine the quality that results in optimal recognition performance achievable by a particular face recognition sys-

tem. In this way, we not only assess the possible recognition performance variation for a given trace quality, but we also improve the evidential value by judiciously choosing the quality of suspect reference set in order to attain optimal recognition performance; thereby addressing the above two limitations.

This paper has been organized as follows: First, we briefly describe the standard framework that uses likelihood ratio (LR) for forensic individualization. Building upon this framework, we propose a framework for forensic face recognition based on calibrating the recognition performance of a face recognition system for the quality of image pairs. Finally, we present results of our framework applied to several mock-up forensic cases created entirely from the CMU MultiPIE dataset [4].

2. Related Work

The framework of [6] for forensic individualisation from biometric data (speech, fingerprint, face, etc.) is based on the hypothetical deductive method of [5], which begins with the generation of a set of hypotheses explaining the source of the trace. The criteria of possibility and plausibility are used to define the set of hypotheses that will be tested empirically with the LR approach. These empirical tests will show the degree of support of the evidence for each pair of hypotheses tested.

The dangers of a "poorly designed experiment", when a framework based on hypothetical deductive method is used to forensic individualisation, has been emphatically stated by [5, p.27]. For such a framework, hypothesis testing experiments should be carefully designed so that a hypothesis is neither illegitimately rejected nor unjustifiably accepted.

In a forensic case, there are two mutually exclusive hypotheses. First is the prosecution hypothesis ($\mathrm{H_p}$) which states that the trace/mark Y originated from a source X_1 (i.e. individual X_1 is the source of the trace/mark Y). Second is the defence hypothesis ($\mathrm{H_d}$) which, on the contrary, claims that the trace/mark originated from an alternative source (i.e. the trace/mark Y originated from some other source $X_{k\neq 1}$). The framework of [6] is based on testing these two competing hypotheses ($\mathrm{H_p}$) and ($\mathrm{H_d}$) using the "within-source" and "between-source" distributions of biometric similarity scores.

The "within-source" distribution captures the extent of variation possible when an individual's trace/mark is compared against reference traces/marks from the same individual. This distribution is constructed from a set of scores obtained by comparing the putative source control database (C) against the putative source reference database (R). The Putative Source Control Database (C) is a biometric database of pseudo traces which is "made up of information that is ideally of the same quality as the trace/mark, but originated from the putative source X_1." [6, p.211]. Similarly, the Putative Source Reference Database (R) is a biometric database "made up of information that ideally contains the exhaustive characteristics of interest of the putative source X_1." [6, p.211].

The "between-source" distribution captures the extent of possible matching score variations when a trace/mark is compared against reference samples of potential population database that contains an "exhaustive characteristics of interest of the alternative source $X_{k\neq 1}$" [6, p.210]. Such a biometric database is called the Potential Population Database (P) which is a biometric database that contains all the characteristic features present in the biometric data of the alternative sources. The "between-source" distribution is constructed from a set of scores obtained when the trace/mark Y is compared against the Potential Population Database (P).

From the "within-source" and "between-source" distributions, the numerical value of the likelihood ratio is estimated as: $LR = \frac{P(E|\mathrm{H_p})}{P(E|\mathrm{H_d})}$.

In a real forensic evaluation case involving face recognition, forensic investigators often find it difficult to acquire sufficient face images of the suspect in order to create sufficiently complete P and R database. This prevents estimation of true "within-source" and "between-source" distributions specific to the forensic face recognition case under consideration. In such a case, a common practice is to assume that generic "same-source" (true match) and "different-source" (false match) distributions (generated by the representative population) is a good approximation of the true "within-source" and "between-source" distributions. It is important to realise that validity of this assumption rests on the capability of the face recognition system to properly deal with possible image quality variations. Based on this generic "same-source" and "different-source" distributions, forensic investigators estimate the evidential value (or, likelihood ratio) of evidence (E). Ideally, likelihood ratio should be estimated from true "within-source" and "between-source" distributions. A subject of future study is to investigate the validity of this assumption.

The ability of these frameworks ([5] and [6]) to demonstrate robustness and provide strong evidential value rests upon the three biometric databases (C, P, R) that is used to create a probabilistic view of the evidence (E) under the two competing hypotheses $\mathrm{H_p}$ and $\mathrm{H_d}$. The notion of "poorly designed experiment", suggested by [5, p.27], comes into play if the contents of these databases are not sufficient to estimate the true "within-source" and "between-source" distributions (i.e. if R and P databases are not sufficiently complete)

Building upon the work of [6], we propose a framework specifically designed to address the issue of selecting the quality of face images in these 2 critical databases (P and R) in a forensic face recognition case. Recall that

we already know that the quality of Putative Source Control Database (C) should be similar to that of the trace [6]. In this framework, we perform calibration, with samples for which ground truth is known, to determine the quality of reference set that results in optimal recognition performance. For such calibration, the quality of calibration trace set is fixed to that of the true trace in the forensic face recognition case. Such a calibration helps judiciously chose the quality of face images in the P and R databases in order to achieve optimal recognition performance.

This quality based calibration of a particular face recognition system ensures that the two hypotheses are always testing using experiments based on optimal data. i.e. "well designed experiment". It is important to understand that the adequacy of an experiment to test the two competing hypotheses is also constrained by two other factors: quality of the trace and recognition performance of the face recognition system under consideration for the quality at hand.

3. A Framework for Forensic Face Recognition

We first define some forensic terminologies that will be used to describe our framework. Recall that trace Y is a face image extracted from CCTV footage of the crime scene and the individual X_1 is suspected of being the person in the trace. The prosecution hypothesis states that the trace contains face image of individual X_1 (the putative source) i.e. $\mathrm{H_p} : X_1$ is the source of Y while the defence hypothesis states that it is someone else (alternative sources X_i where $i \neq 1$) i.e. $\mathrm{H_d} : X_1$ is not the source of Y. evidence (E) is the similarity score value obtained when the trace Y is compared to the putative source X_1.

The technical characteristics of a facial image can be grouped into the following four categories [3]: *a*) Defects caused by environment (illumination, background, etc.); *b*) Defects caused by camera conditions (resolution, distortion, etc.); *c*) Defects caused by user's face conditions (expression, make-up, etc.); and, *d*) Defects caused by user-camera positioning (pose, focus, etc.). In this paper, we use the term "image quality" to refer to these technical properties of a face image.

In a typical forensic face recognition case, a particular face recognition system (often commercial) is used to obtain a similarity score (i.e. evidence E) between the trace Y and the putative source X_1. The likelihood ratio (LR) conveys the relative support for observing the evidence (E) when the prosecution hypothesis ($\mathrm{H_p}$) is true, versus the probability of observing the same evidence (E) when the defence hypothesis ($\mathrm{H_d}$) is true. For a particular evidence (E), the numerical value of LR is given by: $LR = \frac{P(E|\mathrm{H_p})}{P(E|\mathrm{H_d})}$.

A critical stage of every forensic evaluation case involving face recognition is to determine the quality of face images in R (Putative Source Reference Database) and P (Potential Population Database) so that the two mutually exclusive hypotheses ($\mathrm{H_p}$ and $\mathrm{H_d}$) are tested using a "well designed experiment". For example, one common dilemma is: should frontal face images of the putative source X_1 and the alternative sources $X_{i\neq 1}$ be used in R and P database respectively because vendor of the commercial face recognition system recommends use of frontal face images for optimal recognition performance?

In our framework, we find an answer to this critical question by assessing recognition performance of the face recognition system on a representative face population database. This representative population database include individuals that have facial characteristics similar to that of the suspects of the forensic face recognition case.

For this performance assessment, which we refer to as "calibration" for a pair of image qualities, we create two databases of the individuals in the representative population. First is the calibration trace set which contains face images having quality similar to the trace. Second is the calibration reference set which is partitioned into subsets with varying image qualities, where quality is the optimization parameter for "calibration".

From this calibration of the face recognition system, we expect optimal recognition performance when the quality (in terms of pose, illumination, imaging device, etc) of the calibration trace set and the calibration reference set images are aligned. Such an alignment of image quality, causes the difference in face images due to identity to become more prominent. However, we do not expect improvement in recognition performance by alignment of non-deterministic degradations like imaging noise.

Based on the results of calibration, facial images in the P and R databases are now transformed to have the quality of optimal calibration reference set. This quality of P and R database ensures that a face recognition system operates at optimal recognition performance level achievable for the given trace quality.

We now describe the processing pipeline of our framework (also depicted in Fig. 1) in detail by dividing the whole process into the following four steps:

3.1. Step 1: Create a Calibration Trace Set having Image Quality Similar to that of the Trace

For a given trace Y, the Quality Assessor (QA) module quantifies the image quality of the trace (q_{trace}) in terms of the predefined face image quality parameters. Using this set of image qualities, we create a calibration trace set that consists of face images of the representative population having image quality similar to that of the trace.

Realisation of such a calibration trace set requires functionality of the following two modules:

- Quality Assessor (QA) module which, given a trace

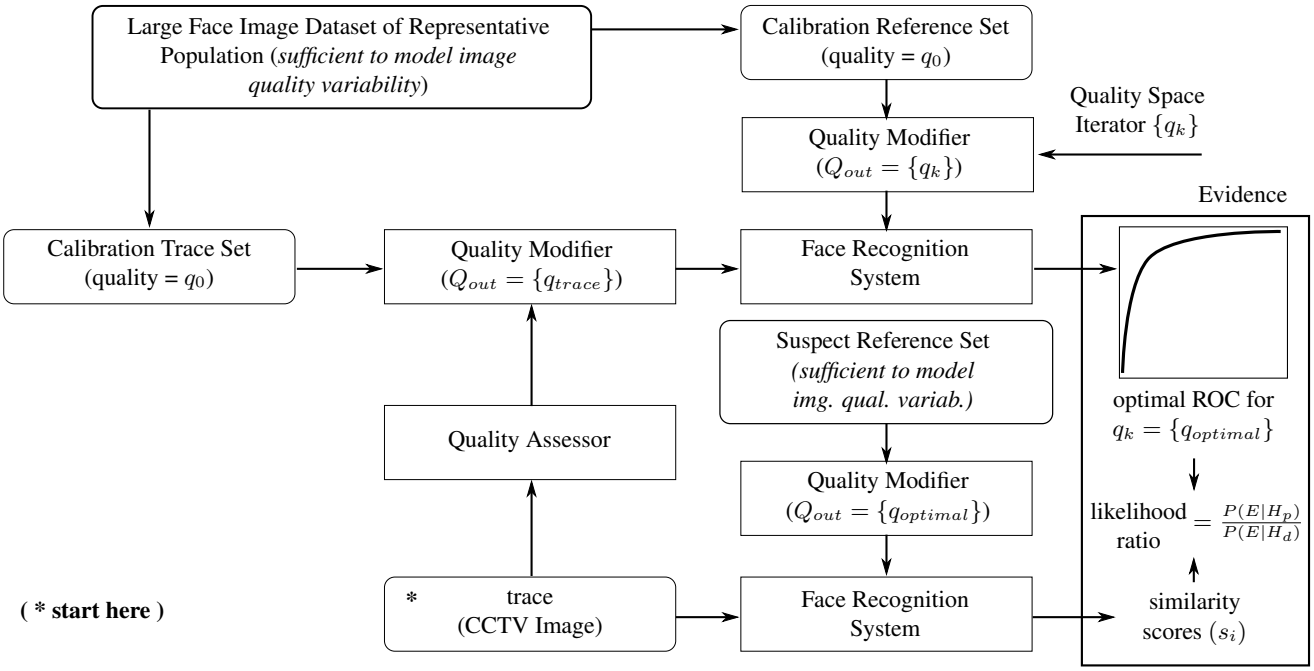


Fig. 1: Framework for forensic face recognition based on calibration of a face recognition system for quality of image pairs, in which, one of the image quality is fixed to that of the trace image and we do not have any control over it.

image, quantifies the image quality in terms of predefined quality parameters.

- Quality Modifier (QM) module which can transform a given face image with baseline quality (q_0) into a new face image having any user defined quality. Such functionality generally requires, for each individual, a set of basis face images that span the quality space.

3.2. Step 2: Search for the Optimal Image Quality of the Calibration Reference Set

Keeping the calibration trace set (that was created in Step 1) fixed, we search for the quality of calibration reference set that results in optimal recognition performance. The optimality of recognition performance is judged by area under ROC curve corresponding to each possible pair of calibration trace set and calibration reference set image quality. Optimality of ROC can also be judged according to requirements of the forensic investigators in terms of desired Verification Rate (VR or True Accept Rate) and False Acceptance Rate (FAR).

At the end of this step, we have a quality pair of sets (formed by a calibrated trace set and a calibrated reference set) for which the face recognition system has optimal recognition performance. This optimal quality pair of sets will be used in Step 3. This calibration also produces "same-source" (true matches) and "different-source" (false matches) distributions which depict score variation when face images of same and different individuals of the representation population are compared. These generic distributions will be used in Step 4 (the last step) to determine evidential value by assuming that the generic distributions provide a good approximation of the true "within-source" and "between-source" distributions.

3.3. Step 3: Create a Suspect Reference Set having Image Quality of the Optimal Calibration Reference Set

We create a suspect reference set such that the quality pair formed by the trace set and the suspect reference set is same as the optimal quality pair of sets obtained from the calibration of Step 2. In other words, we create a suspect reference set having quality of the optimal calibration reference set as determined in Step 2.

Again, realising such a suspect reference set requires the functionality of Quality Modifier (QM) module.

3.4. Step 4: Compute evidence and Likelihood Ratio

With the suspect reference set transformation of Step 3, the trace and suspect reference set form the optimal quality pair (as determined in calibration stage of Step 2). The corresponding optimal "same-source" (true matches) and "different-source" (false matches) distributions (that also generate the optimal ROC curve in Step 2) is used to estimate the likelihood ratio. For example: the method

of [7] can be adapted to estimate LR value from generic "same-source" (true matches) and "different-source" (false matches) distributions.

As stated earlier, the LR value estimated from generic "within-source" and "between-source" distributions assume that these two distributions provide a reasonable approximation to the true LR value when sufficient data is not available to estimate the LR from true "within-source" and "between-source" distributions.

4. Experimental Results

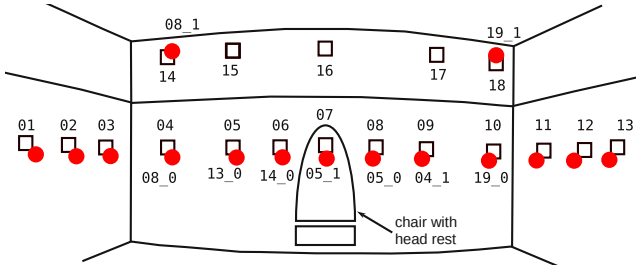


Fig. 2: Position of camera (red circles, e.g. `08_1`) and flash (black squares, e.g. `04`) in the MultiPIE collection room.

Here, we present a proof of concept by applying our framework to a set of mock-up forensic cases because implementation of the QA and QM modules is still an open problem. These mock-up forensic cases are designed such that all the facial images required by our framework are present in the CMU MultiPIE dataset [4].

For the sake of simplicity in illustration, we only consider pose (quality parameter due to user-camera positioning) and illumination (quality parameter due to environment) variation in this experimental evaluation of our framework using the MultiPIE dataset.

Fig. 3: Trace (left) having quality `(03,19_1,18)` and (right) individuals in the suspect reference set

The MultiPIE dataset samples face of 346 individuals over 4 sessions from a discrete set of camera and flash positions as shown in Fig. 2. To create a set of mock-up forensic cases, we randomly select 66 individuals who appear in both session `03` and `04`. In all the cases, the trace has quality `(03,19_1,18)`[1] and suspect reference set contains the randomly selected 66 individuals.

We use Cognitec FaceVACS SDK [1] as the face recognition system for all the experiment results discussed in this section. This SDK does not proceed to face comparison stage if it fails to detect both eyes in face images of the trace set or the reference set. Therefore, we only use images captured by the camera positions labelled in Fig. 2 (red circles).

The calibration involves searching the quality space of the calibration reference set in order to achieve best separability of the "same-source" (true matches) and "different-source" distributions for the given trace quality. As the quality of trace `(03,19_1,18)` remains constant in all the 66 mock-up forensic cases, we only require a single calibration (Step 1 and 2) to determine the optimal quality of the suspect reference set.

For the calibration, we select a representative population of 129 individuals (who were not among the 66 individuals selected for creating the mock-up forensic cases) who are present in all the four sessions `01,02,03,04` of the MultiPIE dataset. The calibration trace set (Step 1) contains images of 129 individuals having quality `(01,19_1,18)` and the calibration reference set (used in Step 2) contains images of the same 129 individuals but is taken from sessions `02,03,04`.

In this experiment, the four different datasets are taken from different sessions in the MultiPIE dataset. This simulates the session variation that is inevitable in real forensic cases.

For our mock-up forensic cases, we now describe each of the four processing stages of our framework in detail:

Step 1: In all the mock-up forensic cases, we avoid the need for QA and QM modules because the trace has been taken from the MultiPIE dataset for which the capture setup is well defined. Therefore, we know that the trace has the quality `(03,19_1,18)`.

As stated earlier, we create a calibration trace set (as shown in Fig. 4) of the representative population having quality similar to the trace from the population of the 129 individuals selected for the calibration in which the ground truth is known.

Fig. 4: Some images from the calibration trace set having image quality, `(01,19_1,18)`, similar to that of the trace.

[1] `(session-id,camera-id,flash-id)` denotes quality of MultiPIE face images where, `session-id` is not a quality parameter.

Step 2: The MultiPIE dataset samples face image of 129 individuals, selected as representative population for the calibration, from a discrete set of camera and flash positions as shown in [4, Fig. 4]. Therefore, instead of iterating over the full image quality space of the calibration reference set, we only consider discrete pose and illumination as defined for the MultiPIE dataset. Ideally, search for optimal quality of calibration reference set should consider all the possible image quality variations.

During the search for optimal quality of the calibration reference set, quality of the calibration trace set remains constant. Also, during this search, the calibration reference set contains images from all the remaining three sessions (i.e. `02,03,04`) of the same 129 individuals in order to include the effect of session variation to recognition performance.

ROC plots for different combinations of the calibration trace and reference set quality is shown in Fig. 5. Due to space limitations, we only show ROC curve for a few calibration reference set quality variations. We observed further degradation in recognition performance for all the remaining pairs of image quality.

The ROC plots in Fig. 5, show that the optimal recognition performance occurs when the calibration trace set and the calibration reference set contain face images captured by the same camera (i.e. `19_1`). Also, in such a case, illumination variation has major contribution towards recognition performance.

Near-optimal recognition performance is obtained when there is small pose variation[2] between the calibration trace set (`19_1`) and the calibration reference set (`19_0`) set face images.

It is important to realise that, due to the nature of MultiPIE dataset, when we match camera between calibration trace set and calibration reference set images, we are not only matching the pose but also matching the imaging condition (camera response, resolution, distortion, etc.). In practical forensic cases, it is usually possible to acquire the camera that captured the trace to capture the suspect reference set. However, in practical calibrations, it is difficult to match both pose and illumination between the trace and suspect reference set. Moreover, it is also difficult to obtain a trace set and reference image set corresponding to the circumstances of the case. Therefore, we analyse further results using the following four cases of calibration reference set quality that are representative of possible scenarios in a real forensic case: *a*) when both camera and illumination match between trace and reference set – ideal case; *b*) only camera matches; *c*) only illumination matches; and, *d*) neither camera nor illumination matches

The ROC curve corresponding to these four possible cases of suspect reference set image quality is shown in Fig. 7. For reference, we have also depicted the baseline performance of [1] when subject to frontal pose and illumination face images of 346 individuals across the four sessions of the MultiPIE dataset in Fig. 7. With this baseline performance ROC, it is evident that the LR computed from performance of a particular face recognition system based on frontal face images of standard face image dataset will lead to a large number of misleading interpretation of the evidence (E).

In Fig. 7, we also observe significant degradation in the recognition performance, when quality of calibration reference set is frontal face images (randomly chosen without using calibration of our framework) for a profile view calibration trace set. Therefore, without calibration based on pair of image quality, the recognition rate drops significantly (i.e. higher rate of misleading interpretation of the evidence) thereby making any evidence from forensic face recognition unusable in the court of law.

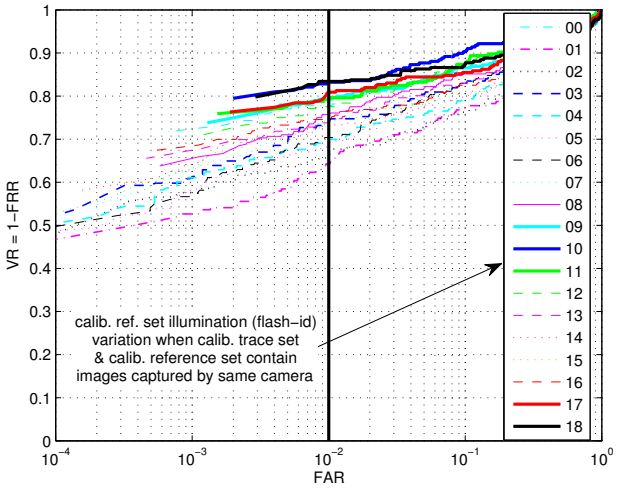
Step 3: With the optimal quality of calibration reference set (for a given calibration trace set quality) to hand, we now shift our attention to the task of matching the trace to the set of 66 suspect images having the quality of optimal calibration reference set (as determined in Step 2). For instance, if the sub-optimal calibration reference set quality refers to the case when neither camera nor illumination matches, the suspect reference set quality is `(04,19_0,17)`.

In Fig. 6, we show one such suspect reference set having quality `(04,19_0,17)` (i.e. neither camera nor illumination matches between calibration trace set and calibration reference set)

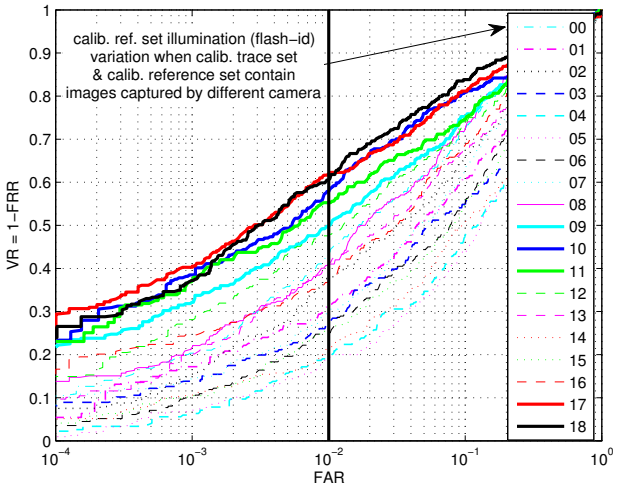
Fig. 6: The case when neither camera nor illumination matches between the trace and the suspect reference set.

Step 4: Recall that, evidence (E) is the similarity score value when the trace Y_k is matched against the putative source X_k, where $k = \{1, 2, \cdots, 66\}$. The generic "same-source" (true matches) and "different-source" (false matches) distributions, that generated the generic ROC of Fig. 7, are used to estimate likelihood ratio (LR) value which is given by $LR = \frac{w_k}{b_k} \times M$, where, k is the score interval corresponding to evidence (E), w_k and b_k are score count for k^{th} interval of generic "within-source"

[2] the angular difference between `19_1` and `19_0` camera positions is $(\theta = 25.9°, \phi = 0.3°)$



(a) Trace = (01,19_1,18), Ref. Set = ({02,03,04},**19_1**,{⋆}).



(b) Trace = (01,19_1,18), Ref. Set = ({02,03,04},**19_0**,{⋆}).

Fig. 5: Some ROC obtained during calibration: search for optimal quality of the calibration reference set (in terms of illumination [flash-id] and pose [camera-id]) when quality of the calibration trace set remains fixed to (01,19_1,18).

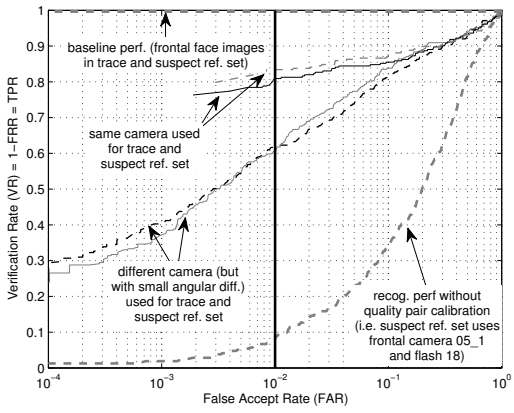


Fig. 7: ROC of four calibration reference set qualities with optimal and near-optimal recognition performance.

and "between-source" distribution respectively, and, $M = \sum w_i / \sum b_i$.

For all the 66 mock-up forensic cases, we depict the rate of misleading interpretation of evidence (E) in the Tippet plots of Fig. 8a and 8b.

4.1. Discussion

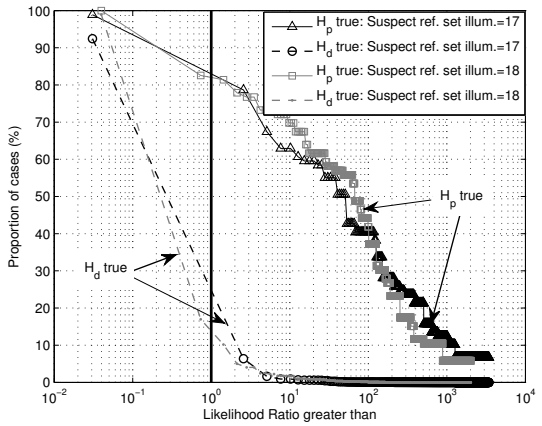
An ideal LR based framework for forensic face recognition generates LR > 1 when H_p is true while LR values < 1 is generated when H_d is true. The framework causes misleading interpretation of evidence (E) when exact opposite LR values are generated.

In Table 1, we show evidence (E) and corresponding LR value for one of the 66 mock-up forensic cases. In this case, the trace contains face image of individual 063 and the suspect reference set consists of 66 individuals (including the person 063). This table shows that our framework supports correct hypothesis when camera (and hence, pose) matches between trace and suspect reference set. When there is a mismatch, the framework causes misleading interpretation of evidence (E) by generating LR value > 1 when H_d is true.

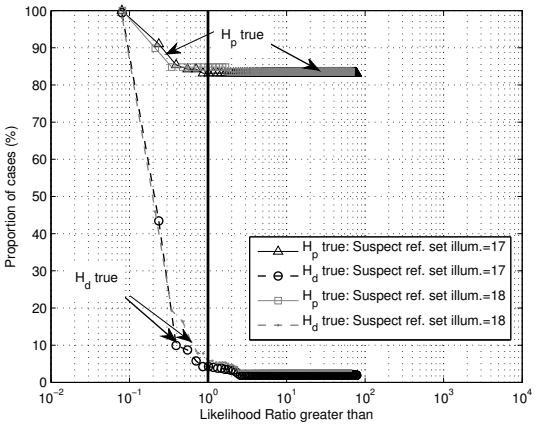
When the trace and the suspect reference set images are captured by same camera (and hence have same pose), we observe reduction in the rate of misleading interpretation of evidence (E) as shown in Fig. 8b. Moreover, these two Tippet plots also show that slight mismatch in illumination between trace and suspect reference set does not significantly contribute to the rate of misleading interpretation of evidence (E).

Table 1: Top 3 Likelihood Ratio (LR) and corresponding evidence (E $\in [0,1]$) when the facial image of individual (063) having quality (03,19_1,18) is the trace (Y)

	Suspect Reference Set Quality			
	Different cam. (i.e. 19_0)		Same cam. (i.e. 19_1)	
	flash 17	flash 18	flash 17	flash 18
(X_1)	(063) : 0.85	(063) : 0.88	(063) : 0.99	(063) : 0.99
(X_2)	(261) : 0.55	(286) : 0.57	(016) : 0.61	(255) : 0.03
(X_3)	(286) : 0.50	(222) : 0.39	(222) : 0.61	(113) : 0.64
	Top 3 LR values when		H_p : X_k is the source of Y H_d : X_k not the source of Y	
k = 1	508.58	252.85	78.14	66.02
k = 2	117.36	99.33	0.69	0.71
k = 3	31.78	4.14	0.69	0.50



(a) Suspect Ref. Set Quality = ({02,03,04},**19_0**,{⋆}).



(b) Suspect Ref. Set Quality = ({02,03,04},**19_1**,{⋆}).

Fig. 8: Tippett plots depicting rate of misleading interpretation of evidence when our framework is applied to 66 mock-up forensic cases.

5. Conclusion

Quality of a trace and capabilities of a particular face recognition system determines the limit of recognition performance achievable in a forensic face recognition case. Preliminary tests on a limited face image dataset (i.e. MultiPIE dataset with 9 viewpoints and 18 illuminations) using a commercial face recognition system [1] shows that such a limit of recognition performance can be achieved by calibrating the quality (in terms of pose, illumination and imaging device) of the trace and the suspect reference set. Our results also affirms the notion that image quality, when assumed of being predictive of recognition performance, is the property of image pair participating in the recognition process and not of an individual image [2].

We also observed gradual improvement in recognition performance when the quality of reference set approached the quality of the trace set. This behaviour is encouraging for practical forensic face recognition cases, where exact match of pose and illumination between the trace and suspect reference set is not possible. Therefore, we can expect near-optimal recognition performance even when the two image qualities have small mismatch.

Our results also show that for a face image quality typical in a CCTV surveillance footage, [1] has very low recognition performance resulting in high rate of misleading interpretation of evidence. Most commercial face recognition systems are not designed for noisy and profile view face images mostly captured by a CCTV camera. With the improvement in face recognition technology and quality of CCTV face images, we believe that the rate of misleading interpretation of evidence would reduce to the extent that our framework is applicable for real forensic face recognition cases.

Mainly, there are two limitations of our framework. First, is that the practical implementation of the Quality Assessor (QA) and Quality Modifier (QM) module is critical to the practicability of this framework. To the best of our knowledge, it is still an open problem. Second, searching the quality space of the calibration reference set is a computationally expensive process. However, this does not limit the practicability of our framework because forensic evaluation do not have real-time requirements.

References

[1] Cognitec FaceVACS SDK version 8.4.0, 2010.

[2] P. J. Beveridge, J. R.; Phillips, G. H. Givens, B. Draper, M. N. Teli, and D. Bolme. When high-quality face images match poorly. *The Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*, page 7, 2011.

[3] X. Gao, S. Li, R. Liu, and P. Zhang. Standardization of face image sample quality. In S.-W. Lee and S. Li, editors, *Advances in Biometrics*, pages 242–251. Springer Berlin / Heidelberg, 2007.

[4] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–8, 2008.

[5] Q. Kwan. *Inference of identity of source*. PhD thesis, University of California: Berkeley, CA, 1977.

[6] D. Meuwly. Forensic individualisation from biometric data. *Science & Justice*, 46(4):205–213, 2006.

[7] D. Meuwly and A. Drygajlo. Forensic speaker recognition based on a bayesian framework and gaussian mixture modelling (gmm). In *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[8] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 346 –353, march 2011.