Predicting Dactyloscopic Examiner Fingerprint Image Quality Assessments

Martin Aastrup Olsen[1], Martin Böckeler[2], Christoph Busch[2]

[1]Faculty of Computer Science and Media Technology, Gjøvik University College, Norway, martin.olsen@hig.no

[2]da/sec Biometrics and Internet Security Research Group, Hochschule Darmstadt, Darmstadt, Germany, christoph.busch@h-da.de, martin.boeckeler@googlemail.com

Abstract: We work towards a system which can assist dactyloscopic examiners in assessing the quality and decision value of a fingerprint image and eventually a fingermark. However when quality assessment tasks of datyloscopic examiners are replaced by automatic quality assessment then we need to ensure that the automatic measurement is in agreement with the examiner opinion. Under the assumption of such agreement, we can predict the examiner opinion. We propose a method for determining the examiner agreement on ordinal scales and show that there is a high level of agreement between examiners assessing the ground truth quality of fingerprints. With ground truth quality information on 749 fingerprints and using 10-fold cross validation we construct models using Support Vector Machines and Proportional Odds Logistic Regression which predicts median examiner quality assessments 35% better than when using the prior class distribution.

1 Introduction

Fingerprint sample quality in the context of forensic applications where an examiner following the Analysis, Comparison, Evaluation, and Verification (ACE-V) protocol is part of the initial information gathering phase where the examiner studies the impression to quantify the present discriminating information and assess the quality and completeness [Exp12].

The quality assessment will among other factors such as the completeness of the fingerprint have an impact on the decision value of the impression, which can be one of: Value for Individualization (VID), which is used when the quantity and quality of the information present is deemed sufficient to determine if the impression is from the same source as another, yet unseen, impression; Value for Exclusion Only (VEO) is used when sufficient information is present to determine that the impression is not from the same source as another impression; No Value (NV) is used when the impression is deemed unsuitable. The process of assigning VID, VEO, and NV is inherently subjective and requires training and experience to perform accurately and consistently. Ulery et al. conducted a study on the accuracy and reliability of 169 forensic examiners who assigned VID, VEO, and NV decisions to 100 latents and found that VID decisions were unanimous in 48% of cases for mated pairs and 33% for non-mated pairs [UHBR11]. In a related study on the

repeatability and reproducibility of decision by individual examiners, it was found that 93% of VID, 85% NV, and 55% VEO decisions were repeated by individual examiners when presented twice with the same impression after a 7 month interval [UHBAR12]. These findings, which indicate a high degree of accuracy with respect to VID, and lower accuracy with respect to VEO mirror the findings by Langenburg [Lan09].

It has been demonstrated that when provided with extraneous contextual information, an examiner might change the method of judging and comparing fingerprints [DCP05]. Additionally, the examiners are vulnerable to biasing information such as evidence of confession of a crime, even in cases where the comparison of the impressions is non-difficult [DC06].

Biometric sample quality has successfully been applied in the context of Automated Fingerprint Recognition Systems (AFIS) to reject samples which are likely to contribute negatively to False Non-Match Rates (FNMR). By rejecting those samples before they are enrolled, a high level of biometric performance is achievable and with it higher levels of satisfaction by the users interacting with the biometric system [WGW04].

We are motivated by the successful application of automated quality assessment in AFIS and by the findings of Ulery et al. and Bradford et al., which highlight the subjective nature of the ACE-V protocol, to determine methods which objectively assesses the quality of an impression in the form of a fingerprint or a fingermark. This paper represents one step in this direction and our main objective is to determine methods which predict the quality assessment that a dactyloscopic examiner gives a particular fingerprint. To achieve this, we leverage a dataset which contains ground truth quality labels on inked impressions as assessed by dactyloscopic examiners from the German Federal Criminal Police Office (BKA), and apply quality assessment algorithms identified or developed in the context of NFIQ 2.0 [Nat14].

The rest of the paper is organized as follows: section 2 outlines state of the art methods for objective quality assessment. In section 3 we propose a method for quantifying examiner agreement, section 4 details the ground truth dataset on which we base our experiments that are described in section 5. We discuss our results in section 6 and conclude in section 7.

2 Fingerprint quality

There exists a large number of fingerprint image quality assessment algorithms in the literature and several reviews have been made, e.g. in the context of optical and capacitive sensors [AFRM$^+$07, AFRM$^+$08]; relation between quality assessed by human experts and algorithms and comparison scores [FAMSAFOG05]; and more recently quality assessment using no-reference algorithms have been investigated [EANCR13]; a comprehensive review of biometric sample quality is provided by Bharadwaj et al. [BVS14] and a quality metric for fingermarks has been proposed [YCLJ13].

We have selected a subset based on those features specified in the NFIQ 2.0 quality feature definitions document version 0.5 [NFI12] which we summarize here: Frequency Domain Analysis (Q_{FDA}) uses the magnitude of the dominant frequency as determined by discrete Fourier transform fingerprint image as a local quality value. Local Clarity Score (Q_{LCS})

determines the clarity of the fingerprint image by estimating how well each block in the fingerprint image can be segmented into ridge and valley region. Orientation Flow (Q_{OFL}) measures the continuity of ridge flows in the fingerprint image by determining the dominant ridge orientation agreement between one block and its neighbouring blocks. Orientation Certainty Level (Q_{OCL}) is a measure of the strength of the ridge orientation within a image block. A high score indicates that the ridge orientation within the block is well defined. Ridge Valley Uniformity (Q_{RVU}) measures the consistency between ridge and valley widths within each image block. The widths of ridges and valleys are expected to be similar across the entire fingerprint image. Radial Power Spectrum (Q_{RPS}) quantifies the energy concentration within a specified band in the Fourier spectrum. The limits of the band are determined by the expected ridge valley frequency. Image mean (Q_{MU}) is the mean value of pixel intensities across the fingerprint image. Image standard deviation (Q_{STD}) is the standard deviation of pixel values across the fingerprint image.

The features Q_{FDA}, Q_{LCS}, Q_{OFL}, Q_{OCL}, Q_{RVU} operate on 32 by 32 pixel non-overlapping regions of the image and thus provide a vector of local quality values. Q_{RPS}, Q_{MU}, Q_{STD} work on the entirety of the image and provide a scalar value. Based on these two groups we define two sets of features: Set A which contains the mean and standard deviation of each of the local quality vectors of Q_{FDA}, Q_{LCS}, Q_{OFL}, Q_{OCL}, Q_{RVU} giving a total of 10 features; set B contains, in addition to the features in set A, Q_{RPS}, Q_{MU}, Q_{STD} for a total of 13 features.

3 Quantifying examiner agreement on ordinal scales

To quantify examiner agreement, which is an essential task to determine to which degree examiners agree on what quality means and to compare and judge how well automatic quality prediction will perform against its human counterpart, the following requirements are specified for an examiner agreement coefficient:

1. an unlimited number of assessments on a single fingerprint shall be considered
2. the agreement of assessments shall be weighted according to their distance
3. assessments that belong to the same decision categories shall be assigned with high weights
4. when the assessment scale varies the measure results shall remain consistent

Ad 1 - the coefficient shall be capable of measuring agreement for fingerprints that were annotated by a minimum of 2 examiners and for fingerprints that are annotated by more than 2 examiners. Ad 2 - assessments that are "closer" to each other shall result in higher agreement. Ad 3 - not only the "distance" of single assessments shall be measured, assessments in equal decision categories shall be weighted higher than assessments that are not in the same decision category. Ad 4 - quality assessments in varying assessment scales shall result in the same agreement if the relative distance of the single ratings are the same. For example, a quality assessment of $x_{11} = 1, x_{12} = 2, x_{13} = 5$ $\{x_{1x} \in \mathbb{N} | 1 \leq x_{1x} \leq 5\}$ and a quality assessment of $x_{21} = 1, x_{22} = 25, x_{23} = 100$ $\{x_{2x} \in \mathbb{N} | 1 \leq x_{2x} \leq 100\}$ shall result in the same agreement coefficient.

Several common statistical measures like the Percentage agreement $\bar{P}$ [UHBR12], the Interquartile Range IQR [ZK99, p.27], the Median Absolute Deviation MAD [HMT82,

p.220] and the Standard Deviation SD [ZK99, p.26] were investigated to determine if they fulfil the specified requirements. Table 1 displays how these measures perform on various assessment examples. It is clear that none of the common measures are able to measure examiner agreement sufficiently as they all violate one of the predefined requirements.

		Assessment example							
		1	2	3	4	5	6	7	8
Assessment	excellent, 1	❶❷❸	❶❷	❶❷	❶	❶	❶	❶❷❸❹❺	❶❷
	very good, 2		❸		❷	❷			❸
	good, 3			❸	❸		❷		❹
	fair, 4								❺
	poor, 5					❸	❸	❻	❻
Agreement	$\bar{P}$	1.000	0.333	0.333	0.000	0.000	0.000	0.667	0.067
	IQR	0.000	1.000	2.000	2.000	4.000	4.000	0.000	3.000
	MAD	0.000	0.000	0.000	1.000	1.000	2.000	0.000	1.500
	SD	0.000	0.471	0.943	0.816	1.700	1.633	1.491	1.491
	$CMCA$	1.000	0.839	0.689	0.422	0.166	0.125	0.765	0.208

Table 1: Examples of agreement values computed using $\bar{P}$, IQR, MAD, SD, $CMCA$ for 8 cases of examiner assessments where the number of experts and their assessments vary. In the top half of the table, each dot represents an examiner assessment ranging from excellent (1) to poor (5). The bottom half of the table shows the agreement value assigned by the 5 metrics for each assessment example.

$\bar{P}$ is equal for assessment examples 2 and 3, and examples 4, 5 and 6 thus violating requirement 2. IQR is equal for the assessment examples 3 and 4 violating requirements 2 and 3 as also for assessment example 5 and 6 which violates requirement 2. MAD is equal for the first 3 examples thus violating requirement 2 and 3 and in the 4th and 5th example requirement 2 is violated. SD violates requirement 3 in examples 2, 3, 7 and 8. Requirement 4 is violated by IQR, MAD, SD as the produced values depend on the range of the scale.

None of the measures described above satisfy the task of measuring examiner agreement as specified by our requirements, and hence we propose a new coefficient which does fulfil the specified requirements: Closest-neighbour Median Cluster Agreement ($CMCA$).

$CMCA$ consists of multiple parts which are calculated as follows. Let x_j be the j^{th} ascending sorted rating on the fingerprint. The closest neighbour distance $ND_j \in [0, 1]$, one part to fulfil requirement 2, of the j^{th} rating on the fingerprint is:

$$ND_j = \min\left(|x_j - x_{j-1}|, |x_j - x_{j+1}|\right) \tag{1}$$

Further, let $\max i$ the maximum possible rating on the rating scale, $\min i$ the minimum possible rating on the scale, $\tilde{x}$ the median of ratings and r the number of ratings on the fingerprint. The distance consensus $D \in [0, 1]$ of the fingerprint made to fulfil requirements

2 and 4 is:

$$D = \frac{\left(\sum_{j=1}^{r}\left(1 - \frac{ND_j}{\max i - \min i}\right) + \left(1 - \frac{|x_j - \tilde{x}|}{\max i - \min i}\right)\right) - 1}{2r - 1} \tag{2}$$

Let c_j be the j^{th} cluster (ratings in the same decision category) and let $|c_j|$ be its cardinality. Let nc be the number of clusters on. The average cluster size difference $CSD \in [0, r-2]$, which compares the size of each assessment cluster against each other and calculates the mean of their size differences is:

$$CSD = \frac{\sum_{j=1}^{nc}\sum_{k=j}^{nc}||c_j| - |c_k||}{\frac{(nc-1)\cdot nc}{2}} \tag{3}$$

To fulfil requirement 3, the distance consensus D is multiplied by the cluster size power $CSP \in [1, r-1]$, which is calculated as:

$$CSP = nc - \frac{CSD}{r} \tag{4}$$

The $CMCA$ on the fingerprint is finally calculated by:

$$CMCA = \begin{cases} 1.0 & , nc = 1 \\ 1.0 - \frac{|x_1 - x_2|}{\max i - \min i} & , r = 2 \\ D^{CSP} & , \text{otherwise} \end{cases} \tag{5}$$

The $CMCA \in [0, 1]$ over a set of fingerprints I is calculated as the arithmetic mean over the $CMCA$ agreements of each fingerprint in I.

The measurement results of $CMCA$ applied to the 8 assessment examples are shown in the last row of table 1. Assessment examples 5 and 6, show that $CMCA$ fulfils requirement 2 by considering the distance of between single assessments. It also fulfils requirement 3, heavy weighted assessments that belong to the same decision category. The example shows that the $CMCA$ fulfils requirement 1, measuring agreement for any number of assessments on a single fingerprint.

After all, the $CMCA$ was designed to measure examiner agreement per fingerprint, see eq. (5). Other coefficients like Cohens Kappa [Coh60] or Fleiss Kappa [Fle71] were designed to measure inter-rater agreement over the whole assessment population which is expressed by the arithmetic mean of the set of $CMCA$. Nonetheless, Cohens Kappa has the disadvantage that it can only measure pairwise rater agreement on nominal data. In addition, Fleiss Kappa is able to measure inter-rater agreement for more than 2 raters, but was also designed for nominal data, so additional information of the natural order of categories on ordinal data like it is present in our case of fingerprint quality, would not be considered properly.

4 Ground truth data

The ground truth data of this paper, representing the human examiner quality assessment part, originates form the NIST special database 14 [Wat93], containing 54 000 fingerprints from live-scan and scanned ink impressions and from the NIST special database 29 [Wat01], containing 4 320 fingerprints. In 2009, a team of 9 dactyloscopic examiners from the BKA annotated for the purpose of a conformance testing study [BLTK09] several fingerprint characteristics, like minutia points, singular points and the overall fingerprint quality which is investigated in this paper. To establish an objective annotation, the examiner team was equipped with a simple graphical user interface that omitted the support of the automatic or semi-automatic minutia extraction functionality of the AFIS. To further increase the objectivity and anonymity of the process, each examiner was assigned with an ID that was not known to other examiners. The examiners rated the overall fingerprint quality within 5 decision categories, ranging from excellent (1), very good (2), good (3), fair (4) to poor (5). A total of 749 fingerprints were annotated by at least 2 examiners. Figure 1 shows the logical partitioning of the annotated samples, where the first level contains the set of 749 fingerprints; the second level shows how many samples were annotated grouped by the number of examiners. The third level shows the number of samples each distinct group of examiners annotated, e.g. there were 713 fingerprints annotated by 3 examiners, which came from two groups of 3 examiners annotating respectively 361 and 352 fingerprints ($\mathbf{S}_2$ and $\mathbf{S}_3$). Table 2 shows the number of images that were annotated by each of the 9 examiners. We note that examiners 11 to 16 have each annotated nearly 400 samples, while examiners 17 to 19 have annotated fewer than 20 samples.

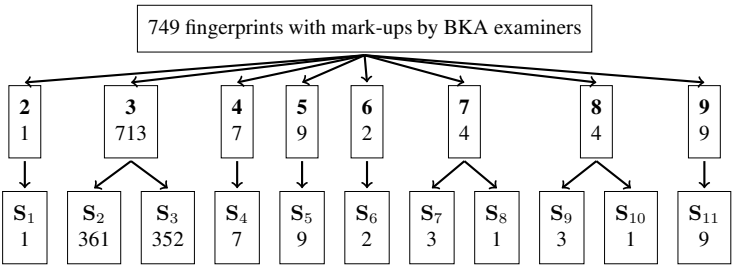

Figure 1: Examiner markup tree. The bold digits in the second tree level display how many examiners annotated a specific number of fingerprints. The last tree level shows how many fingerprints where annotated by distinct groups of examiners.

	Examiner								
	11	12	13	14	15	16	17	18	19
Annotations	396	393	397	378	388	371	10	17	17

Table 2: The number of images annotated by each of the 9 examiners.

Table 3 summarizes the $CMCA$ computed on median ground truth quality levels and for all fingerprint images in the ground truth data set. The mean $CMCA$ is .88 across all

		$CMCA$				
Quality	n	min	max	mean	median	std
1	41	.69	1.00	.82	.84	.08
2	306	.42	1.00	.86	.84	.15
3	305	.42	1.00	.90	.84	.10
4	92	.42	1.00	.88	.84	.11
5	5	.84	.84	.84	.84	.00
All	749	.42	1.00	.88	.84	.13

Table 3: Summary statistics of $CMCA$ at each median ground truth quality level and across all samples.

fingerprint images, indicating that there is generally consensus between examiners when subjectively assessing the quality level of a fingerprint image. At the individual levels we note some differences in the $CMCA$, in particular that mean $CMCA$ at quality level 1 is .82 while for level 3 it is .90.

Figure 2 shows an example fingerprint for each median quality assessment. For median quality levels 1 through 4 the examiners were in agreement with $CMCA = 1.0$ while for the fingerprint illustrating level 5 where $CMCA = 0.838$. We observe that at level 1 (left most figure) the ridge lines are clearly separated around the core, where the examples for levels 4 and 5 appear without clearly defined ridge lines and with blurry or low contrast regions.

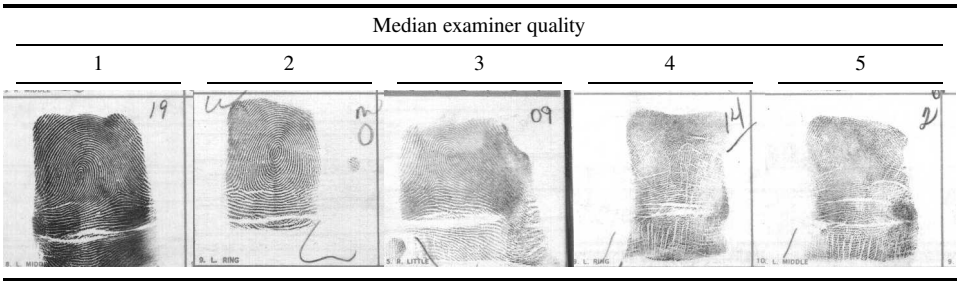

Figure 2: Median examiner quality assessment with example fingerprints. All images are from the NIST special database 14 [Wat93] (file names from left to right: f0000118, f0000109, f0000095, f0000969, f0000968).

To assess whether human examiner quality assessments are indicative of the eventual genuine comparison score of the fingerprint sample, we computed the genuine comparison scores for all samples and grouped them according to the examiner who made the assessment and the quality score that was assigned. Box plots showing the relation between assigned quality and genuine comparison score [id314] for each of the 6 examiners who annotated the most images (see table 2), as well as the median examiner quality are depicted in fig. 3. The plots show that generally a higher quality score is associated with a higher comparison score, however, for some cases we note some irregularities. For Ex-

aminer 11 (left most plot), we note that no images were assigned quality level 5 and that those samples which received a quality level of 1 were involved in comparisons resulting in scores similar to those samples which were assessed as being quality level 2.

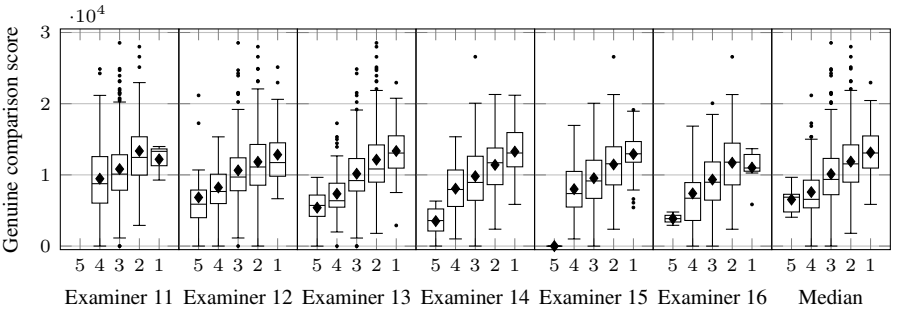

Figure 3: Boxplots of genuine comparison score for each of 5 quality levels assigned by 6 examiners and for the median of assigned quality levels.

5 Experiments

Our goal is to predict the quality level that an examiner will assign to a given fingerprint. We note from table 3 that our proposed $CMCA$ coefficient indicates a high degree of agreement between examiners as to the assigned ground truth quality levels for the data set.

We perform a series of experiments in order to assess to which degree the quality scores assigned by individual examiners or the median quality score is predicted. From table 3 we note that the distribution of median quality levels is not uniform with the majority of samples being assigned levels 2 or 3 and with only 5 samples in level 5. Due to the low annotation count by examiners 17, 18, and 19 we do not attempt to predict their assessments (see table 2) and instead only assess examiners 11 to 16 individually.

We train our predictive models using Multi-class Support Vector Machine (SVM) [CV95] and Proportional Odds Logistic Regression (POLR) [MZ75] where the response variable is the assigned quality level (either individual examiner or median of examiners), and the explanatory variables are features in sets A or B (see section 2).

The experiments are performed using 10-fold cross validation, i.e., we divide the available data in each experiment randomly into 10 disjunct partitions of equal size. Over the 10 possible permutations we perform training of SVM and POLR on the 9 folds and test the performance on the remaining fold. In the case of SVM we use Radial Basis Function as kernel and perform a grid search for optimal sample influence radius (γ) and cost (C) over $\gamma \in \{0.001, 0.01, ..., 1000\}$ and $C \in \{0.001, 0.01, ..., 1000\}$ given the training folds. In the case of POLR no parameter optimization is performed.

The predictive capability of the constructed models in each experiment setting is determined by calculating the mean and standard deviation of the F_1 score and Cohen's Kappa

(κ) [Coh60] over the 10 permutations.

Cohen's Kappa quantifies the class agreement between the model predictions and the ground truth by taking into account the observed probabilities of the classes. When $\kappa = 0$ the agreement is equal to that which can be achieved by random chance based on the priors - when $\kappa = 1$ then the agreement is complete. F_1 is the harmonic mean of precision and recall and the lowest score is achieved when $F_1 = 0$ and highest when $F_1 = 1$.

Experiments were performed using R [R C14] with SVM from e1071 [MDH+14]; POLR from MASS [VR02]; cross validation and miscellaneous functions from caret [KWW+15] and xtable [Dah14] packages.

6 Results

Our analysis of the ground truth data set (section 4) showed that there is a high level of agreement in assessing quality levels across examiners, and that the higher quality levels are associated with higher genuine comparison scores.

Following the protocol described in section 5 we performed a total of 14 experiments and have summarized the mean and standard deviations of F_1 and κ for each of them in table 4. Each line in the table alternates between feature set *A* and *B* in *Set* column with the *Target* column indicating what is being predicted where *Median* indicates that it is the median of examiner quality assessments per fingerprint that is predicted, and *Examiner 11* indicates that it is the quality assessments of Examiner 11 which are predicted. The remaining columns are first grouped by F_1 and κ, next by method *SVM* or *POLR* and finally arithmetic mean (*mean*) and standard deviation (*std*) of testing results across the 10 fold cross validation.

		F_1				κ			
		SVM		POLR		SVM		POLR	
Set	Target	mean	std	mean	std	mean	std	mean	std
A	Median	.58	.05	.59	.06	.31	.07	.33	.07
B	Median	.60	.08	.60	.06	.34	.12	.35	.07
A	Examiner 11	.72	.05	.70	.03	.21	.17	.13	.10
B	Examiner 11	.72	.05	.71	.06	.28	.14	.18	.16
A	Examiner 12	.45	.06	.48	.05	.17	.07	.23	.07
B	Examiner 12	.51	.05	.52	.07	.26	.08	.29	.08
A	Examiner 13	.49	.07	.49	.08	.21	.09	.21	.12
B	Examiner 13	.53	.08	.52	.08	.30	.10	.26	.09
A	Examiner 14	.57	.10	.59	.09	.30	.14	.33	.13
B	Examiner 14	.57	.09	.60	.09	.29	.14	.36	.13
A	Examiner 15	.57	.06	.60	.09	.29	.08	.35	.14
B	Examiner 15	.58	.07	.61	.08	.32	.09	.37	.12
A	Examiner 16	.63	.11	.64	.09	.33	.21	.36	.16
B	Examiner 16	.64	.12	.65	.11	.36	.21	.38	.19

Table 4: Results of experiments in predicting median and individual examiner quality assessment using SVM and POLR on feature sets A and B.

The best prediction results when using Median as target is achieved with POLR and feature

set B when considering either of F_1 and κ as evaluation criteria. We see that the mean F_1 is .60 for both SVM and POLR, but the standard deviation is smaller in the case of POLR with .06 over .08 achieved with SVM. For κ we note a mean of .35 for POLR and .34 for SVM, again with a smaller standard deviation favouring POLR.

We note that both F_1 and κ spans a wide range when using individual examiner quality assessments as target for the predictions. Examiner 12 appears to be hardest to predict with a mean κ of .17 and .23 and F_1 of .45 and .58 for respectively SVM and POLR when using feature set A. Prediction of the scores assigned by Examiner 16 using feature set B gives the highest mean F_1 of .64 and .65 and κ of .36 and .38, however in both cases the standard deviation over the 10 folds is the highest of the experiments performed.

Generally using the global features present in feature set B lead to marginal increases in the mean F_1 score while κ is increased further for both SVM and POLR.

In addition to SVM and POLR listed in table 4 we also used Recursive Partioning, however that algorithm had difficulty working with the relatively small dataset with 10-fold cross validation and was thus not able to complete all intended experiments.

7 Conclusion

In this paper we have made steps towards assisting dactyloscopic examiner in assessing the quality of a given fingerprint or fingermark with the aim of determining its decision value in the ACE-V protocol. We address the objective quality of fingerprints and relation to examiner opinion with a continued goal to extend the quality assessment evaluation to fingermarks which pose the greatest challenge in the forensic evaluation.

We proposed the $CMCA$ coefficient as a general method for quantifying examiner agreement on ordinal scales containing any number of categories. Using $CMCA$ we have shown that there is a high level of agreement between examiners as to what constitutes a high quality fingerprint and further that the ground truth assessments made by dactyloscopic examiners are indicative of genuine comparison scores.

On our limited dataset containing 749 finger images and using 13 quality features we have constructed a model which predicts examiner quality assessments around 35% better than random chance given the prior quality class probabilities as measured using Cohen's Kappa.

Our future work includes refinement of the quality feature set to improve the predictive capabilities; evaluation of the importance of the individual features to gain insights as to which image covariates are important to examiners; evaluation of the trained model to determine the degree that the assigned quality levels are indicative of biometric performance; and finally how the trained models can be incorporated in ACE-V protocol to assist the decision making of dactyloscopic examiners when working with fingerprints or fingermarks.

8 Acknowledgements

We thank the team of dactyloscopic examiners with the German Federal Criminal Police Office for providing ground truth fingerprint information and National Institute of Stan-

dards and Technology for providing the selection of images from SD14 and SD29.

References

[AFRM+07] Fernando Alonso-Fernandez, Fabio Roli, Gian Luca Marcialis, Julian Fierrez, and Javier Ortega-Garcia. Comparison of fingerprint quality measures using an optical and a capacitive sensor. In *Proc. IEEE Conference on Biometrics: Theory, Applications and Systems, BTAS*, pages 1–6, September 2007.

[AFRM+08] F. Alonso-Fernandez, F. Roli, G.L. Marcialis, J. Fierrez, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Performance of fingerprint quality measures depending on sensor technology. *SPIE Journal of Electronic Imaging, Special Section on Biometrics: Advances in Security, Usability and Interoperability*, 17(1), January-March 2008.

[BLTK09] Christoph Busch, Dana Lodrova, Elham Tabassi, and Wolfgang Krodel. Semantic Conformance Testing for Finger Minutiae Data. *Proceedings of the 1st International Workshop on Security and Communication Networks (IWSCN)*, pages 1 – 7, 2009.

[BVS14] Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. Biometric quality: a review of fingerprint, iris, and face. *EURASIP Journal on Image and Video Processing*, 2014(1), 2014.

[Coh60] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[Dah14] David B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2014. R package version 1.7-4.

[DC06] Itiel E. Dror and David Charlton. Why experts make errors. *Journal of Forensic Identification*, 56(4):600, 2006.

[DCP05] Itiel E. Dror, David Charlton, and Ailsa E. Péron. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, 156(1):74–78, January 2005.

[EANCR13] M. El Abed, A Ninassi, C. Charrier, and C. Rosenberger. Fingerprint quality assessment using a no-reference image quality metric. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5, Sept 2013.

[Exp12] Expert Working Group on Human Factors in Latent Print Analysis. *Latent print examination and human factors : improving the practice through a systems approach*. US Department of Commerce, National Institute of Standards and Technology, February 2012.

[FAMSAFOG05] J. Fierrez-Aguilar, L. M. Muñoz-Serrano, F. Alonso-Fernandez, and J. Ortega-Garcia. On the effects of image quality degradation on minutiae- and ridge-based automatic fingerprint recognition. In *Proc. IEEE Intl. Carnahan Conf. on Security Technology, ICCST*, pages 79–82, October 2005.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[HMT82] David C. Hoaglin, Frederick Mosteller, and John W. Tukey, editors. *Understanding Robust and Exploratory Data Analysis (Wiley Series in Probability and Statistics)*. Wiley, 1 edition, 12 1982.

[id314] id3 Technologies. id3 Fingerprint SDK v. 1.4.1. http://www.id3.eu/, 2014.

[KWW+15] Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, and Luca Scrucca. *caret: Classification and Regression Training*, 2015. R package version 6.0-41.

[Lan09] Glenn Langenburg. A Performance Study of the ACE-V Process: A Pilot Study to Measure the Accuracy, Precision, Reproducibility, Repeatability, and Biasability of Conclusions Resulting from the ACE-V Process. *Journal of Forensic Identification*, 59(2):219–257, 2009.

[MDH+14] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-4.

[MZ75] Richard D. McKelvey and William Zavoina. A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, 4(1):103–120, 1975.

[Nat14] National Institute of Standards and Technology. Development of NFIQ 2.0. http://www.nist.gov/itl/iad/ig/development_nfiq_2.cfm – accessed January 2015, 2014. Development of NFIQ 2.0.

[NFI12] NFIQ 2.0 Team. Development of NFIQ 2.0 – Quality Feature Definitions – Version 0.5. http://biometrics.nist.gov/cs_links/quality/NFIQ_2/NFIQ-2_Quality_Feature_Defin-Ver05.pdf – accessed January 2015, 2012.

[R C14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[UHBAR12] Bradford T. Ulery, R. Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. *PLoS ONE*, 7(3), March 2012.

[UHBR11] Bradford T. Ulery, R. Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, 108(19):7733–7738, 2011.

[UHBR12] Bradford T. Ulery, R. Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. *PLoS ONE*, 7(3):e32800, 03 2012.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[Wat93] Craig I. Watson. Nist special database 14. Technical report, National Institute of Standards and Technology, 1993.

[Wat01] Craig I. Watson. Nist special database 29. Technical report, National Institute of Standards and Technology, 2001.

[WGW04] C. L. Wilson, M. D. Garris, and C. I Watson. Matching Performance for the US-VISIT IDENT System Using Flat Fingerprints NISTIR 7110. Technical report, NIST, May 2004.

[YCLJ13] Soweon Yoon, Kai Cao, Eryun Liu, and AK. Jain. LFIQ: Latent fingerprint image quality. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013.

[ZK99] Daniel Zwillinger and Stephen Kokoska. *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press, 1 edition, 12 1999.