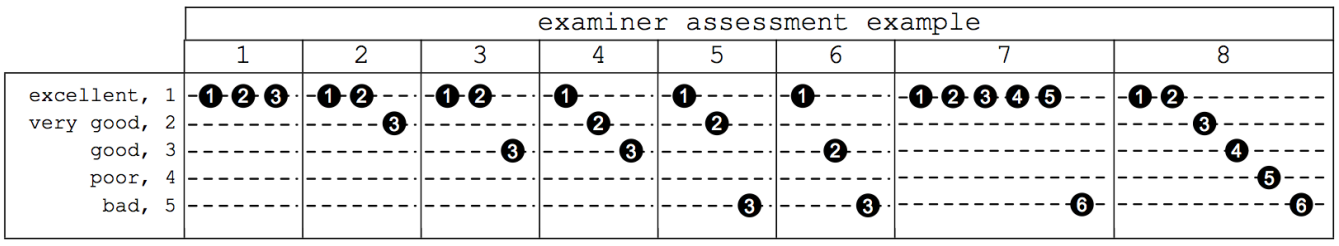
Automated fingerprint quality assessment as a replacement for dactyloscopic expert assessment

Martin Böckeler

Abstract
One of the most explored, permanent, unique and widely accepted biometric modalities is the fingerprint. In the forensic sector fingerprints are of special interest as they are often left unintentionally as latents on objects at crime scenes.
To use reconstructed latents in court cases for identification, they have to go through an analysis, comparison, evaluation, and verification (ACE-V) process that heavily relies on subjective dacyloscopic expert opinion which is itself influenced by human factors like health problems, stress or inadequate training.
Various studies had shown that the fingerprint quality computed by an algorithm highly correlates with its comparison score. Nonetheless, no known study has compared the correlation between automatic and dactyloscopic expert fingerprint quality assessment. This thesis addresses the lack, explores correlations between automatic and expert fingerprint quality assessment and further investigates if dactyloscopic expert assessments can be predicted.



metric							
$\tilde{P}$	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

Figure 1: Visualization of 8 examiner assessment examples with corresponding metric results. Each black dot represents an examiner assessment, ranging from excellent (1) to bad (5). The second table summarizes the results of different statistical metrics applied to the examples starting with the Percentage metric (P), the interquartile range (IQR), the median absolute deviation (MAD), the standard deviation (SD) and the proposed Closest-neighbor Median Cluster Agreement (CMCA).

Motivation
Even when courts accepted fingerprints as evidence over a 100 years ago and despite the fact that human quality assessment is very expensive and time consuming, it is also not objective.
Studies have shown that inter (whether multiple examiners reach the same decision on the same fingerprint) and intra (whether one examiner consistently reaches the same decision on the same fingerprint over a period of time) examiner quality assessment is inconsistent.
Studies had shown that the fingerprint quality computed by an algorithm highly correlates with its comparison score but no known study has compared the correlation between automatic and dactyloscopic expert fingerprint quality assessment which is able to overcome the problems of inter and intra examiner disagreement.

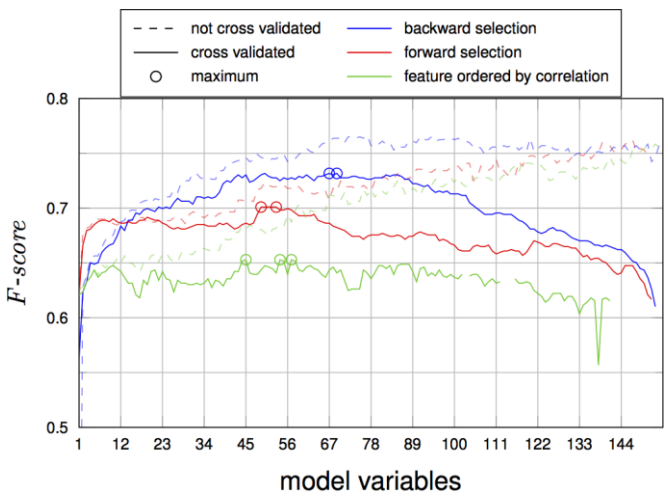
Goal
The ground truth dataset used in this work consists of a number of fingerprint images that have been assessed by dactyloscopic experts where at least one quality value was assigned based on the experts opinion.
The goal of this thesis is to explore and identify features present in fingerprint images which can be used as predictors for sample quality and relate them to the assessments of dactyloscopic experts. Therefore individual features of the NFIQ 2.0 framework shall be investigated and combined in a way to predict expert assessment.

Expert consensus
To explore the relationship between automatic and human fingerprint quality assessment it is essential to quantify expert consensus to determine if experts agree on what quality means. The higher the consensus of human fingerprint quality assessment, the easier it will be to judge if automatic quality assessment is able to produce the same assessment results as its human counterpart.
Therefore several existing metrics where investigated to determine if they are capable to sufficiently measure examiner agreement. As no investigated metric satisfactorily fulfills this task, a new metric, called Closest-neighbor Median Cluster Agreement (CMCA) is proposed in this thesis.

Ground truth analysis
In February 2009 an expert crew from the BKA annotated ground truth data from subsets of the NIST SD14 and SD29 datasets for the purpose of semantic conformance testing.
The experts annotated several fingerprint characteristics such as the fingerprint type, the overall fingerprint quality and the completeness of the whole print, ranging from 1 (excellent), 2 (very good), 3 (good), 4 (fair) up to 5 (poor).
The range of experts that made an assignment to the same fingerprint goes from a minimum of 2 experts up to a maximum of 9 experts per print and forms thereby several assessment subsets which consensus are measured using the CMCA metric.
High consensus between dactyloscopic experts of every subset was measured, resulting in the decision to take the median expert quality assessment per fingerprint as the dependent / outcome variable for the expert assessment prediction model.

Automatic expert quality assessment prediction
To predict the median expert quality assessment, 155 automatic calculated feature values from the NFIQ 2.0 framework serve as independent / predictor variables.
Since the quality categories with which an expert can rate a fingerprint are ordinal dependent, the ordered logistic regression was chosen as statistical prediction model.
One method to select (choose a subset of all available predictor variables) a well performing prediction model is to build all possible models and then choose the one that has the best measure of goodness of fit to the observed data.
This procedure however is very cost intensive as the number of possible models that must be investigated exponentially increases by $2^n - 1$ where n is the number of possible model variables.
To manage the 155 NFIQ 2.0 features and all their combinations, 3 algorithms are compared against each other.
The first one is driven by the assumption that a feature with high correlation to the median expert assessment will perform better in predicting the expert than a feature with low correlation.
Accordingly the first algorithm selects the NFIQ 2.0 feature with the highest correlation to the median expert assignment for a model with 1 variable. A model with 2 variables contains the feature with the highest correlation and the feature with the second highest correlation and so on.
The second algorithm is a kind of backward greedy search that performs a backward elimination of variables without stopping rule. The algorithm starts with a full model containing all n available variables. At each iteration, all possible models with n−1 variables will be evaluated and the model with the highest F-score will be taken to the next iteration. The algorithm operates until a model with 1 variable is reached.
The third algorithm is a kind of greedy forward search that performs a forward selection of variables without stopping rule. It starts with an empty model with n=0 variables. At each iteration, all possible models with n+1 variables will be evaluated and the model with the highest F-score will be taken to the next iteration. The algorithm stops if non of the n+1 models converged or the full model with all variables is reached.

Evaluation
To reliable measure the examiner assessment model prediction performance, the F-score with β=1 as the harmonic mean between precision and recall was measured in a 10 fold cross validation test at each iteration of the 3 model selection algorithms. The best result was achieved with the backward model selection algorithm at 67 model variables and an F-score of 0.7316.



h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Kontakt
Martin Böckeler
E-Mail: martin.boeckeler@googlemail.com

Projektpartner
CASED - Center for Advanced Security Research Darmstadt

Referent: Prof. Dr. Christoph Busch
Korreferent: Martin Aastup Olsen