生物特征质量评估

顾姗

2020/8/6

生物特征质量估计

- 生物特征识别的性能取决于生物特征样本（例如指纹、虹膜和人脸）的质量
- 生物特征样本质量：a measure of a sample’s utility to automatic matching
- 生物特征质量测量应作为识别性能的指标，理想特性是与错误率相关（如FNMR、FMR或识别漏检率等）

- 人脸：Best-Rowden (TIFS18), FaceQnet (ICB19), SER-FIQ (CVPR20)
- 指纹：Cao et al. (ICB16), Crowd-based Learning (TIFS18)

人脸质量评估

- 应用：
  - 拒绝低质量人脸用于识别
  - 对多张人脸图像（序列或视频）根据图像质量进行融合
  - 动态识别算法：低质量图像可用速度较慢的但更鲁棒的识别算法
- 影响人脸质量的因素：
  - 图像获取条件：光照、位置、背景等
  - 人脸本身特性：姿态、遮挡、表情
- 相关人脸质量评估算法：
  - GT定义：
    - 专家标注：可能不能和实际识别算法系统性能完全相关
    - 人脸识别算法性能：利用匹配分数定义质量，但匹配分数和一对图像的质量相关
  - 特征提取：手工定义特征；基于深度学习的特征
  - 输出：定性类别标签；数值分数

Best-Rowden

- Automatic face image quality of unconstrained face images using the Labeled Faces
  - 通过crowdsourcing为无约束人脸图像数据库LFW进行人工标记
  - 研究人工标记人脸质量对自动人脸识别性能的效用
  - 利用CNN特征训练一自动质量评估的模型
  - 比较(1)人工标记质量与(2)利用相似度分数定义质量作为GT的性能

L. Best-Rowden and A. K. Jain. Learning face image quality from human assessments. IEEE Transactions on Information Forensics and Security, 13(12):3064–3077, Dec 2018. 2, 4,5, 6

Best-Rowden

Ground Truth

- Human Quality Values (HQV)
  - 标注一对图像的相对质量，194人*1001对人脸图像
  - 利用矩阵补全算法从稀疏的成对分数矩阵推断每个图像的单独质量分数->194*13233
  - 使用中位数作为最后的质量->1*13233
- Matcher Quality Values (MQV)
  - 匹配分数归一化： $z_{ij} = (s_{ij}^{G} - \mu_{ij}^{I})/\sigma_{ij}^{I},$
  - 基于假设：相似度分数主要受低质量图像的影响
  - 手工挑选1680张高质量图像作为库图像，剩余7484张图像作为probe图像；利用匹配分数定义probe图像的质量

Best-Rowden

Method

- 特征：
  - 人脸图像质量应能反映人脸识别性能-> 直接使用训练好的人脸识别算法提取特征
  - 人脸检测 -> 人脸校正 ->提取特征 Deep-320
- 模型：
  - 支持向量回归（SVR）预测质量分数

Best-Rowden

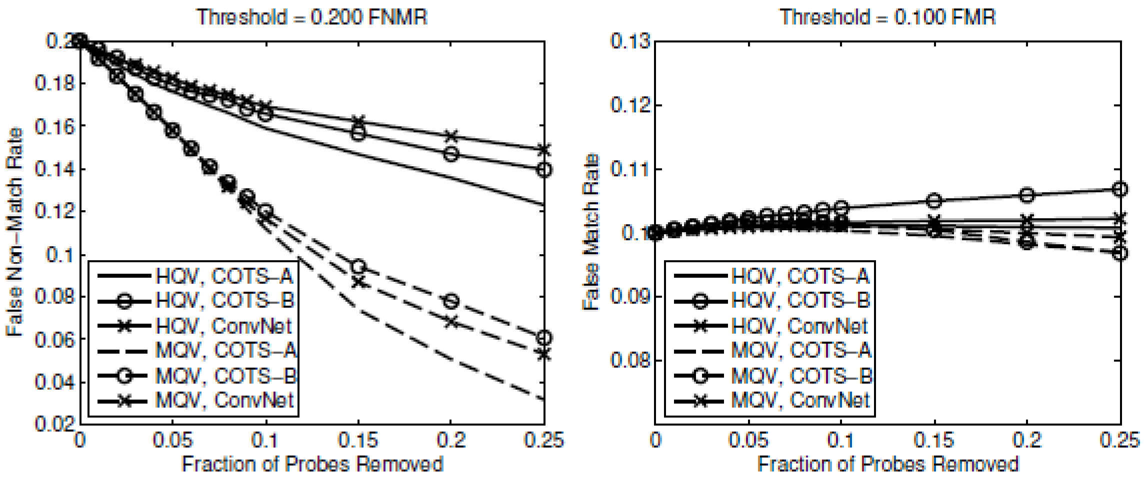
Experiments

- 评价指标：
  - Error versus Reject (EvR) curve
  - 固定（FMR,FNMR）时错误率（FNMR， FMR）versus 拒绝低质量样本的比例
- 数据集：
  - CASIA-WebFace (10575 subjects, 494414 images) (training)
  - Labeled Faces in the Wild (LFW) (5749 subjects, 13233 images) (training&testing)
  - IARPA Janus Benchmark-A (IJB-A) (500 subjects, 5712 images&2085 videos)
- Matcher: COTS-A, COTS-B, ConvNet
  - 获得MQV的GT
  - 评价质量预测算法对人脸识别算法的效用
  - ConvNet 还用于提取人脸deep 特征

Best-Rowden

Experiments

- GT 质量对识别性能作用



(a) FNMR (b) FMR

A 移除低质量图像可以降低所有匹配器的FNMR，表明人工标注质量与识别性能相关，但MQV在降低FNMR方面更有效。
B 移除低质量图像对FMR影响不大。

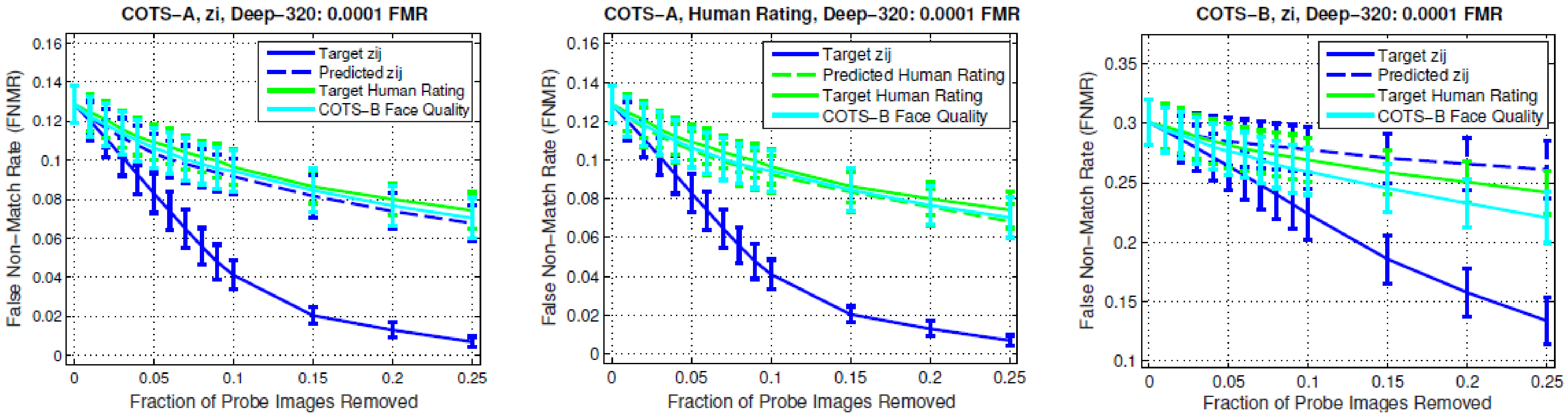
Best-Rowden

Experiments

- 人脸质量预测

TABLE III
SPEARMAN RANK CORRELATION (MEAN ± STANDARD DEVIATION OVER 10 RANDOM SPLITS OF LFW IMAGES) BETWEEN TARGET AND PREDICTED MQV AND HQV

	Matcher		
	COTS-A	COTS-B	ConvNet
MQV	0.558 ± 0.023	0.442 ± 0.026	0.459 ± 0.022
HQV	0.585 ± 0.019		



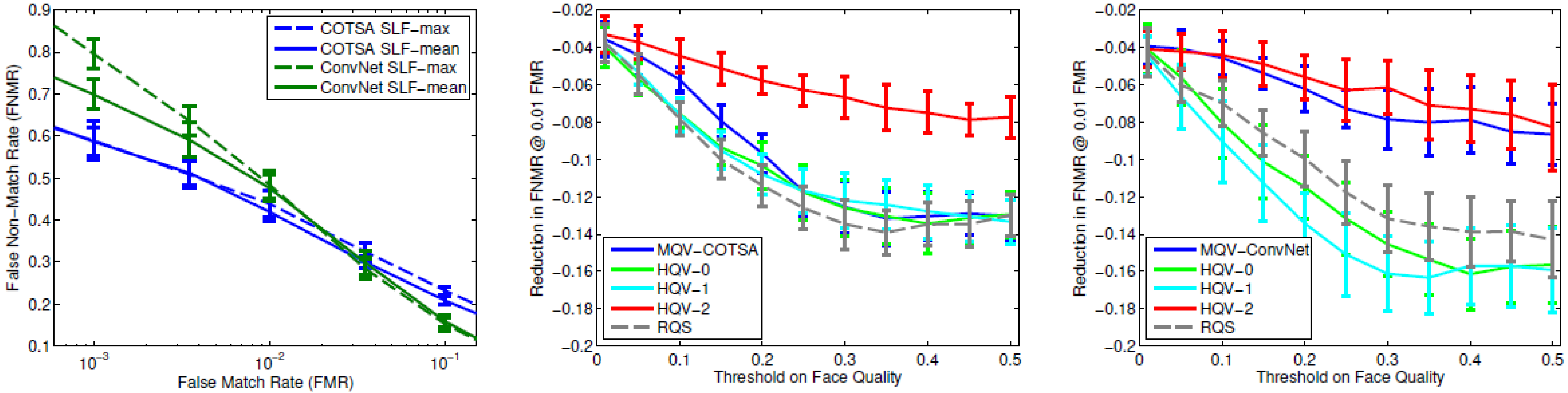
(a) COTS-A, z_{ij} (b) COTS-A, Human (c) COTS-B, z_{ij}

Fig. 9. Error vs. Reject curves for target and predicted face image quality values (MQV and HQV) for the LFW database. The curves show the efficiency of rejecting low quality face images in reducing FNMR at a fixed FMR of 0.01%. The models used for the face quality predictions in (a)-(c) are SVR on the deep-320 features from ConvNet in [17].

Best-Rowden

Experiments

- 人脸质量预测
  - 给定人脸质量阈值，每个对象template只包含质量超过阈值的图像
  - 固定1%FMR时，FNMR随图像质量变化时的下降程度



(a) COTS-A and ConvNet (b) COTS-A (c) ConvNet

Fig. 13. Results for the verification protocol of the IJB-A database [6]. All curves in (a)-(c) show mean performance and error bars give standard deviation in performance over the 10 splits in the protocol. (a) Receiver Operating Characteristic (ROC) for COTS-A and ConvNet [17] matchers, where score-level fusion (SLF) is applied to the multiple face samples per subject for template-based matching of the IJB-A protocol. Using thresholds on face image quality measures to determine which face samples in a template to use for matching, (b) and (c) plot reduction in FNMR at 1% FMR, showing that FNMR decreases as the face quality thresholds are increased. Flowcharts providing details of each face quality method (MQV, HQV-0, etc.) are given in Fig. 12. The RQS method is proposed by Chen *et al.* [35].

FaceQnet

- Best-Rowden的方法：
  - HQV需要大量人工标注
  - MQV需要人工挑选高质量图像
  - 耗时耗力且会引入人工偏差
- ->
  - 使用第三方软件获得ICAO分数，自动选取高质量图像
  - GT只使用匹配分数建立与质量的联系
  - Finetune与训练好的人脸识别网络进行质量估计

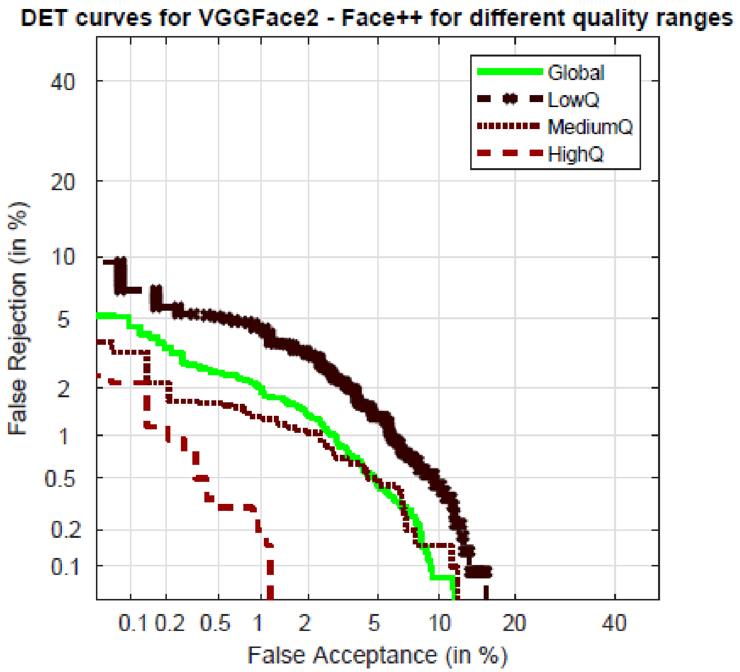
Javier Hernandez-Ortega, Javier Galbally, Julian Fi´errez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In IEEE International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019, Jun. 2019. 2, 3, 5, 7

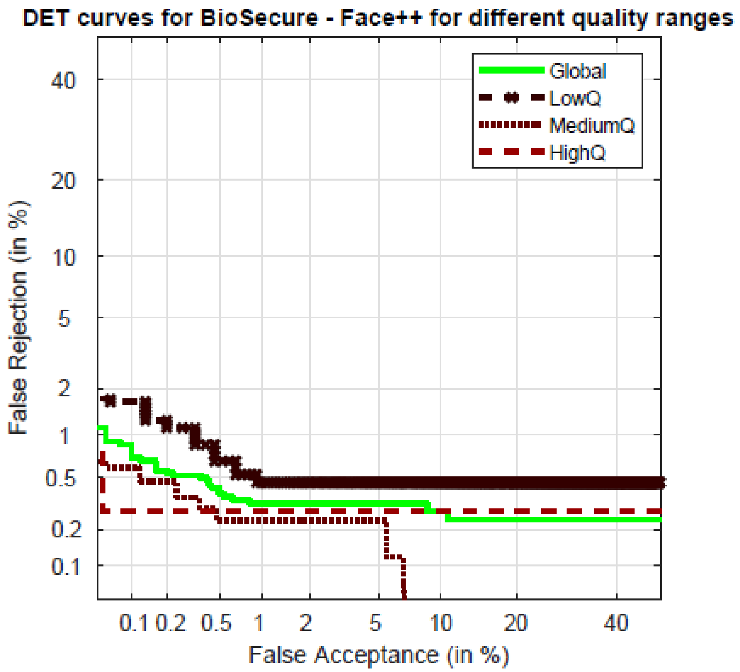
FaceQnet

- GroundTruth
  - VGGFace2中300subject，使用BioLab框架获得ICAO分数，每个subject得到一张高质量的gallery图像
  - 使用预训练的FaceNet计算剩余图像和gallery图像相似度，作为质量的GT
- Regression Model
  - 质量与准确率息息相关
  - 选择预训练的ResNet-50模型，将最后的分类层替换为两个全连接层
  - 固定前序网络参数，只训练全连接层参数

FaceQnet

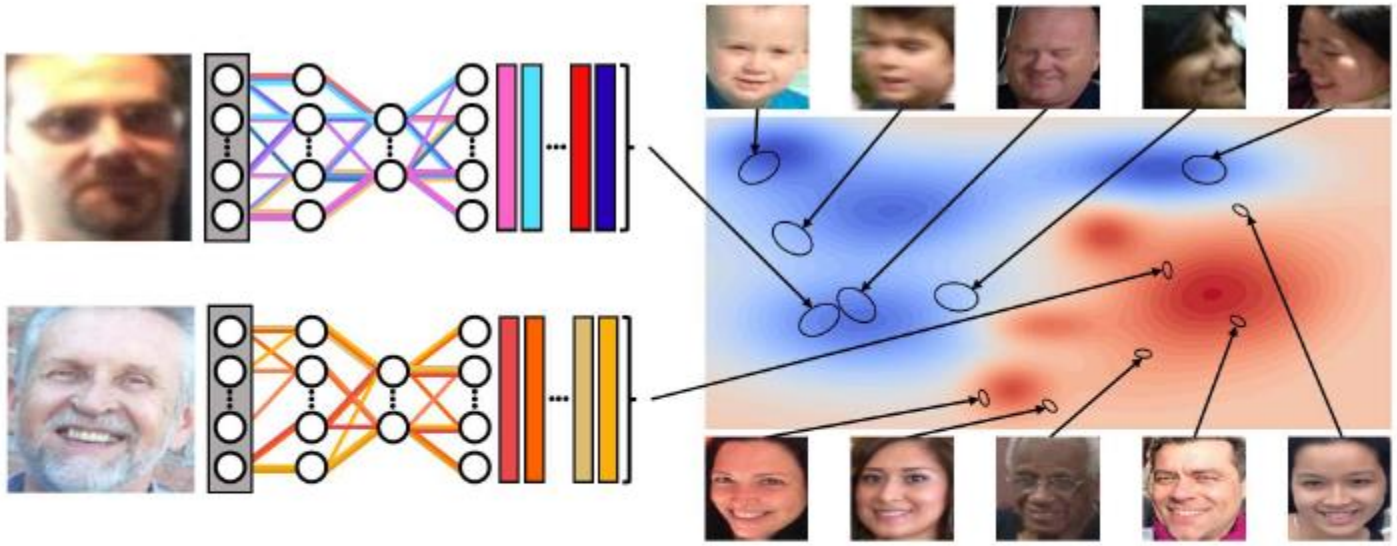
- Experments
- 数据集：
  - VGGFace2 (300 subjects for finetune FaceQnet , 100 subjects for test)
  - BioSecure Multimodal Database (BMDBA) (140 subjects, 1459 images selected)
- 不同质量图像的DET曲线





SER-FIQ

- 目前利用监督的算法，人脸质量的GT不能真实反应人脸识别算法性能
- -> 提出一种无监督的人脸质量评估算法
- 使用网络模型中随机子网络提取特征，计算特征之间的偏差



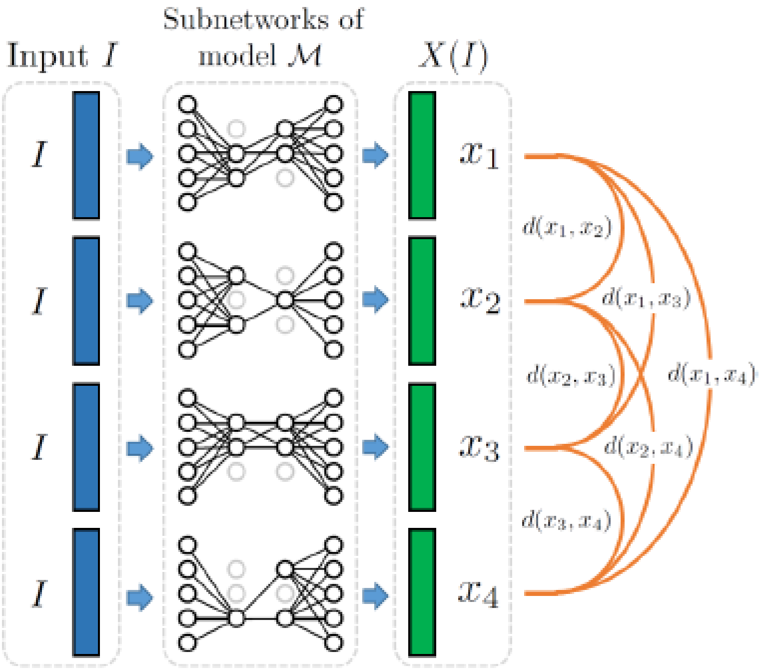
Terhorst, Philipp, et al. "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

SER-FIQ

Method

- 给定一包含dropout的人脸识别model，使用dropout提取m=100次特征，人脸质量用特征的鲁棒性定义：

$$q(X(I)) = 2\,\sigma\Big(-\frac{2}{m^2}\sum_{i<j} d(x_i, x_j)\Big),$$



SER-FIQ

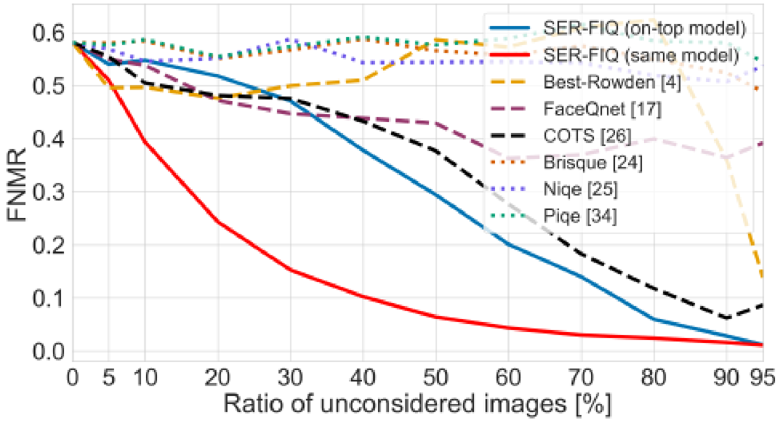
Experiment

- 数据集：
  - ColorFeret (1199 subjects,14126 high-resolution images) (training)
  - Adience (2284 subjects, 26580 images)
  - LFW (5749 subjects, 13233 images)
- 评价指标：
  - Error versus Reject (EvR) curve
- 人脸识别网络： FaceNet / ArcFace
  - On top model：适用于任何人脸识别网络， $n_{emb}/128/512/n_{emb}/n_{ids}$
  - Same model ： 需要网络本身包含dropout

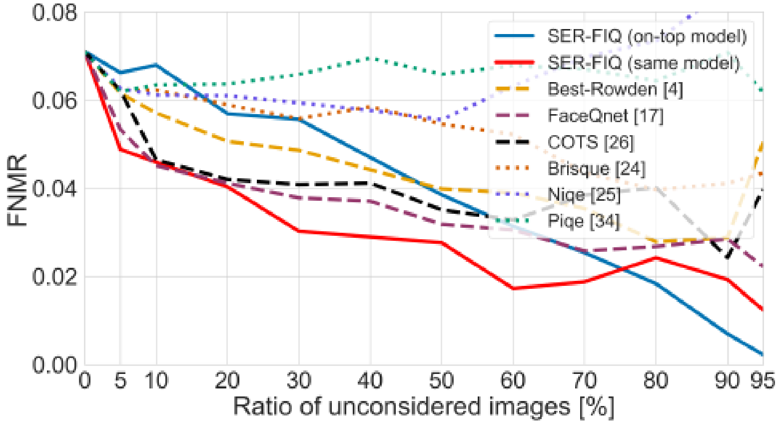
SER-FIQ

Experiment

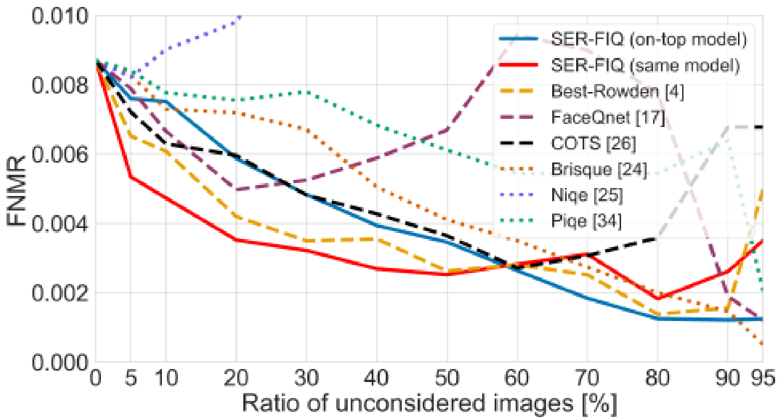
- 0.001 FMR



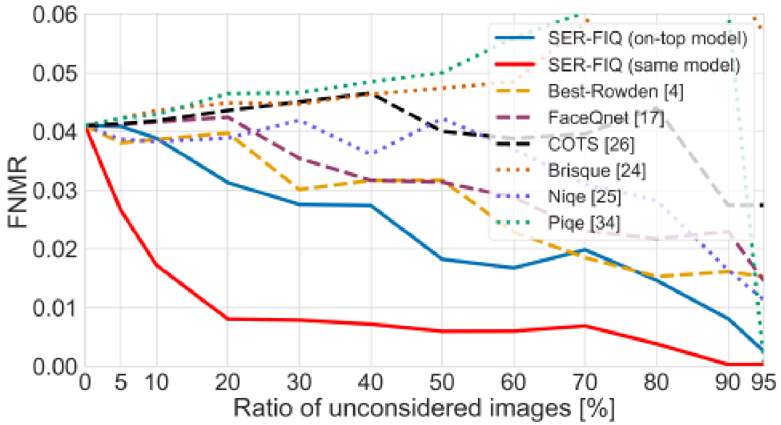
(a) Adience - FaceNet



(b) Adience - ArcFace



(c) LFW - FaceNet



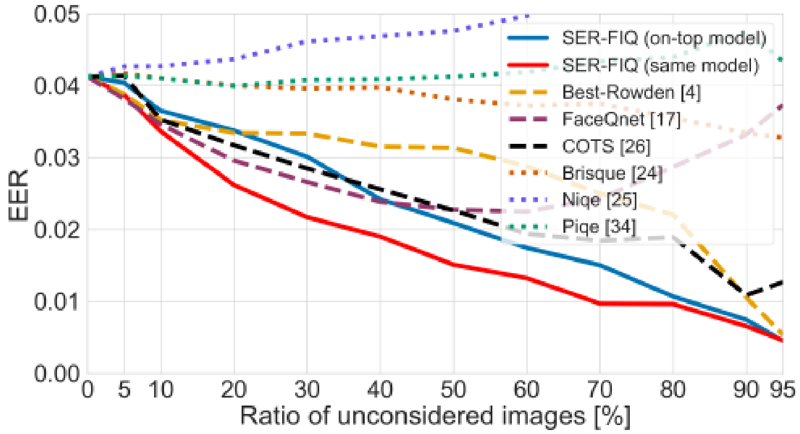
(d) LFW - ArcFace

A 仅预测图像本身质量对人脸识别性能提升有限
B U型曲线：高质量图像区分度较差
C 针对特定人脸识别模型预测质量能取得最好的效果

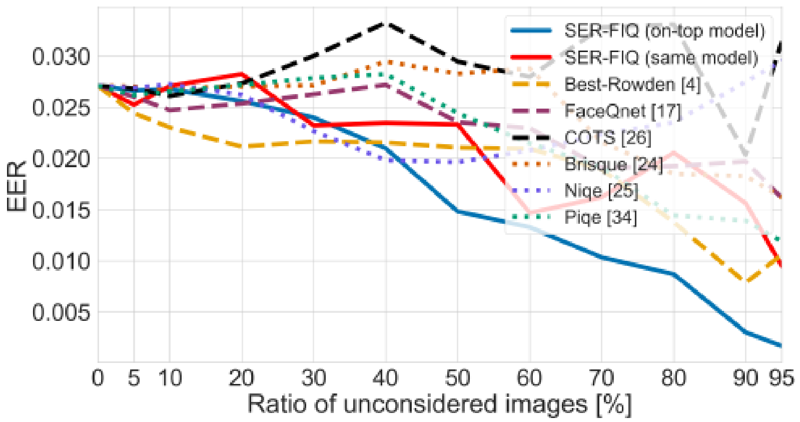
SER-FIQ

Experiment

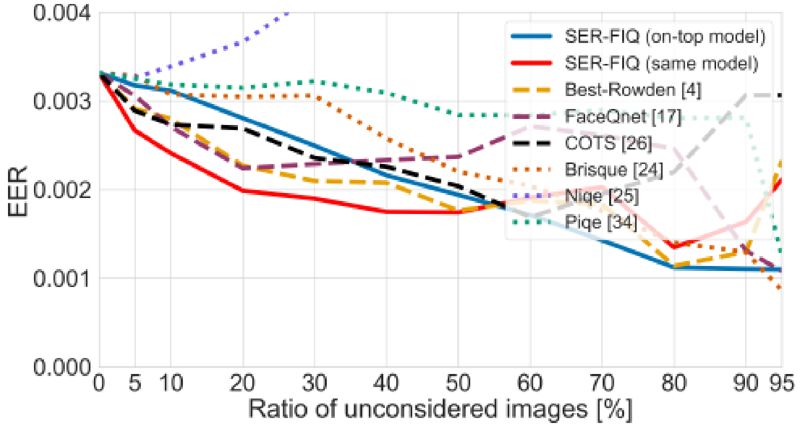
- EER



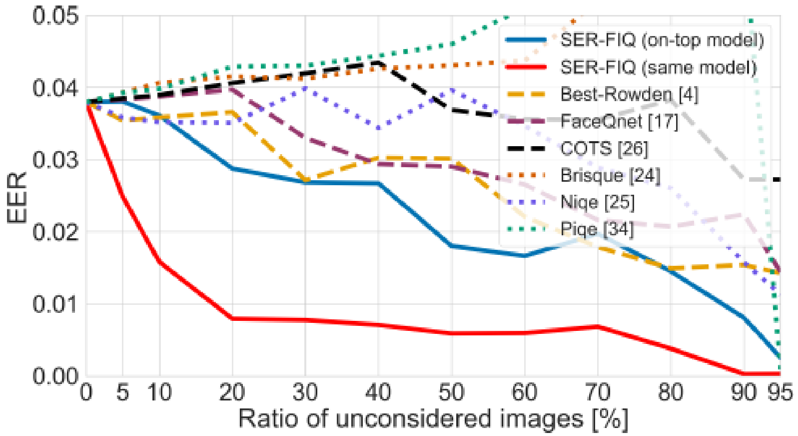
(a) Adience - FaceNet



(b) Adience - ArcFace



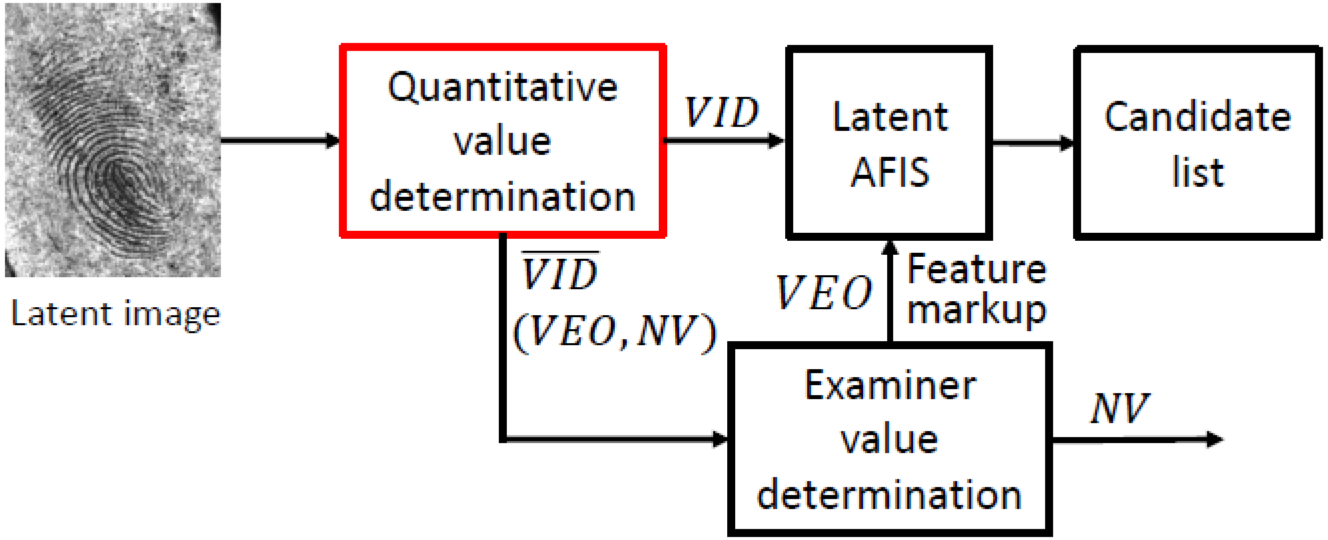
(c) LFW - FaceNet



(d) LFW - ArcFace

现场指纹质量估计

- Analysis, Comparison, Evaluation and Verification (ACE-V) methodology
- 手工标注现场指纹费时费力；高质量现场指纹能用全自动的方式处理；
- 人工定义现场指纹质量，客观性和一致性不能保证。Repeatability（intra）为84.6%，Reproducibility（inter）仅为75.2%。
- 人工定义的现场指纹质量和AFIS性能不一定一致

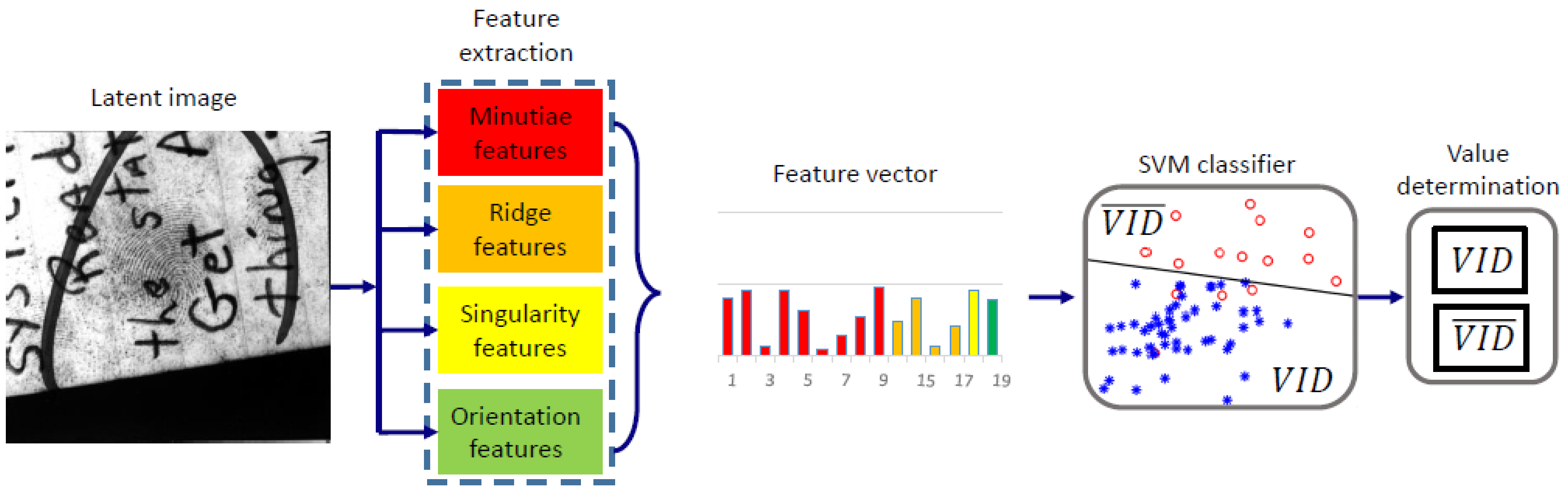


基本概念

- 在ACE-V的分析中，现场指纹通常被人为定义为3类:
  - value for individualization (VID)
  - value for exclusion only (VEO)
  - no value (NV)
- VID VEO被认为可用于后续评价，NV被认为应被抛弃

Automatic Latent Value Determination

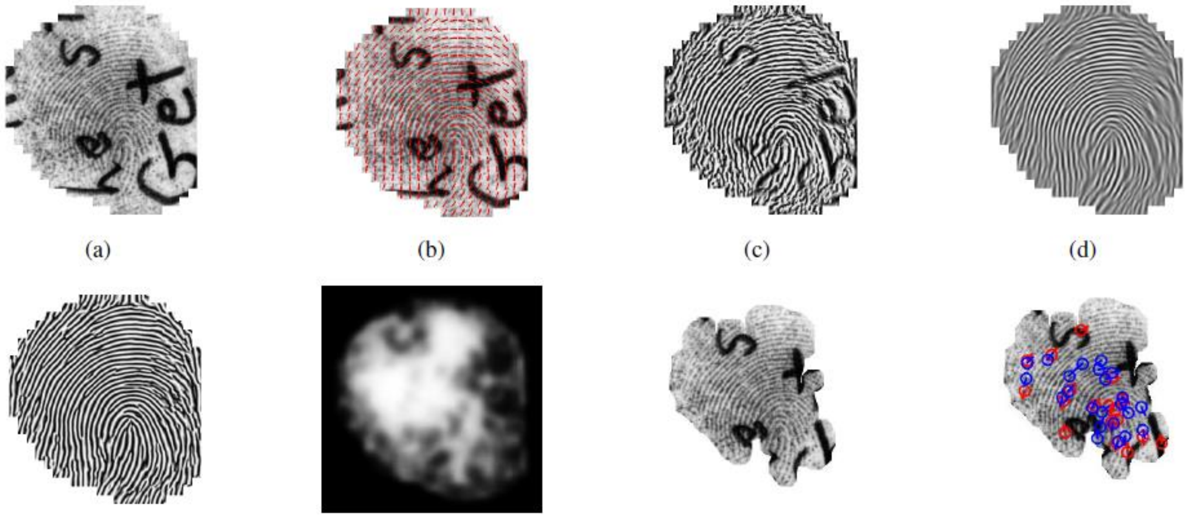
- 全自动的现场指纹评估算法（手工ROI）
  - 19维特征+ 2分类（VID 与~VID ） SVM/RF
  - GT： 手工标注/ latent AFIS算法rank-1



K. Cao, T. Chugh, J. Zhou, E. Tabassi, and A. K. Jain, “Automatic Latent Value Determination,” in 9th IAPR International Conference on Biometrics (ICB), 2016, pp. 1–6.

Automatic Latent Value Determination

- 脊线质量估计： 排除错误的细节点
  - R_c： $I_N(c)$与$E_N(e)$ patch之间的相似度
  - $R_f = \{Coh(I_N) + \mathrm{Coh}\{E\}\}$:
    $I_N(c)$与$E\ (d)$ 的orientation coherence map： $Coh(I_N)$与$\mathrm{Coh}(E)$
  - $R = R_c + R_f - T$



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5: Illustration of the main steps in automatic minutiae extraction. (a) Input latent (I) with ROI, (b) estimated ridge flow, (c) normalization of I (I_N), (d) enhancement of I_N (E), (e) normalization of E (E_N), (f) estimated quality map, (g) cropped latent image based on ridge quality map, and (h) detected minutiae (red) and manually annotated minutiae (blue). In (g) and (h), the automatical cropping explains the irregular boundary.

Automatic Latent Value Determination

- 细节点可靠性：
  - R_M 细节点M处的脊线质量
  - f 细节点紧凑性：
  - 手指位置w： 关于细节点坐标和奇异点坐标的距离的二维高斯分布
  - $Q_M = R_M\,(1 - f)\,w$

$$f = \begin{cases} 0 & \text{if } d > d_1 \\ 1 & \text{if } d < d_2 \\ \frac{(d_1 - d)}{d_1 - d_2}, & \text{otherwise,} \end{cases}$$

Table 2: Feature vector used for quantitative value assessment.

Feature no.	Description
1	Number of minutiae extracted in the latent
2-8	Sum of reliability of minutiae that have reliability value $\geq t$, $t = 0, 0.1, ..., 0.6$
9	Average area of the triangles in minutiae Delaunay triangulation
10	Area of the convex hull of minutiae set
11-17	Sum of ridge quality of blocks that have quality value $\geq t$, $t = 0, 0.1, ..., 0.6$
18	Number of singular points (core and delta) [16]
19	Standard deviation of the ridge flow (orientation map) in the foreground

Automatic Latent Value Determination

- 实验：
  - 数据集 NIST27 & WVU，10重交叉验证，训练+测试
  - Reference database: 100,000 (707 mated)

Table 3: Latent database description. Values determined by examiners were reported in [12] and values determined by AFIS (in parentheses) are obtained using the hit rate of a latent AFIS[3].

	NIST SD27	WVU
No. of latents	258	449
Capture environment	operational casework	laboratory environment
No. of VID latents	210 (176)	370 (317)
No. of $\overline{\text{VID}}$ latents	48 (82)	79 (132)

数据集

Table 6: Confusion matrix illustrating the differences in the value determination by examiners and AFIS on latents in NIST SD 27 and WVU databases (total of 707 latents).

Method	VID by AFIS	$\overline{\text{VID}}$ by AFIS
VID by examiners	458	122
$\overline{\text{VID}}$ by examiners	35	92
Consistency (%)	77.8%	

两种GT的一致性

Automatic Latent Value Determination

- 实验：

Table 4: Confusion matrix, classification accuracy (%) and AUC of the proposed method and LFIQ when the ground truth is provided by examiners. Numbers in the brackets indicate s.d. based on 10-fold cross validation. Each fold has ~70 latents.

Method	Proposed*		LFIQ [23]**	
	$\hat{y}_{\mathrm{VID}}$	$\hat{y}_{\overline{\mathrm{VID}}}$	$\hat{y}_{\mathrm{VID}}$	$\hat{y}_{\overline{\mathrm{VID}}}$
y_{VID}	534	46	546	34
$y_{\overline{\mathrm{VID}}}$	56	71	51	76
Classification accuracy	85.6% (2.4%)		88.0% (3.8%)	
AUC	0.892 (0.033)		0.903 (0.045)	

* Automatically extracted features.
** Manually annotated minutiae.

Table 5: Confusion matrix, classification accuracy (%) and AUC of the proposed proposed and LFIQ when the ground truth is provided by AFIS (Rank-1 retrieval).

Method	Proposed*		LFIQ [23]**	
	$\hat{y}_{\mathrm{VID}}$	$\hat{y}_{\overline{\mathrm{VID}}}$	$\hat{y}_{\mathrm{VID}}$	$\hat{y}_{\overline{\mathrm{VID}}}$
y_{VID}	447	46	433	60
$y_{\overline{\mathrm{VID}}}$	99	115	103	111
Classification accuracy	79.5% (7.2%)		76.9% (4.1%)	
AUC	0.824 (0.069)		0.835 (0.059)	

* Automatically extracted features.
** Manually annotated minutiae.

Crowd-based Learning

- Crowdsourcing based 理解专家标记指纹质量时的潜在基础；并用于学习一定量的指纹质量预测器
  - Input：单个指纹的类别；成对指纹的相对类别
  - 利用Multidimensional Scaling (MDS)学习专家标记时的潜在基础；使用Lasso建立特征与潜在基础间的关系

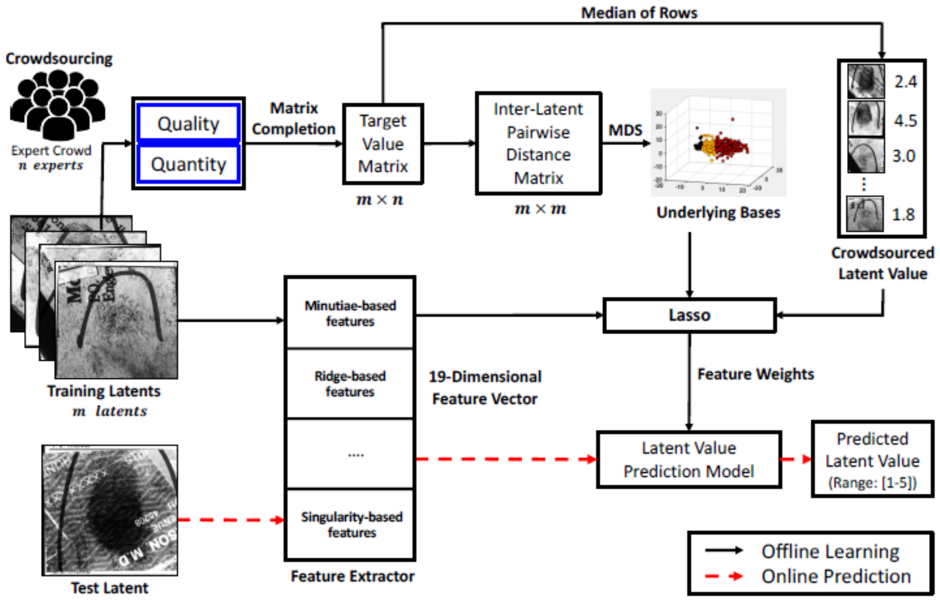
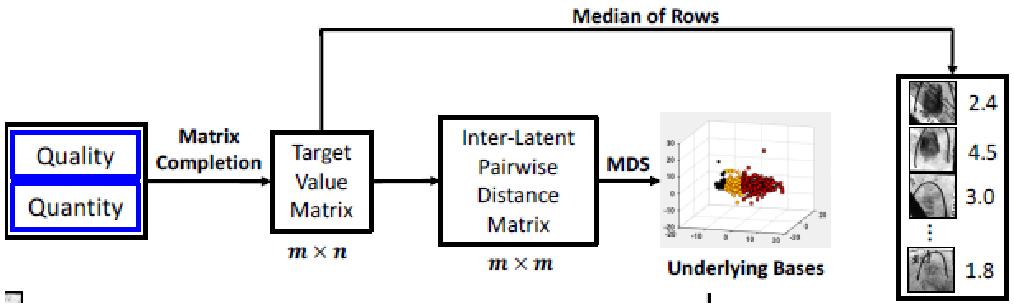


Fig. 5: Overview of the proposed crowdsourcing-based learning approach for latent value assignment.

T. Chugh, K. Cao, J. Zhou, E. Tabassi, A.K. Jain, Latent fingerprint value prediction: crowd-based learning, IEEE Trans. Inform. Forensics Secur. 13 (1) (2018) 20–34, doi:http://dx.doi.org/10.1109/TIFS.2017.2721099

Crowd-based Learning

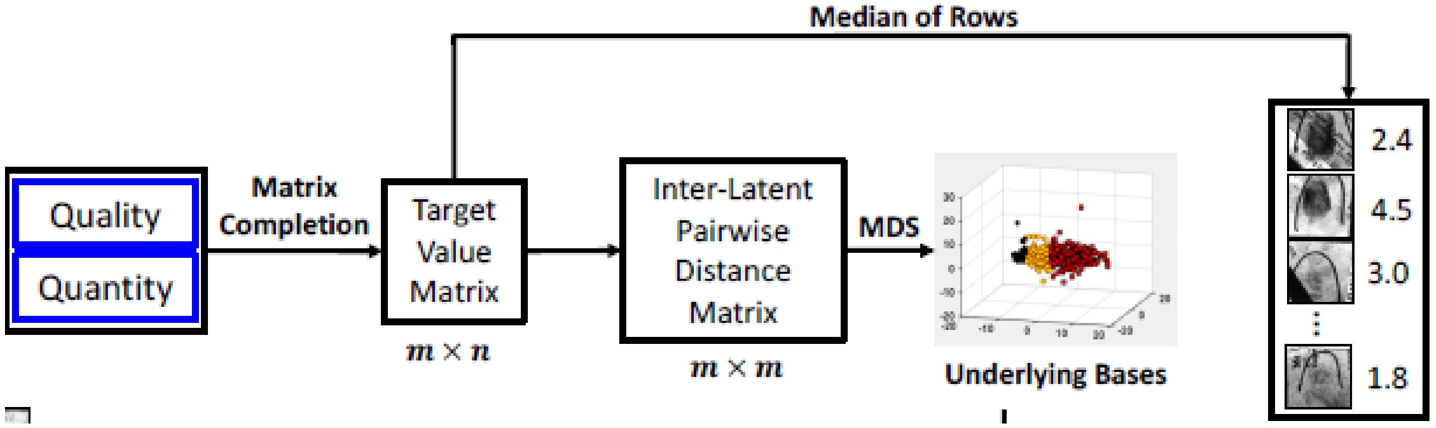
- Crowdsourcing
  - 516个现场指纹，随机选择100对，标注每个指纹的单独quality质量&相对quantity质量
  - 31个专家
  - -> 6200单独类别标签，3100相对标签
- Matrix Completion
- Latent Value
  - 寻找对quality和quantity线性组合的最佳权重 $\hat{V} = \frac{\hat{Q} + \hat{C}}{2}$

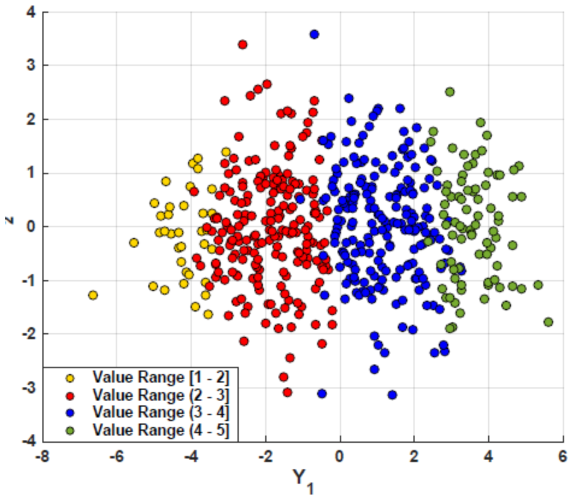


Crowd-based Learning

- Bases for Explaining Expert Ratings
  - MDS算法：使每对样本在高维空间的距离与在构建的低维空间中的样本相似性尽可能保持一致
- Interpreting Bases in terms of Latent Features
  - 每个指纹提取19维特征向量
  - 利用Lasso建立特征向量与MDS embedding之间的关系

$$y_k = \beta_{k1} \cdot x_1 + \beta_{k2} \cdot x_2 + \ldots + \beta_{k19} \cdot x_{19} + \beta_{k20}$$





Crowd-based Learning

- Learning Latent Value Predictor
  - 利用Lasso建立MDS base和专家标记的质量分数之间的关系
  - -> 质量分数与特征之间的关系

$$
\begin{aligned}
V_{pred} = 1.22\ x_1 + 0.50\ x_2 + \ 0.55\ x_3 + 0.40\ x_4 + 0.29\ x_5 \\
+ 1.14\ x_9 + 0.43\ x_{10} + \ 0.29\ x_{11} + 0.34\ x_{12} \\
+ 0.92\ x_{18} + \ 1.19\ x_{19} + 1
\end{aligned}
$$

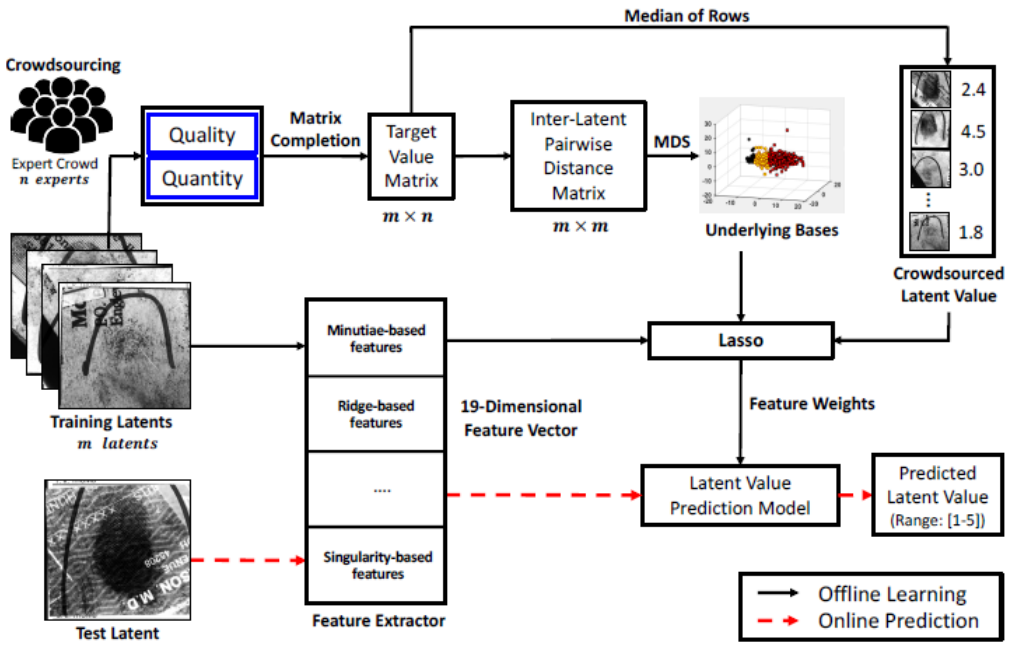


Fig. 5: Overview of the proposed crowdsourcing-based learning approach for latent value assignment.

Crowd-based Learning

实验：

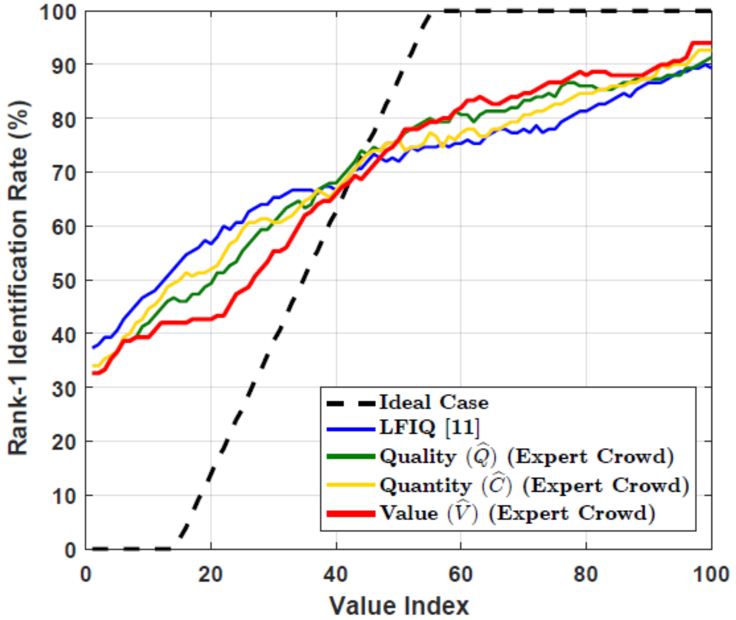
- Evaluating Target Latent Value from Expert Crowd
- NIST27/MSP258
  - 516个指纹按分数排序，分为100个可重叠的bin，每个bin中包含100个现场指纹；
  - 在每个bin中，比较利用state of the art AFIS识别成rank-1的数量
- NIST27 （210 VID, 41 VEO, 7 NV）, 166/258 retrieved at rank1
  - 

TABLE VI: Number of latents retrieved at Rank-1, using a state-of-the-art latent AFIS, for NIST SD27 latents that were determined as VID (210), VEO (41) and NV (7) by examiners [8]. For value determination by the expert crowd, the median inferred latent value ($\widehat{v}$) is thresholded as VID > 2.16 and NV < 1.66. The range for the median latent value is $[1 - 5]$.

	VID	VEO	NV
Value Determination by Examiners [8]	155/210	11/41	0/7
Value Determination by Expert Crowd ($\widehat{V}$)	161/210	5/41	0/7

Crowd-based Learning

实验：

- Correlation between Predicted Latent Value and AFIS Performance
  - MSP 400, WVU 139, MOLF 2926, with high mates (NFIQ=1) 保证算法性能只受现场指纹质量影响
  - Background： 250,000 roll prints

TABLE VII: Correlation of predicted latent value by LFIQ, and by the proposed method with the performance of a state-of-the-art AFIS for three independent latent databases.

	LFIQ [11]	Proposed Method
MSP Latent Database (400 latents)	0.49	0.70
WVU Latent Database Subset [46] (139 latents)	0.47	0.73
IIITD MOLF Latent Database Subset [47] (2926 latents)	0.42	0.67