计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100210

视频人脸识别进展综述

白子轶　毛懿荣　王瑞平
中国科学院计算技术研究所智能信息处理重点实验室　北京 100190
中国科学院大学计算机科学与技术学院　北京 100049
(ziyi.bai@vipl.ict.ac.cn)

摘　要　人脸识别是生物特征识别领域的一项关键技术,长期以来得到研究者的广泛关注。视频人脸识别任务特指从一段视频中提取出人脸的关键信息,从而完成身份识别。相较于基于图像的人脸识别任务来说,视频数据中的人脸变化模式更为多样且视频帧之间存在较大差异,如何从冗长而复杂的视频中抽取到人脸的关键特征成为当前的研究重点。以视频人脸识别技术为研究对象,首先介绍了该技术的研究价值和存在的挑战;接着对当前研究工作的发展脉络进行了系统的梳理,依据建模方式将传统基于图像集合建模的方法分为线性子空间建模、仿射子空间建模、非线性流形建模、统计建模四大类,同时对深度学习背景下基于图像融合的方法进行了介绍;另外对现有视频人脸识别数据集进行分类整理并简要介绍了常用的评价指标;最后分别采用灰度特征和深度特征在 YTC 数据集及 IJB-A 数据集上对代表性工作进行评测。实验结果表明:神经网络可以从大规模数据中提取到鲁棒的视频帧特征,从而带来识别性能的大幅提升,而有效的视频数据建模能够挖掘出人脸潜在的变化模式,从视频序列包含的大量样本中找到更具判别力的关键信息,排除噪声样本的干扰,因此基于视频的人脸识别具有广泛的通用性和实用价值。

关键词:视频人脸识别;图像集合建模;子空间学习;流形学习;深度学习
中图法分类号　TP391

Survey on Video-based Face Recognition

BAI Zi-yi,MAO Yi-rong and WANG Rui-ping
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China
School of Computer Science and Technology,University of Chinese Academy of Sciences,Beijing 100049,China

Abstract　Face recognition is a key technology in the field of biometrics,which has been widely concerned by researchers in the past decades. Video-based face recognition task refers specifically to extract the key information of human faces from a video to complete the personal identification. Compared with the image-based face recognition task,the changing patterns of faces in videos are much more diverse,and there are great differences among the whole video frames as well. Current research focuses on how to extract the key features of faces from lengthy videos. Firstly,this paper introduces the research value and challenges of video-based face recognition. Then,the developing venation of the current research work is explored. Based on the video modeling manners,traditional image set based methods are divided into four categories:linear subspace modeling,affine subspace modeling,nonlinear manifold modeling and statistical modeling. Besides,the methods based on image fusion under the background of deep learning are also introduced. This paper also briefly reviews existing datasets for video-based face recognition and the commonly used performance metrics. Finally,gray features and deep features are used to evaluate the representative works on YTC dataset and IJB-A dataset. Experimental results show that deep neural network can extract robust features of each frame after being trained with large-scale data,which greatly improves the performance of video-based face recognition. Moreover,the effective video modeling can help to identify the potential human face changing patterns. Therefore,more discriminative information can be found from the large number of samples contained in the video sequence,and the inference of noise samples can be eliminated,which suggests the advantages of video-based face recognition to be applied to a large range of practical application scenarios.

Keywords　Video-based face recognition,Image set modeling,Subspace learning,Manifold learning,Deep learning

到稿日期:2020-12-05　返修日期:2021-01-27
基金项目:国家自然科学基金(61922080,U19B2036,61772500)
This work was supported by the National Natural Science Foundation of China (61922080,U19B2036,61772500).
通信作者:王瑞平(wangruiping@ict.ac.cn)

1 引言

人脸识别是计算机视觉及模式识别领域长期研究的课题之一。传统的人脸识别主要通过分析两张或多张包含人脸的图像，并对图像进行脸部关键特征提取来完成身份识别。静态图像的采集过程中人脸往往是受控的，但在很多实际应用场景下，需要对非受限拍摄环境下采集的视频进行人脸识别，如公安部门通过分析监控视频进行嫌疑人排查、互联网公司对海量视频数据进行以人为核心的分析与管理。因此，视频人脸识别存在其自身特有的技术难点：1）人脸表观的变化更加剧烈。除了视频拍摄设备以及摆拍距离较远带来的光照条件不佳、低分辨率、抖动模糊等多变性外，还包括拍摄对象自身不受控带来的头部姿态多变性、面部表情多变性以及大面积遮挡的情况。2）需要对视频数据进行整体建模。视频数据可以看作由多帧图像构成的集合，进行视频人脸识别时，一一比对两段视频中的所有帧的效率是非常低的，因此往往需要对视频进行整体建模。然而视频数据长短不一，且图像帧之间存在较大的差异性，如何从冗长而复杂的视频中提取到人脸关键信息成为另一大挑战。与此同时，视频数据也具备其天然的优势：视频数据的多样性提供了更为丰富的人脸信息，研究者可以通过构造更加复杂的人脸表示来提升人脸识别的性能[1]；另外，视频数据同时具有空间和时间两个维度，可以提供静态图像不具备的人脸动态信息，从而为人脸识别带来更多可能性。

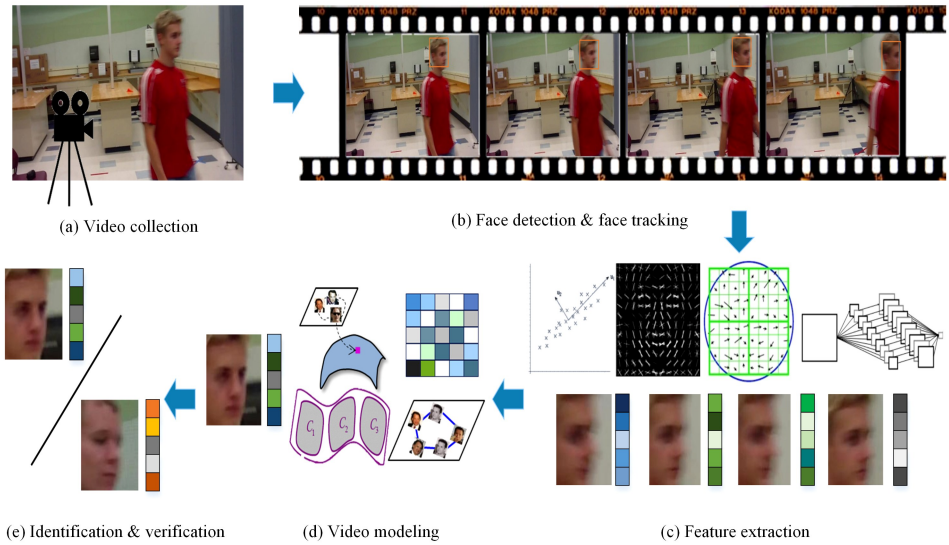


图 1 视频人脸识别流程

Fig. 1 Process of video face recognition

完整的视频人脸识别流程主要包括人脸视频数据的采集、人脸检测、人脸跟踪、人脸特征提取及视频建模等多个环节，其整体框架图如图 1 所示。在完成视频数据的收集之后，首先对视频进行人脸检测及跟踪[2]，并对人脸区域进行裁剪，得到一系列人脸区域图像；人脸特征提取步骤是指对视频中的每一帧图像采用 PCA 降维[3]、LBP[4]、SIFT[5]、HoG[6]等手工设计的浅层特征或深度卷积神经网络提取得到人脸关键信息的特征表达；而视频建模步骤则是利用所有图像特征进行建模得到统一的人脸特征表示，用于后续的判别分析；最后，识别任务可以分为 1∶1 的人脸验证（Verification）任务及 1∶N 的人脸辨识（Identification）任务。人脸验证任务主要面向刷脸进站、刷脸支付等身份核实场景；而人脸辨识主要面向法医鉴定、门禁系统等身份认证场景。系统中包含一组已知身份的人脸库（Gallery），在测试阶段会将新的人脸数据（Probe）与人脸库进行比对，以此来确认身份。如果测试数据身份在人脸库中出现过，则称为闭集人脸识别，否则称为开集人脸识别。

近年来，很多研究者采用不同的视频建模方式，充分利用视频数据的优势尝试应对视频人脸识别任务中的各项挑战。本文以视频人脸识别技术为研究对象，对该领域的整体发展脉络进行梳理，并对现有研究方法及常用数据集进行全面而系统的分类介绍，最后通过对比实验分析各方法的独特性，总结得出未来的研究方向。

2 基于图像集合建模的视频人脸识别

在深度学习出现之前，研究者主要考虑如何建模视频中潜在的人脸变化模式。一类方法保留了视频完整的时序关系，利用时序信息进行视频建模，另一类方法则将视频中的每一帧看作独立的图像，用无序的图像集合表示一段视频。本文将重点关注基于图像集合建模的相关研究工作，并依据集合的建模方式对这类方法进行归纳总结。

2.1 线性子空间建模方法

早期，研究者[7-11]利用视频帧对应特征向量可以张成线性子空间的想法对图像集合进行线性子空间建模。对于一个图像特征集合$\boldsymbol{X}_i=\{\boldsymbol{x}_{i,1},\boldsymbol{x}_{i,2},\cdots,\boldsymbol{x}_{i,n_i}\}\in\mathbb{R}^{d\times n_i}$（其中$n_i$表示该集合由$n_i$张图像构成，$d$为特征维度），采用 PCA 得到前 m 个最大的特征向量构成线性子空间 $span(\boldsymbol{Y}_i)$，$\boldsymbol{Y}_i\in\mathbb{R}^{d\times m}$。那么视频人脸识别问题就转化为两个线性子空间距离的度量。

线性子空间的距离一般采用主夹角（Principal Angles）来度量。Kim 等[7]通过学习投影矩阵将原始的线性子空间投影到低维空间，并采用最大化类间主夹角同时最小化类内主夹角，使得特征更具判别性。然而该模型只关注子空间的最小

主夹角，在很多数据中会出现最小主夹角趋于零的情况；除此之外，最小主夹角也不满足距离度量的条件，不能很好地挖掘线性子空间背后的数据分布结构。

因此，涌现出更多[8-11]将线性子空间与格拉斯曼流形进行关联的方法，两者之间的关系如图 2 所示。格拉斯曼流形可以看作由固定维度的线性子空间构成的集合，反过来，线性子空间可以看作流形上的一点，两个线性子空间的距离度量可以转化为流形上两点之间测地线的距离。主夹角和测地线距离之间满足如下关系：

$$d_G^2(\boldsymbol{Y}_1,\boldsymbol{Y}_2)=\sum_i\theta_i^2 \tag{1}$$

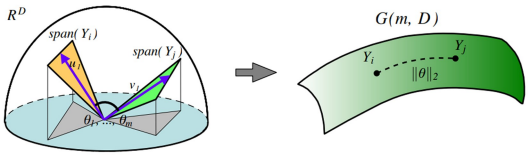


图 2 线性子空间主夹角与格拉斯曼流形距离[8]

Fig. 2 Principal angles of linear subspace and grassmann distances[8]

由此推导出一系列子空间距离[8]，包括投影度量(Projection metric)和比奈-柯西度量(Binet-Cauchy metric)。

由于传统的判别学习方法不适用于格拉斯曼流形，因此 Hamm 等[8]利用核技巧将流形嵌入满足欧氏几何的高维希尔伯特空间再进行判别分析。他们分别基于上述的投影度量和比奈-柯西度量设计了投影核(Projection Kernel)和比奈-柯西核(Binet-Cauchy Kernel)，具体公式如下：

$$\begin{aligned}K_P(\boldsymbol{Y}_1,\boldsymbol{Y}_2)&=\|\boldsymbol{Y}_1^{\mathrm{T}}\boldsymbol{Y}_2\|_F^2\\K_{BC}(\boldsymbol{Y}_1,\boldsymbol{Y}_2)&=(\det\boldsymbol{Y}_1^{\mathrm{T}}\boldsymbol{Y}_2)^2\end{aligned} \tag{3}$$

然后采用这些核函数完成后续的格拉斯曼流形的判别分析。文献[10]更进一步地探索了多种格拉斯曼流形上的核函数，文献[11]则提出直接在格拉斯曼流形上进行判别学习。

2.2 仿射子空间建模方法

仿射子空间是一种更为通用的集合建模方式，它可以建模向量的所有仿射组合，因此有助于挖掘更为复杂的视频数据结构。Cevikalp 等[12]提出将图像集合建模为仿射包：

$$H=\{\boldsymbol{x}=\sum_{i=1}^{n}\alpha_i\cdot\boldsymbol{x}_i\mid\sum_{i=1}^{n}\alpha_i=1\} \tag{4}$$

对样本进行归一化后，仿射包可以简化为 $H=\{\boldsymbol{\mu}+\boldsymbol{U}\boldsymbol{v}\mid\boldsymbol{v}\in\mathbb{R}^l,l<n\}$，其中 $\boldsymbol{\mu}$ 为样本中心，$\boldsymbol{U}$ 为对归一化样本进行奇异值分解得到的一组正交基。那么两个集合之间的距离可以利用仿射包间最近点对之间的距离度量：

$$d(\boldsymbol{H}_i,\boldsymbol{H}_j)=\min_{\boldsymbol{v}_i,\boldsymbol{v}_j}\|(\boldsymbol{\mu}_i+\boldsymbol{U}_i\boldsymbol{v}_i)-(\boldsymbol{\mu}_j+\boldsymbol{U}_j\boldsymbol{v}_j)\|_2^2 \tag{5}$$

由于仿射包的自由度较高，噪声样本很容易干扰判别分析过程，因此文献[12]还对仿射包系数进行了上下界约束 $L\leqslant\alpha\leqslant U$，最终将求解两个仿射包距离的问题转化成了有约束的凸优化问题：

$$\begin{aligned}(\alpha_i^*,\alpha_j^*)=\arg\min_{\alpha_i,\alpha_j}\|\boldsymbol{X}_i\alpha_i-\boldsymbol{X}_j\alpha_j\|^2\sum_{k=1}^{n_i}\alpha_{ik}=1=\sum_{k'=1}^{n_j}\alpha_{jk'},\\L\leqslant\alpha_{ik},\alpha_{jk'}\leqslant U\end{aligned} \tag{6}$$

文献[13]在寻找最近点对时，在组合系数中加入稀疏的约束使得组合得到的点维持在原图像帧集合附近，但系数约束为 L1 范数，计算复杂度较高。

Yang 等[14]和 Zhu 等[15]提出正则化最近点(Regularized Nearest Points)模型，对仿射包系数的 p-范数进行约束，即 $\|\alpha_{l_p}\leqslant\sigma\|$。如图 3 所示，正则化可以有效地将解空间从超平面约束到绿色区域内，避免仿射包包含过多距离样本中心太远的点，在保证性能的基础上大幅度提升计算效率。当约束中取 2-范数时，仿射包的距离可以形式化为马氏距离：

$$\begin{aligned}d_M(\boldsymbol{X}_i,\boldsymbol{X}_j)&=\|\boldsymbol{P}(\boldsymbol{X}_ia_i^*-\boldsymbol{X}_ja_j^*)\|_2^2\\&=(\boldsymbol{X}_ia_i^*-\boldsymbol{X}_ja_j^*)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{X}_ia_i^*-\boldsymbol{X}_ja_j^*)\end{aligned} \tag{7}$$

其中，$\boldsymbol{P}$ 为给定的线性投影矩阵，$\boldsymbol{M}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}$。

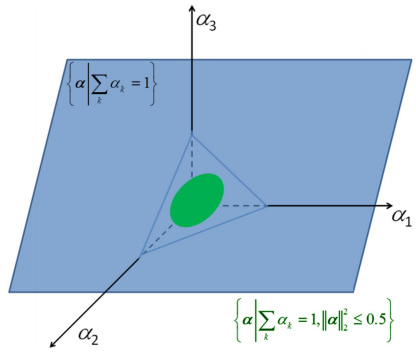


图 3 正则化仿射包的解空间[14]

Fig. 3 Solution space of regularized affine hull[14]

2.3 非线性流形建模方法

当人脸的表观变化同时受到光照、尺寸、姿态、表情等因素的影响时，上述简单的子空间建模方式很难完整地刻画人脸所有的变化模式。因此，Wang 等[16-17]提出利用非线性流形来进行视频建模，将流形看成多个不重叠的局部线性子空间的集合，则两个流形之间的距离可以通过计算两两局部线性子空间的距离得到。定义流形-流形距离(Manifold to Manifold Distance)如式(8)所示：

$$d(M_1,M_2)=\min_{C_i\in M_1}\min_{C_j'\in M_2}d(C_i,C_j') \tag{8}$$

其中，C 表示一个最大线性块(Maximal Linear Patch)。

进一步地，文献[16]提出可以利用流形上两个局部区域的欧氏距离矩阵(Euclidean Distance Matrix)和测地距离矩阵(Geodesic Distance Matrix)的比例反映该局部的线性扰动(Linear Perturbation)程度。给定阈值 θ，随机选取流形上的一点，通过计算该点与其 K 近邻点形成的局部区域的线性扰动程度(扰动程度小于阈值就认为属于线性局部块)，最终逐步形成最大线性块。定义两个最大线性块的距离后，基于式(8)就可以计算得到流形之间的距离。

文献[17]提出在判别特征空间内度量流形-流形距离可以进一步提升模型的识别性能。其采用分层分裂聚类(Hierarchical divisive clustering)的算法计算流形上的最大线性块，并构造对应的本征图(Intrinsic Graph)G 和惩罚图(Penalty Graph)G'，用以刻画两两最大线性块的相似关系。流形判别分析过程为学习嵌入矩阵 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_t]$，$l\ll d$，最大化目标：

$$J(\boldsymbol{v})=\frac{|\boldsymbol{S}_b|}{|\boldsymbol{S}_w|}=\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}_b\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}_w\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}} \tag{9}$$

其中，$\boldsymbol{L}_b$，$\boldsymbol{L}_w$ 分别是惩罚图及本征图的拉普拉斯矩阵。

Cui 等[18]提出流形对齐的方法，将所有流形统一对齐到

一个参考流形再计算流形-流形距离；Chen 等[19]通过稀疏表示划分局部线性子空间，采用联合的稀疏表示搜索最近线性子空间。

2.4 统计建模方法

由于人脸视频中的每一帧图像均可以看作对人脸的一次采样，同时视频中包含了丰富的人脸变化样本，因此一部分基于统计建模的工作希望统计视频数据潜在的模式对人脸进行建模。

一些工作利用高斯模型和混合高斯模型对人脸样本进行统计建模。文献[20]对高斯分布的参数进行估计，用 KL 散度(Kullback-Leibler Divergence)度量两个分布 $g_i=(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ 及 $g_j=(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ 之间的距离：

$$KLD(g_i \parallel g_j)=\frac{1}{2}(\mathrm{tr}(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i)+(\boldsymbol{\mu}_j-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_i)-\ln\left(\frac{\det\boldsymbol{\Sigma}_i}{\det\boldsymbol{\Sigma}_j}\right)-D) \tag{10}$$

其中，D 为特征维度。文献[21]采用混合高斯模型进行建模，同时考虑了更多高斯分布的度量，如 Bhattacharyya 距离、Hellinger 距离等，并推导出对应的正定核函数，将高斯模型映射到高维的希尔伯特空间进行判别分析，进一步提升了视频人脸识别的性能。

Wang 等[22]将图像集合的二阶信息(协方差矩阵 $\boldsymbol{C}$)作为统计量进行建模。协方差矩阵是对称正定矩阵，位于黎曼流形上，因此文献[22]利用流形上的对数欧氏度量(Log-Euclidean Distance)导出黎曼核函数，将流形上的点映射到高维欧氏空间中进行核判别分析。其公式如下：

$$d_{LED}(\boldsymbol{C}_1,\boldsymbol{C}_2)=\|\log(\boldsymbol{C}_1)-\log(\boldsymbol{C}_2)\|_F \tag{11}$$

其中，$\log(\boldsymbol{C})=\boldsymbol{U}\log(\boldsymbol{\Sigma})\boldsymbol{U}^{\mathrm{T}}$，$\boldsymbol{U}$ 为协方差矩阵的特征向量，$\boldsymbol{\Sigma}$ 是特征值构成的对角矩阵。

进一步地，文献[23]引入图像深度特征，实现了端到端的学习。文献[24]指出这种度量方式会破坏原始流形的几何结构，同时计算效率也很低。为了解决这些问题，文献[24]提出学习从原始流形切空间到更具判别力的切平面空间的映射 $DF(S)$，直接在流形上进行判别分析：

$$T_S\mathbb{S}_+^d \to T_{F(S)}\mathbb{S}_+^k,\ \xi \mapsto DF(S)[\xi_s] \tag{12}$$

3 基于图像融合的建模方法

近年来，随着深度学习带来单帧图像特征表达能力的显著提升，研究者将研究重点转移到如何对深度网络提取得到的大量图像特征表示进行合理有效的融合，从而整合视频中所有帧包含的判别信息，得到主体的统一人脸特征。

由于视频数据中人脸的变化过程往往是连续的，其中可能包含大量冗余图像信息，因此一些工作在图像层面对冗长的视频数据进行整合。Hassner 等[25]采用分类的方式对图像进行整合，他们提出将视频中的所有图像按照头部姿态划分为不同子集，并通过 3D 头部姿态对齐减少外观变化，再经过统一的深度网络提取得到图像特征表示，最后利用求平均的方式快速得到人脸视频的统一表示。Rao 等[26]提出利用所有图像帧训练生成对抗式网络(Generative Adversarial Network)虚拟地合成一张尽可能与真实视频帧贴近但同时更具辨别力的图像来代表整个图像集合，再进行后续的判别分析。这样可以很好地避免因复杂的视频数据在特征空间中的分布过于杂乱而降低模型判别的鲁棒性。Shi 等[27]将图像看成对该人物在人脸特征空间的一次观察值采样，利用多元高斯分布对观察过程进行建模，则分类问题可以看作估计条件概率：

$$p(\boldsymbol{z}|\boldsymbol{x}_i)=N(\boldsymbol{z};\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i^2\boldsymbol{I}) \tag{13}$$

其中，$\boldsymbol{x}_i$ 为图像特征，$\boldsymbol{z}\in Z$ 表示该人物在人脸特征空间的唯一表示，$\boldsymbol{\mu}_i$ 和 $\boldsymbol{\sigma}_i^2$ 为神经网络需要学习的模型参数。则由多个图像构成的视频分类问题可以看作联合条件概率的估计：

$$p(\boldsymbol{z}|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{n+1})=\alpha\frac{p(\boldsymbol{z}|\boldsymbol{x}_{n+1})}{p(\boldsymbol{z})}p(\boldsymbol{z}|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n) \tag{14}$$

为了保证人脸识别的效率，以上研究直接在图像层面上进行了信息整合，而更多研究则是在特征层面进行融合以保留更多判别信息。一些研究[28-29]希望对视频中的每帧图像进行质量评估，并以质量分数为权重对图像特征进行加权融合。这种方式一方面可以保证保留高质量图像中的判别性信息，同时也能避免低质量的噪声图像参与融合。Liu 等[28]提出了两分支网络：一个用来提取每个人脸图像的外观特征，另一个用来估计视频中每帧人脸图像的质量分数。视频级特征 $\boldsymbol{R}_a(S_i)$ 的计算公式如下：

$$\boldsymbol{R}_a(S_i)=\sum_{l=1}^{N_i}Q(\boldsymbol{x}_{i,l})\boldsymbol{F}(\boldsymbol{x}_{i,l})/\sum_{l=1}^{N_i}Q(\boldsymbol{x}_{i,l}) \tag{15}$$

其中，$Q(\boldsymbol{x}_{i,l})$ 为第 i 个图像集合中第 l 帧图像 $\boldsymbol{x}_{i,l}$ 的质量分数，$\boldsymbol{F}(\boldsymbol{x}_{i,l})$ 为对应特征向量。之后，其利用视频级特征计算三元组损失(Triplet Loss)优化网络参数：

$$L_{QAN}=\max(\|\boldsymbol{R}_a(S_a)-\boldsymbol{R}_a(S_p)\|_2^2-\|\boldsymbol{R}_a(S_a)-\boldsymbol{R}_a(S_n)\|_2^2+\delta,0) \tag{16}$$

其中，$R_a(S_a)$，$\boldsymbol{R}_a(S_p)$，$\boldsymbol{R}_a(S_n)$ 为一个三元组。经过推导，这种加权融合方式可以自适应地降低与视频代表性特征图像相差较大的帧的质量分数。Yang 等[29]提出了 NAN(Neural Aggregation Network)方法，网络中采用两个级联的注意力机制模块，第一级评估全局图像质量，第二级进一步得到基于图像内容的质量分数。图 4 是其对图像质量的估计情况，可以看到高清正脸图像的质量分数更高，难以分辨的图像的质量分数更低。

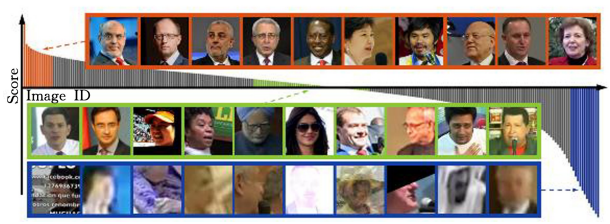


图 4 NAN 方法对视频序列内图像质量的评估结果[29]

Fig. 4 Evaluation results of image quality in video sequence by NAN method[29]

Zhang 等[30]训练了一个判别力蒸馏网络(Discriminability Distillation Network)来回归图像的判别性，图像判别性的计算方法如下：

$$D_i=\frac{d_{ip}}{\max\{d_{in}\mid n\in[1,K],n\neq p\}}$$
$$d_{ip}=\frac{\boldsymbol{f}_i\cdot\boldsymbol{C}_p}{\|\boldsymbol{f}_i\|_2\|\boldsymbol{C}_p\|_2} \tag{17}$$
$$d_{in}=\frac{\boldsymbol{f}_i\cdot\boldsymbol{C}_n}{\|\boldsymbol{f}_i\|_2\|\boldsymbol{C}_n\|_2}$$

其中，K 代表类别数，$\boldsymbol{C}_p$，$\boldsymbol{C}_n$ 分别代表正样本和负样本的类别中心。由此可见，离正样本中心越近、离负样本中心越远的样本的判别性越强。依据归一化后得到的判别力分数（D-score），模型可以过滤掉图像集合中判别性较低的噪声样本，保留判别性较强的样本进行识别。

Zhong 等[31]借鉴 NetVLAD[32]的思想，首先对所有图像的深度特征进行聚类，得到 $K+G(K\ll N)$ 个聚类中心$\{\boldsymbol{c}_k\}$，其中 G 为额外分配的聚类中心，Zhong 等称其为伪中心（Ghost Center），N 为视频中的图像数量。之后基于这些聚类中心对同一视频的所有图像特征进行融合。

$$V(j,k)=\sum_{i=1}^{N}\frac{e^{\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_k}}{\sum_{k'=1}^{K}e^{\boldsymbol{a}_{k'}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_{k'}}}(\boldsymbol{x}_i(j)-\boldsymbol{c}_k(j)) \tag{18}$$

其中，$\{\boldsymbol{a}_k\}$，$\{\boldsymbol{b}_k\}$，$\{\boldsymbol{c}_k\}$为可训练的参数。第一项计算得到聚类中心的影响权重；第二项计算残差量，其中伪中心参与权重计算，但不参与后续融合，Zhong 等认为这种方式可以增强模型融合特征的鲁棒性。

Liu 等[33]指出，这种独立给出每张视频帧的质量分数的方法会导致模型给予高清正脸图像过高的分数，从而无法从模糊图像获取人脸的轮廓信息。因此，权重的计算应该考虑视频序列之间的相关性，于是他们采用强化学习的方式，根据序列之间的相关关系抽取具有人物判别性信息的视频帧代表。

还有研究在融合特征时考虑了局部特征的空间信息。Xie 等[34]设计了一种联合人脸检测及识别进行人脸校验的对比网络（Comparator Network），其指出直接对图像特征进行融合没有考虑以下两个方面：1）人脸校验在相同视点的条件下更加容易；2）人脸校验应该对人脸的关键点进行比对。因此该研究首先对人脸的关键点进行了检测，将关键点的归一化响应值作为注意力分数，在整体特征上计算哈达玛积作为某个关键点处的局部特征。在进行人脸比对时，首先逐一对 K 个关键点对应的局部特征进行比对，再通过通道级的最大池化操作得到最终的校验结果。

Gong 等[35]提出了更为细粒度的特征分量级的特征融合网络（Component-wise Feature Aggregation Network），其简要证明了将每张图像对应的特征看作样本，采用质量分数作为权重进行融合相比简单的平均池化可以更大程度地降低噪声信息。因此他们将结论推广到每个特征分量，将每个局部特征看作样本进行加权融合，进一步降低了局部特征的噪声信息，得到更具判别性的人脸特征。Liu 等[36]将同一视频序列内其他视频帧特征对应位置处的相关信息作为当前帧的残差特征进行重构，使得每一帧的特征更具判别性。

4 数据库及评测协议

4.1 数据库

在过去近二十年中，研究者构建并发布了多个基于视频的人脸识别数据集。早期的数据集除了规模较小外，其拍摄环境大多为光照条件较好的室内环境，同时视频内容往往是人物完成一些指定动作。拍摄环境和视频内容的单一性使得早期数据集的人脸变化模式较为固定。为了增加人脸变化模式的多样性，研究者考虑从更具挑战性的非受限场景下获取数据，如真实的监控场景及互联网在线视频数据（主要来源于 YouTube，YouKu，iQIYI 等视频网站），这些场景下人物对数据的采集并不知情，因此人脸的状态完全不受限制，如此得到的视频数据更加贴近真实应用需求。图 5 给出了部分不同场景下采集得到的数据示例。

图 5 视频人脸识别数据集示例

Fig. 5 Examples of datasets for video-based face recognition

根据数据集中人脸变化受限的程度，我们将常用的数据集分成3种：全受限、半受限及非受限，并简要介绍了这些数据集的收集方式和各自的特点。

4.1.1 全受限数据集

Honda/UCSD数据集[37-38]是典型的全受限视频人脸数据集，视频在同一室内场景下进行采集。每段视频中，人物可以随意进行二维、三维的头部旋转，因此人脸的姿态变化非常多样。整体数据包含两个数据子集，第一个子集包含20个人物的45段视频，第二个子集包含20个人物的52段视频。XM2VTSDB数据集[39]包含295个人物的1 180段视频。数据包含两种镜头，一种镜头要求人物面对摄像机朗读句子，另一种镜头则要求人物按顺序向左右上下旋转头部。MobFaces数据集[40]包含50个人物的750段视频，该数据集主要面向手机面部识别任务，因此采集了人物在不同光照、不同背景环境下进行5种任务时的面部视频来模拟手机登录场景。全受限数据集为室内环境中采集得到的数据，主要考虑头部姿态及光照对视频人脸识别带来的影响。但是由于环境条件的设置较为固定，因此人脸变化模式比较单一。

表1 视频人脸识别数据集特征

Table 1 Characteristics of various datasets for video-based face recognition

Constraints	Dataset	# Subjects	# Videos	# Frames	Highlights
Fully Constrained	1st Honda/UCSD	20	45	>600 *	Indoor environment
	2nd Honda/UCSD	20	52	>300 *	Indoor environment
	XM2VTSDB	295	—	—	Rotation/speech shot
	MobFaces	50	750	—	Captured by front camera of mobile phone
Less Constrained	CMU FIA	180	—	600 *	Simulate passport checking scenarios
	ChokePoint	29	48	—	Walking through portals
	MBGC	146	770	—	Walking or take activities such as conversation
	COX	1 000	3 000	~411 000	Walking along S-shape route in an indoor gym
	PaSC	265	2 802	~600 000	Record by controlled camera & handheld camera
	IJB-S	202	2 424	—	Real-world surveillance environments
Unconstrained	YTC	47	1 910	304 689	Collected from YouTube
	YTF	1 595	3 425	620 940	Collected from YouTube
	Celebrity1000	1 000	7 021	2 452 519	Large-scale/ Collected from YouTube/YouKu
	IJB-A	500	2 042	51 434	Template-base
	IJB-B	1 845	7 011	227 641	More uniform geographic distribution than IJB-A
	IJB-C	3 531	11 779	469 418	Large-scale
	UMDFaces-Video	3 107	22 075	3 800 000	Large-scale/Collected from YouTube
	iQIYI-VID	5 000	600 000	—	Multimodality

注：* 表示引用原文数据

4.1.2 半受限数据集

CMU FIA数据集[41]主要模拟护照检查场景，由180名参与者拍摄得到的长度约为20 s的视频组成。数据采用3个不同角度下的6个摄像机进行拍摄，拍摄环境有两种，一种是光线较好的室内环境，另一种为更具挑战的开放室外场景。Chokepoint数据集[42]则面向真实监控场景，利用3个不同角度架设的摄像机同时拍摄，模拟行人通过检查点进行人脸识别的过程。MBGC数据集[43]包含821个人物的3 764段视频，每段视频要求参与者做出如行走、交谈等指定动作。COX数据集[44]同样模拟视频监控场景，并且考虑了静态图片与视频序列之间的匹配任务，数据集包含了1 000名参与者的3 000段视频（每名参与者录制3段视频），同时也包含每名参与者的一张正面静态图像。PaSC数据集[45]包含265个人的2802段在多个场景下采集的视频，视频要求每个人做出更多的规定动作，并采用手持相机和固定相机分别拍摄，其中手持相机拍摄会造成视频的抖动和模糊，具有较大的挑战性。IJB-S数据集[46]面向更为复杂的视频监控场景，参与者会在学校、公交站台、地铁站、医院等多个监控点走动，且摄像机架设得较高，基本可以完全模拟真实的视频监控场景。半受限数据集对拍摄环境的约束有了明显的降低，包含了很多更加贴近真实监控场景的视频数据，人脸的分辨率普遍降低，但是参与者仍然是在拍摄者设定的环境下进行拍摄，因此距离真实场景还有一定距离。

4.1.3 非受限数据集

YTC数据集[47]采集了YouTube网站上47位名人的1 910段视频，这些视频的质量往往较差，具有分辨率低、压缩比高的特点。YTF数据集[48]是首个提出非受限视频人脸识别概念的数据集，包含1 595个人物的3 425段视频。其数据来源是在线YouTube视频。这些视频大多由业余爱好者创作，因此光照及视频质量都较差。Celebritiy-1000数据集[49]是大规模非受限视频人脸识别数据集，包含来自YouTube及YouKu网站上的1 000位名人的7 021段视频。每段视频被裁剪为多个片段，构成159726个视频片段。视频内容包含名人的采访、演讲及新闻播报等。IJB系列数据集[50-52]经过不断地扩展，包含IJB-A，IJB-B，IJB-C3个数据集。这个系列的数据集提出模板（Template）匹配的概念，模板既包含人物的静态图片，同时也包含人物的视频人脸序列。数据收集自互联网，每个人物经过人工确认保证地理分布多样化，因此人脸的变化模式多样。UMDFaces数据集[53]的规模非常庞大，该数据集包含两部分，一部分是静态人脸数据集，另一部分UMDFaces-Videos[54]收集了静态图像人脸数据集batch-1子集中包含的3107位人物的22075段视频，因此可以用于从头训练深度卷积网络。IQIYI-VID数据集[55]同样是大规模数据集，包含5000位名人的600000个视频序列，数据收集于爱奇艺视频平台，包括电影、电视剧、综艺及新闻等各种媒体形式，名人主要来自亚洲。该数据集还提供了头部、身体、语音

标注等多模态信息。非受限数据集的数据来源主要集中在各大网站上收集得到的在线视频，因此视频采集环境的多样性及人物姿态表情的多变性都给视频人脸识别带来了巨大的挑战，且更加贴近实际应用需求。

4.2 评价指标

引言中介绍的人脸识别分为人脸校验任务及人脸辨识任务。在人脸校验任务中，通常采用 ROC(Receiver Operating Characteristic)及 ACC(Estimated Mean Accuracy)进行模型性能的评价指标。其中 ROC 反映了在给定的错误接受率(False Accept Rate，FAR)下模型的正确接收率(True Accept Rate，TAR)。误识率越低，正确识别率越高，说明模型越鲁棒。ACC 即为正确识别身份的百分比。在闭集人脸识别任务中通常采用 rank-N 正确率及 CMC(Cumulative Match Characteristic)曲线作为评价指标。rank-N 正确率是指用测试数据检索系统人脸库时返回的置信度排序的前 N 个结果中身份正确的比例。CMC 曲线反映的是在 rank-N 中给定不同的 N 时计算得到的正确率变化。在开集人脸识别任务中，模型需要拒绝没有在人脸库注册过的测试数据，评价指标采用给定假阳识别率(False Positive Identification Rate，FPIR)时的真阳识别率(True Positive Identification Rate，TPIR)，FPIR 越低时，TPIR 越高，模型越鲁棒。

5 算法评测

深度学习兴起之前，视频人脸识别算法大多基于手工设计的特征对图像集合进行建模。随着深度学习的发展，很多方法直接利用深度神经网络提取的图像特征进行信息融合得到视频级特征表示。接下来本文将通过对比实验，分别评测基于传统集合建模的代表算法和基于图像融合的算法在手工设计的特征及深度特征下的表现。

5.1 实验设置

本文采用 MTCNN 算法[56]进行人脸检测，裁剪出 224×224 的人脸区域。之后分别选取灰度特征，由 10 层卷积网络(Conv-10)提取得到的浅层深度特征和 VGG 研究组发布的 50 层深度残差网络[57](ResNet-50)提取得到的深层深度特征作为图像的特征表示。具体地，Conv-10 在大规模人脸图像数据集 CASIA-WebFace[58]中进行预训练，神经网络中每两个卷积层后接一个平均池化层，卷积层都默认连接 BN 层[59]和 ReLU 层[60]。ResNet50 则遵循原文设置。

考虑到手工设计特征的表示能力有限，本文首先选取了视频人脸识别中的经典数据集 YTC，并分别采用灰度特征及浅层深度特征对比不同算法在该数据集上的表现。另外，我们选取更具挑战性的 IJB-A 数据集进一步测试算法利用深度特征完成人脸识别的性能。

用于评测的对比算法选取了第二节中各类图像集合建模方法中的代表性算法：线性子空间建模选用 DCC[7]和 GDA[8]；仿射子空间建模选用 AHISD[12]，CHISD[12]和 RNP[14]；非线性流形中选择了 MMD[16]和 MDA[17]；基于统计建模的方法选择了 CDL[22]和 LEML[24]；基于图像特征融合的方法选择了 NAN[29]，并将直接进行平均池化(AVE)作为该类方法的基准(Baseline)。实验同时总结了更多近几年已发表文章的实验结果。

5.2 实验结果分析

表 2 列出了上述代表性算法在 YTC 数据集上的识别准确率结果。对比两列结果可以发现，采用浅层深度特征的模型性能在所有建模方式下相比灰度特征都有明显提升，且采用简单的平均池化就能取得较好的识别结果。这充分说明了神经网络经过大规模数据的预训练后可以有效提取人脸的判别性特征，并具有较强的鲁棒性。

表 2 YTC 数据集上的性能对比

Table 2 Performance comparison on YTC dataset

Algorithm	Gray features	Conv-10
DCC	0.640	0.977
GDA	0.657	**0.986**
AHISD	0.637	0.959
CHISD	0.663	0.962
RNP	**0.703**	0.972
MMD	0.639	0.971
MDA	0.657	0.958
CDL	0.698	0.983
LEML	0.669	0.978
AVE	0.525	0.975
NAN	0.536	0.985

除此之外，对比传统图像集合建模的方法及基于特征融合的方法在灰度特征上的表现，可以发现传统图像集合建模的方法更具优势，其中 RNP 取得了最佳结果，并且相比仅利用最近点对度量仿射子空间的方法有明显的提升，表明正则化约束仿射子空间的解空间可以有效降低噪声样本的干扰。同时，对比所有方法与平均池化(AVE)方法的结果可以发现，在底层图像特征表达能力不足时，所有方法的结果都优于 AVE，这充分说明对视频进行有效建模可以从复杂的视频内容中抽取到人脸的潜在变化模式。

基于特征融合的方法在灰度特征中表现较差，但在深度特征下结果较好，这说明基于图像融合的视频建模方式对深度特征有较强的依赖性。NAN 方法相比基准方法 AVE 有一定性能提升，这说明了模型在视频建模阶段可以依据图像质量找到高质量代表样本并排除噪声样本的干扰。

表 3 列出了采用浅层深度特征作为图像特征表示在更具挑战性的 IJB-A 数据集上的人脸识别结果。

表 3 IJB-A 数据集上的性能对比

Table 3 Performance comparison on IJB-A dataset

Algorithm	TAR@FAR=1×10^{-3}	TAR@FAR=1×10^{-2}
DCC	0.561	0.778
GDA	0.754	0.920
AHISD	0.557	0.800
CHISD	0.590	0.791
RNP	0.790	0.936
MMD	0.739	0.921
MDA	0.542	0.743
CDL	0.729	0.915
LEML	0.578	0.838
AVE	0.750	0.936
NAN	**0.849**	**0.942**

与表 2 第二列的结果相比，大部分算法在 IJB-A 数据集上的识别准确率相比简单的 YTC 数据集都有大幅度的降低，

特别是传统集合建模的方法，如 DCC，AHISD，CHISD，MDA 等。这一方面说明了数据集规模的增大和拍摄的约束程度降低都会加大识别难度；同时也说明了由于神经网络提取的深度特征本身具有较强的判别性，集合内部样本特征较为集中，使得传统集合建模方法不能很好地基于这些深度特征建模人脸的变化模式。相较于传统集合建模方法，NAN 及 AVG 在 IJB-A 数据集上的优势更加突出，充分说明了深度特征在非受限数据集中仍然具有较强的判别性，且 NAN 方法能够更进一步地提高模型的判别能力。

表 4 列出了采用 ResNet-50 得到的深层深度特征在 IJB-A 数据集上的结果，表 4 同时汇总整理了更多近几年代表性工作在人脸验证及人脸辨识任务上的性能对比。可以看到，传统基于图像集合建模的方法表现较差，而基于深度学习的方法采用更深层的深度特征后的识别正确率已经非常高了。因此，研究者应积极挖掘更具挑战性的测试数据集及识别任务来不断贴近实际应用需求。

表 4 代表性工作在 IJB-A 数据集上的性能对比

Table 4 Performance comparison on IJB-A dataset of representative researches

Algorithm	Backbone	Training Data	1:1 Verification TAR		Close-set identification		Open-set identification TPIR	
			FAR=1×10^{-3}	FAR=1×10^{-2}	rank1	rank5	FPIR=1×10^{-2}	FPIR=1×10^{-1}
DCC	ResNet-50	VGGFace2(3M)	0.857	0.957	—	—	—	—
GDA	ResNet-50	VGGFace2(3M)	0.520	0.782	—	—	—	—
AHISD	ResNet-50	VGGFace2(3M)	0.807	0.968	—	—	—	—
RNP	ResNet-50	VGGFace2(3M)	0.959	0.983	—	—	—	—
MDA	ResNet-50	VGGFace2(3M)	0.844	0.939	—	—	—	—
CDL	ResNet-50	VGGFace2(3M)	0.533	0.798	—	—	—	—
LEML	ResNet-50	VGGFace2(3M)	0.764	0.938	—	—	—	—
DAN*[26]	SENet-50	VGGFace2(3M)	0.910	0.941	0.980	0.990	—	—
NAN*[29]	GoogleNet	Crawled(3M)	0.881	0.941	0.985	0.980	0.817	0.917
QAN*[28]	—	Ext. VGGFace(5M)	0.893	0.942	—	—	—	—
GhostVLAD*[31]	SENet-50	VGGFace2(3M)	0.935	0.972	0.977	**0.991**	0.884	0.951
DAC*[33]	GoogleNet	Crawled(3M)	—	0.954	0.973	—	0.855	0.934
C-FAN*[35]	Face-ResNet	MS1M(10M)	0.916	0.940	0.946	0.963	0.869	0.929
PIFR*[36]	ResNet-50	VGGFace2(3M)	0.955	0.983	**0.990**	—	**0.908**	**0.969**
DDN*[30]	ResNet-101	MS1M(10M)	**0.984**	**0.988**	—	—	—	—

注：* 表示引用原文数据

结束语 本文对国内外视频人脸识别的研究工作进行了综述，视频人脸识别的主要技术挑战在于从冗长而复杂的视频数据中提取出人脸的关键信息。我们首先对基于传统图像集合建模的研究工作进行分类总结；其次对深度学习背景下的图像融合工作进行介绍；最后介绍了视频人脸识别数据库及评测协议，并选择了一些代表性算法进行对比评测。

实验结果一方面表明了经过预训练的深度神经网络可以有效提取视频中每一帧图像的判别性特征，基于图像融合的视频建模方法可以从这些大量人脸样本中整合有效信息，排除噪声信息的影响，建立起统一的人脸特征表示。但目前基于深度特征的视频建模算法对特征的依赖性较强，一旦特征表示能力不足，将会大幅度影响模型的识别性能。另一方面，传统集合建模的方法在特征表达能力不足时可以有效建模人脸的变化模式，但在特征本身具有较强的判别性时，同一人物的整体人脸特征变化缩小会导致部分方法的识别性能有所降低。总体来看，尽管当前基于视频的人脸识别方法在小规模数据集上可以取得不错的识别准确率，但当视频数据规模逐步扩大同时人脸受限程度逐步降低时，人脸识别的性能还有待提升。因此如何对神经网络提取的深度特征进行高效高质的整合，从而得到统一的人脸特征表示仍是当前该领域的研究重点。如何将传统图像集合建模方式刻画人脸变化模式的能力与深度特征的判别能力进行更有机的结合值得研究者们的进一步探索。另外，更加贴近实际应用需求的测试数据及识别任务值得进一步挖掘。

参 考 文 献

[1] CHEN S, MAU S, HARANDI M T, et al. Face recognition from still images to video sequences: a local-feature-based framework [J]. Journal on Image and Video Processing, 2011, 2011(1): 1-14.

[2] LI Z, ZHANG J, ZHANG K, et al. Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4478-4489.

[3] SIROVICH L, KIRBY M. Low-dimensional procedure for the characterization of human faces[J]. Josa A, 1987, 4(3): 519-524.

[4] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(7): 971-987.

[5] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[6] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005: 886-893.

[7] KIM T K, KITTLER J, CIPOLLA R. Discriminative learning and recognition of image set classes using canonical correlations [J]. IEEE Transactions on Pattern Analysis and Machine Intel-

ligence, 2007, 29(6): 1005-1018.

[8] HAMM J, LEE D D. Grassmann discriminant analysis: a unifying view on subspace-based learning[C]// Proceedings of the 25thInternational Conference on Machine Learning. 2008: 376-383.

[9] HARANDI M T, SALZMANN M, JAYASUMANA S, et al. Expanding the family of grassmannian kernels: An embedding perspective[C]// European Conference on Computer Vision. Springer, Cham, 2014: 408-423.

[10] HARANDI M T, SANDERSON C, SHIRAZI S, et al. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2011: 2705-2712.

[11] HUANG Z, WANG R, SHAN S, et al. Projection metric learning on Grassmann manifold with application to video based face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. 2015: 140-149.

[12] CEVIKALP H, TRIGGS B. Face recognition based on image sets[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2567-2573.

[13] HU Y, MIAN A S, OWENS R. Sparse approximated nearest points for image set classification[C]// 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2011: 121-128.

[14] YANG M, ZHU P, VAN GOOL L, et al. Face recognition based on regularized nearest points between image sets[C]// 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 2013: 1-7.

[15] ZHU P, ZHANG L, ZUO W, et al. From point to set: Extend the learning of distance metrics[C]// Proceedings of the IEEE International Conference on Computer Vision. 2013: 2664-2671.

[16] WANG R, SHAN S, CHEN X, et al. Manifold-manifold distance with application to face recognition based on image set[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[17] WANG R, CHEN X. Manifold discriminant analysis[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 429-436.

[18] CUI Z, SHAN S, ZHANG H, et al. Image sets alignment for video-based face recognition[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2626-2633.

[19] CHEN S, SANDERSON C, HARANDI M T, et al. Improved image set classification via joint sparse approximated nearest subspaces[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 452-459.

[20] SHANKHNAROVICH G, FISHER J W, DARRELL T. Face recognition from long-term observations[C]// European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2002: 851-865.

[21] WANG W, WANG R, HUANG Z, et al. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2048-2057.

[22] WANG R, GUO H, DAVIS L S, et al. Covariance discriminative learning: A natural and efficient approach to image set classification[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2496-2503.

[23] WANG W, WANG R, SHANS, et al. Discriminative covariance oriented representation learning for face recognition with image sets[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5599-5608.

[24] HUANG Z, WANG R, SHAN S, et al. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification[C]// International Conference on Machine Learning. 2015: 720-729.

[25] HASSNER T, MASI I, KIM J, et al. Pooling faces: Template based face recognition with pooled face images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 59-67.

[26] RAO Y, LIN J, LU J, et al. Learning discriminative aggregation network for video-based face recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 3781-3790.

[27] SHI Y, JAIN A K. Probabilistic face embeddings [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 6902-6911.

[28] LIU Y, YAN J, OUYANG W. Quality aware network for set to set recognition[C]// Proceedings of the IEEE Confe-rence on Computer Vision and Pattern Recognition. 2017: 5790-5799.

[29] YANG J, REN P, ZHANG D, et al. Neural aggregation network for video face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4362-4371.

[30] ZHANG M, SONG G, ZHOU H, et al. Discriminability distillation in group representation learning[C]// European Conference on Computer Vision. Springer, Cham, 2020: 1-19.

[31] ZHONG Y, ARANDJELOVIC R, ZISSERMAN A. GhostVLAD for set-based face recognition[C]// Asian Conference on Computer Vision. Springer, Cham, 2018: 35-50.

[32] ARANDJELOVIC R, GRONAT P, TROII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.

[33] LIU X, VIJAYA K B V K, YANG C, et al. Dependency-aware attention control for unconstrained face recognition with image sets[C]// Proceedings of the European Conference on Computer Vision. 2018: 548-565.

[34] XIE W, SHEN L, ZISSERMAN A. Comparator networks[C]// Proceedings of the European Conference on Computer Vision. 2018: 782-797.

[35] GONG S, SHI Y, KALKA N D, et al. Video face recognition: Component-wise feature aggregation network (c-fan)[C]// 2019 International Conference on Biometrics. IEEE, 2019: 1-8.

[36] LIU X, GUO Z, LI S, et al. Permutation-invariant feature restructuring for correlation-aware image set-based recognition [C]// Proceedings of the IEEE International Conference on

Computer Vision. 2019：4986-4996.

[37] LEE K C，HO J，YANG M H，et al. Video-based face recognition using probabilistic appearance manifolds[C]// 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Proceedings. IEEE，2003：I/313-I/320.

[38] LEE K C，HO J，YANG M H，et al. Visual tracking and recognition using probabilistic appearance manifolds[J]. Computer Vision and Image Understanding，2005，99(3)：303-331.

[39] MESSER K，MATAS J，KITTLER J，et al. XM2VTSDB：The extended M2VTS database[C]// Second International Conference on Audio and Video-based Biometric Person Authentication. 1999：965-966.

[40] FATHY M E，PATEL V M，CHELLAPPA R. Face-based active authentication on mobile devices[C]// 2015 IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2015：1687-1691.

[41] GOH R，LIU L，LIU X，et al. The CMU face in action (FIA) database[C]// InternationalWorkshop on Analysis and Modeling of Faces and Gestures. Berlin，Heidelberg：Springer，2005：255-263.

[42] WONG Y，CHEN S，MAU S，et al. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition[C]// CVPR 2011 WORKSHOPS. IEEE，2011：74-81.

[43] PHILLIPS P J，FLYNN P J，BEVERIDGE J R，et al. Overview of the multiple biometrics grand challenge[C]// International Conference on Biometrics. Berlin，Heidelberg：Springer，2009：705-714.

[44] HUANG Z，SHAN S，WANG R，et al. A benchmark and comparative study of video-based face recognition on cox face database[J]. IEEE Transactions on Image Processing，2015，24(12)：5967-5981.

[45] BEVERIDGE J R，PHILLIPS P J，BOLME D S，et al. The challenge of face recognition from digital point-and-shoot cameras [C]// 2013 IEEE Sixth International Conference on Biometrics：Theory，Applications and Systems. IEEE，2013：1-8.

[46] KALKA N D，MAZE B，DUNCAN J A，et al. IJB-S：IARPA Janus surveillance video benchmark[C]// 2018 IEEE 9th International Conference on Biometrics Theory，Applications and Systems. IEEE，2018：1-9.

[47] KIM M，KUMAR S，PAVLOVIC V，et al. Face tracking and recognition with visual constraints in real-world videos[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2008：1-8.

[48] WOLF L，HASSNER T，MAOZ I. Face recognition in unconstrained videos with matched background similarity[C]// CVPR 2011. IEEE，2011：529-534.

[49] LIU L，ZHANG L，LIU H，et al. Toward large-population face identification in unconstrained videos[J]. IEEE Transactions on Circuits and Systems for Video Technology，2014，24(11)：1874-1884.

[50] KLARE B F，KLEIN B，TABORSKY E，et al. Pushing the frontiers of unconstrained face detection and recognition：Iarpa janus benchmark-a[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015：1931-1939.

[51] WHITELAM C，TABORSKY E，BLANTON A，et al. Iarpa janus benchmark-b face dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017：90-98.

[52] MAZE B，ADAMS J，CUNCAN J A，et al. Iarpa janus benchmark-c：Face dataset andprotocol[C]// 2018 International Conference on Biometrics. IEEE，2018：158-165.

[53] BAMSAL A，NANDURI A，CASTILLO C D，et al. Umdfaces：An annotated face dataset for training deep networks[C]// 2017 IEEE International Joint Conference on Biometrics. IEEE，2017：464-473.

[54] BAMSAL A，CASTILLO C，RANJAN R，et al. The do's and don'ts for cnn-based face verification[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017：2545-2554.

[55] LIU Y，PENG B，SHI P，et al. iqiyi-vid：A large dataset for multi-modal person identification[J]. arXiv：1811. 07548，2018.

[56] ZHANG K，ZHANG Z，LI Z，et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters，2016，23(10)：1499-1503.

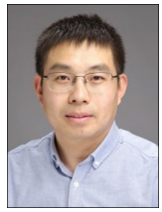
[57] CAO Q，SHEN L，XIE W，et al. Vggface2：A dataset for recognising faces across pose and age[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE，2018：67-74.

[58] YI D，LEI Z，LIAO S，et al. Learning face representation from scratch[J]. arXiv：1411. 7923，2014.

[59] IOFFE S，SZEGEDY C. Batch normalization：Accelerating deep network training by reducing internal covariate shift[J]. arXiv：1502. 03167，2015.

[60] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

BAI Zi-yi，born in 1997，postgraduate. Her main research interests include computer vision and pattern recognition.



WANG Rui-ping，born in 1981，Ph. D，professor，Ph. D supervisor，is a senior member of China Computer Federation. His main research interests include computer vision and pattern recognition.