视频人脸识别中高效分解卷积与时间金字塔网络研究

周书田，颜 信，谢镇汕*

(电子科技大学格拉斯哥学院 成都 611731)

【摘要】随着大量视频监控和摄像头网络的架设，非受限场景下的连续视频帧人脸识别愈发引人关注。传统的连续视频帧人脸识别方法大多存在识别结果易波动和计算资源消耗密集的问题。因此，该文对比了不同的帧间汇聚方式，采用注意力机制优化帧间汇聚过程，并采用 3D 分离卷积进行视频人脸建模，有效降低了视频人脸识别的计算消耗，提高了识别准确率。此外，提出了一种时间金字塔网络，可以进一步有效挖掘帧间互补信息，以提高识别准确率。该方法的有效性在 YTF、PaSC 数据集上得到了验证。

关 键 词 卷积神经网络； 分解卷积； 人脸识别； 时间金字塔网络； 视频分析

中图分类号 TP394.1 **文献标志码** A doi:10.12178/1001-0548.2020319

Efficient Decomposition Convolution and Temporal Pyramid Network for Video Face Recognition

ZHOU Shu-tian, YAN Xin, and XIE Zhen-shan*

(Glasgow Collge, University of Electronic Science and Technology of China Chengdu 61173)

Abstract With a large number of video surveillance and camera networks, face recognition of continuous video frames in unrestricted scenes is becoming more and more attractive. Most of the traditional face recognition methods for continuous video frames have the problem of fluctuating recognition results and intensive computing resources. In this paper, an efficient 3D decomposition convolution is designed, which can effectively reduce the computational consumption of video face recognition and improve the recognition accuracy. Finally, we also propose a temporal pyramid network to further effectively mine complementary information between frames to improve the recognition accuracy. The performance has been tested on YTF and PaSC datasets.

Key words convolutional neural network; decomposition convolution; face recognition; temporal pyramid network; video analysis

随着越来越多的摄像采集与监控设备被部署，视频中人脸识别的需求大幅上升。这些系统迫切地需要可靠与准确的人脸识别。与静态配合式人脸识别[1-3]不同的是，视频中的人脸识别往往是非受限的，即捕捉姿势、图像质量等在帧间都会呈现巨大的变化。一方面，帧间信息可以形成信息互补，并通过多帧对人脸进行更加准确的识别；另一方面，帧间部分低质量，如运动模糊、极端采集角度及低分辨率的图像，又会干扰视频人脸识别的结果。如果直接使用静态人脸识别方法，那么这些低质量帧将会带来误识别。因此如何对帧之间的信息进行有效汇聚，从而形成更加鲁棒的视频人脸特征表达，成为一个关键问题。

目前对视频中进行人脸识别最流行的方法是将视频人脸图像帧表示为无序的特征向量[4-8]，并把各帧的特征向量进行汇聚成为视频级特征。验证时，对视频级特征进行相似性检索即可。常见汇聚的方式有平均汇聚[6]、最大池化汇聚与注意力机制汇聚[7]。但这些方式需要对所有视频帧进行特征提取，消耗了大量的计算资源，并不高效。因此近来，视频识别领域提出了如 3D 卷积[8]等新颖的视频分析框架，可以有效地对连续帧信息进行捕捉，但 3D 卷积同样会引入巨大的计算量。

本文首先比较了基于 2D 卷积网络与不同帧间

收稿日期：2020－05－30；修回日期：2020－08－29
作者简介：周书田 (1998－)，男，主要从事计算机视觉方面的研究.
通信作者：谢镇汕，E-mail：ivanxie1022@gmail.com

汇聚的方法，并介绍了一种在视频中使用 3D 分解卷积的连续帧人脸识别方法。该方法不需要逐帧地对人脸数据进行提取，而是将多帧输入一个 3D 分解卷积结构，得出一个全局的特征向量。与逐帧进行特征提取并汇聚的算法相比，该方法可以大幅提高计算效率，且保持了竞争力的识别精度。最后，本文提出了用于视频人脸识别的时间金字塔网络，可以对帧间互补信息进行有效建模。3D 分解卷积与时间金字塔网络的有效性在 YouTubeFace[9]、PaSC 测试集得到了验证。

1 视频人脸识别

本节介绍视频人脸识别的整个流程及各项的详细配置。视频人脸识别系统可以分为 3 部分：视频特征编码器、优化视频编码器的损失函数及将视频进行匹配与检索的查找方法。首先将视频切成连续且非重叠的视频片段$\{c_k\}$，每个片段包含有T帧，对每个片段进行特征抽取。片段特征抽取器将片段作为输入，并且输出D维度的特征向量$\boldsymbol{f}_c$。视频总体的特征为所有视频片段特征的平均汇聚。

在对视频间特征进行比对时，本文采用余弦相似度，并且设定阈值，当阈值大于一定值时，认为两段视频中的人为同一人，亦或是视频中的人脸与底库当中的相匹配。视频特征向量$\boldsymbol{x}_i$与$\boldsymbol{y}_i$的余弦相似度为：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(\boldsymbol{x}_i\times\boldsymbol{y}_i)}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{x}_i)^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{y}_i)^2}}$$

判定是否为同一人的阈值，可根据应用场景进行设定。在评价性能时，往往采用一定 FPR (false positives rates) 下的 TPR(true positives) 进行比较。

对于视频片段编码器，采用 2D 卷积网络+特征聚合方法或者 3D 卷积网络直接将视频剪辑编码为特征向量$\boldsymbol{f}_c$。2D 卷积网络则首先提取每帧的图像特征$\{\boldsymbol{f}_c^t\},t\in[1,n]$，并且通过特征聚合方法，将其聚合为单个视频特征向量$\boldsymbol{f}_c$。

接下来，本文分别介绍基于 2D 卷积神经网络+特征聚合以及 3D 卷积神经网络进行视频人脸识别的方法，并提出一种高效分解卷积，以在保持识别精度的同时，降低计算消耗。

2 基于帧间特征汇聚方法

2.1 特征提取

本文采用标准的 ResNet-50 网络[10] 作为特征提取模型，对一个给定的视频帧序列$\{\boldsymbol{f}_c^t\},t\in[1,n]$，将每帧进行特征提取，得到一个维度的特征$T\times D$向量矩阵，其中$T$是视频帧的数量，$D$为每个视频帧的特征向量维度。

2.2 池化汇聚

对视频帧的特性进行池化，以获取视频层特征。池化汇聚主要包括最大池化与平均池化。平均池化和最大池化分别为：$\frac{1}{T}\sum_{t=1}^{n}\boldsymbol{f}_c^t$、$\max_t(\boldsymbol{f}_c^t)$。

2.3 注意力机制汇聚

另外一种汇聚的方式是使用注意力机制。注意力机制旨在自适应地在帧特征之间寻找权重，以给“关键帧”更高的汇聚权重[11]。基于注意力机制的帧间汇聚表示为：

$$\boldsymbol{f}_c=\frac{1}{T}\sum_{t=1}^{n}a_c^t\boldsymbol{f}_c^t$$

式中，a_c^t为T帧的注意力权重。在卷积神经网络的最后一层，得到$[w,h,2\,048]$的特征图，采用宽为w、长为h、输入通道数为 2 048、输出通道为d的卷积核对特征图进行卷积；然后使用输入通道数为 2 048、输出通道数为 1 的全连接层将特征映射为注意力权重$s_c^t,t\in[1,T]$；最后将所有帧间的注意力权重过 softmax 层得到最终的注意力分数：

$$a_c^t=\frac{e^{s_c^t}}{\sum_{i=1}^{T}e^{s_c^i}}$$

3 基于时序卷积的方法

3.1 3D 卷积

对于视频连续帧，本文直接将其输入 3D 卷积神经网络——ResNet-50 网络。将包含有n帧的视频片段c进行卷积生成为$\boldsymbol{f}_c$。相比于正常的 2D 卷积，3D 卷积在时间维度上多了一维度，正常的 2D 卷积核可以表示为$[c,h,w]$，而 3D 卷积核则可表示为$[c,t,h,w]$。因为时间通道的加入，可以建模时序帧间的特征信息。

3.2 时序分解卷积

采用 3D 卷积可以建模帧间信息，但 3D 卷积会引入巨大的计算量与显存消耗，给部署带来巨大

挑战。文献 [12] 对 3D 卷积分解进行了一定的探索。如图 1 所示，忽略掉通道维度，一个$t\times d\times d$的 3D 卷积，可以分解为一个$1\times d\times d$的卷积再加上一个$t\times 1\times 1$的卷积，使计算量减小，并且分解后的卷积本质上为 2D 卷积加 1D 卷积，可以使用工业界更加成熟的卷积优化算子进行加速。

3D 分解卷积其本质是先对单帧的空间信息进行建模 (2D 卷积)，后对帧间信息进行卷积汇聚 (1D 卷积)。本文将 ResNet-50 网络中所有的 3D 卷积都替换成为此分解卷积，实验证明，其实现了更快的推理速度与有竞争力的准确度。

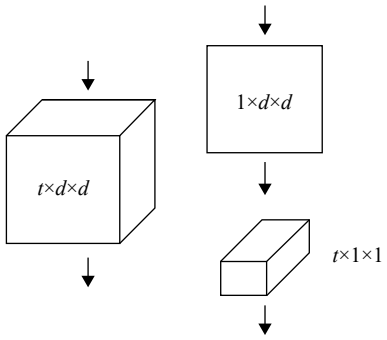

图 1　将卷积核为 $t\times d\times d$ 的 3D 卷积分解为 $1\times d\times d$ 与 $t\times 1\times 1$ 的卷积

3.3　时间金字塔网络

对于视频人脸识别，核心关键点是如何对视频帧间的信息进行建模，如何高效地利用帧间的互补信息。使用 3D 卷积分离对连续帧进行提取，并没考虑到视频速度的变化，即一个人在视频中是缓慢的摇晃头部，还是快速地摇晃头部，这种速度的变化并没有被建模到网络中。为了对多种速度的脸部运动进行建模，本文以不同的时间帧率对视频进行采样，并将不同帧率的输入以时间金字塔的方式进行汇聚。

本文分别对输入视频以不同的速率进行采样，如图 2 所示。假设采取M个不同的采样速率$\{r_1, r_2,\cdots,r_M, r_1<r_2<\cdots<r_M\}$，将对不同采样速率得到的视频帧率送入不同分支的 3D 分流卷积中进行特征提取，得到特征序列$\{\boldsymbol{F}^{(1)}_{\text{base}},\boldsymbol{F}^{(2)}_{\text{base}},\cdots,\boldsymbol{F}^{(M)}_{\text{base}}\}$，其中特征序列的维度分别为$\left\{C\times\frac{T}{r_1}\times W\times H, C\times\frac{T}{r_2}\times W\times H,\cdots, C\times\frac{T}{r_m}\times W\times H\right\}$。

不同分支的网络会经过一个时间金字塔结构，由时间分辨率低的分支进行时序上连续线性插值$(\boldsymbol{I})$，与高时间分辨率分支进行融合，融合过程可以表示为：

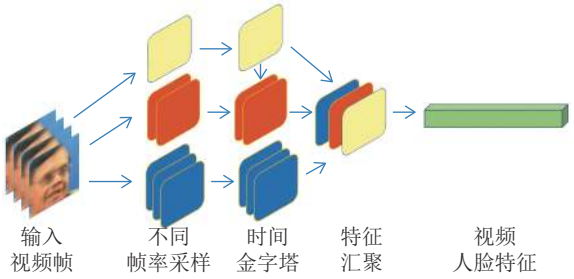

图 2　用于视频人脸识别的时间金字塔网络

$$\boldsymbol{F}^{t}_{\text{base}}=\text{concat}(\boldsymbol{F}^{t}_{\text{base}},\{\boldsymbol{I}(\boldsymbol{F}^{(1)}_{\text{base}}),\boldsymbol{I}(\boldsymbol{F}^{(2)}_{\text{base}}),\cdots,\boldsymbol{I}(\boldsymbol{F}^{(t-1)}_{\text{base}})\})$$

然后，各时间分辨率网络将会进行特征汇聚，最终由最大池化 (max-pooling) 形成视频人脸特征。时间金字塔网络综合了不同时间分辨率信息，从而可以捕捉更丰富的帧间信息。

4　实 验

4.1　损失函数

对于单帧与基于时序卷积的网络的训练，本文采用了三元组损失与 softmax 交叉熵损失相结合的方式。三元组损失旨在将负样本对的距离拉远，而保持正样本对的距离靠近。softmax cross-entropy 损失则为经典的分类损失函数。在两个 loss 的监督下，可以使卷积神经网络形成健壮的特征表达。

4.2　实验数据

本文采用 UMDFace[13] 作为训练集，其包含了 3 107 476 个带注释的视频帧，这些视频帧来自于包含 3 107 个人物的 22 075 个视频。数据集预先包含了估计的人脸姿势和 21 个关键点位置及性别等信息。本文采用 RetinaFace 对视频帧的人脸进行预处理，采用仿射变换进行剪裁，所有的人脸都被裁切到 224×224。

本文在 YouTubeFace[9] 数据集上对提出的方法与基线方法进行了测试。YouTubeFace 数据集包含来自 1 595 个体的 3 424 个视频，平均每个人 2.15 个视频，是现在广泛采用的视频人脸识别测试数据集。视频的长度从 48 帧到 6 070 帧不等，平均为 181.3 帧。本文采用了标准的测试流程对 6 000 对给定的正负样本对进行预测，当预测结果与样本对标签结果一致时，计为正确。汇报结果为在 6 000 对样本的预测准确率。

为了验证该方法的泛化能力，本文还在 PaSC[14] 数据集上进行了实验。

4.3 实验参数

本文采用 SGD 作为网络训练的优化器，初始学习率被设定为 0.01，且随着损失的饱和而乘以 0.1。对于 2D 卷积网络，网络训练的批大小被设定为 32。对于 3D 卷积网络及 3D 分离卷积网络，批大小设定为 8。在 8 块 NVIDIA 2080TI GPU 进行了实验。

4.4 实验结果

在 YoutubeFace 数据集上对分解卷积以及各项基线方法进行了验证，其结果如表 1 所示。

表 1 不同方法在 YoutubeFace(YTF) 上性能的比较

方法	YTF准确率/%
文献[15]	84.8
文献[16]	94.5
文献[17]	94.3
2D卷积+最大池化	95.2
2D卷积+平均池化	95.0
2D卷积+注意力机制	96.1
3D卷积	94.2
3D分解卷积	95.7
3D分解卷积+时间金字塔	96.5

可以看到，在性能比较上，直接利用 3D 卷积实现了最差的性能，而利用改进的分解卷积，可以使得测试准确率提高 1.5%。在预先采用 2D 卷积网络提取帧间特征并使用不同汇聚方法进行汇聚的比较中，注意力机制取得到了最好的结果，达到 96.1%。最大池化相比于平均池化提高了 0.2% 的性能。对 3D 卷积进行高效分解后，其实现了与 2D 卷积网络+帧间汇聚相竞争力的结果。进一步将 3D 分解卷积与时间金字塔网络进行结合，达到了最佳的性能，为 96.5%。

在 PaSC 数据集上验证了本文方法的泛化性，如表 2 所示。可以看到，采用 3D 分离卷积+时间金字塔结构在 PaSC 数据集上取得了最优成绩，比常用的 2D 卷积+平均池化提高了 4.51% 的性能

本文比较了不同方法计算消耗的大小，以包含 32 帧的视频片段为例，对于 2D 卷积网络，其单模型输入为 224×224，对于 3D 卷积网络，其单模型输入为 32×224×224。如表 3 所示，3D 卷积网络方法与 3D 分离卷积方法显著减小了推理阶段的计算消耗。时序金字塔网络在提高精确的同时，相比于 2D 方法，还减少了 50% 的算力消耗，这有利于工业界的大规模应用。

表 2 不同方法在 PaSC 上性能的比较

方法	PaSC准确率/%
文献[18]	68.76
文献[19]	80.12
文献[20]	80.52
2D卷积+最大池化	88.70
2D卷积+平均池化	91.81
2D卷积+注意力机制	95.72
3D卷积	94.72
3D分解卷积	96.01
3D分解卷积+时间金字塔	96.32

表 3 不同方法在计算消耗比较

方法	Madd
2D卷积+最大池化	263.04
2D卷积+平均池化	263.04
2D卷积+注意力机制	267.87
3D卷积	81.01
3D分解卷积	76.92
3D分解卷积+时间金字塔	137.29

5 结 束 语

本文针对视频中人脸识别问题，比较了 2D 卷积网络特征提取+帧间特征汇聚 (平均池化、最大池化与注意力机制) 与 3D 卷积的方法，采用 3D 分解卷积提高了视频人脸识别的准确率与效率，并提出了一种时间金字塔网络，实现了更高精度的视频人脸识别。在未来的工作中，将会考虑如组卷积及 Depth-wise 卷积等更加轻量化的 3D 卷积实现方式，以实现更加高效的视频人脸识别。

参 考 文 献

[1] DENG J, GUO J, XUE N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 4690-4699.

[2] WANG H, WANG Y, ZHOU Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah: IEEE, 2018: 5265-5274.

[3] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, Massachusetts: IEEE, 2015: 815-823.

[4] 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. Las Vegas, Nevada: IEEE, 2016: 59-67.

[5] 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. Honolulu, Hawaii: IEEE, 2017: 3781-3790.
[6] DING C, TAO D. Trunk-branch ensemble convolutional neural networks for video-based face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 1002-1014.
[7] 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. Honolulu, Hawaii: IEEE, 2017: 4362-4371.
[8] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.
[9] WOLF L, HASSNER T, MAOZ I. Face recognition in unconstrained videos with matched background similarity[C]//CVPR 2011. Colorado, USA: IEEE, 2011: 529-534.
[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada: IEEE, 2016: 770-778.
[11] YAO H, ZHANG S, HONG R, et al. Deep representation learning with part loss for person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2860-2871.
[12] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah: IEEE, 2018: 6450-6459.
[13] BANSAL 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(IJCB). Glasgow, UK: IEEE, 2017: 464-473.
[14] BEVERIDGE J R, PHILLIPS P J, BOLME D S, et al. The challenge of face recognition from digital point-and-shoot cameras[C]//IEEE Sixth International Conference on Biometrics: Theory. New York: IEEE, 2013: 872-891.
[15] LI H, HUA G, SHEN X, et al. Eigen-pep for video face recognition[C]//Asian Conference on Computer Vision. Cham: Springer, 2014: 17-33.
[16] TAIGMAN Y, YANG M, RANZATO M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio: IEEE, 2014: 1701-1708.
[17] 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. Honolulu, Hawaii: IEEE, 2017: 3781-3790.
[18] MOHAGHEGHIAN E. An application of evolutionary algorithms for WAG optimisation in the Norne Field[D]. St. John: Memorial University of Newfoundland, 2016.
[19] HUANG Z, VAN GOOL L. A riemannian network for spd matrix learning[C]//The 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 362-371.
[20] HUANG Z, WU J, VAN GOOL L. Building deep networks on grassmann manifolds[C]//The 32nd AAAI Conference on Artificial Intelligence. New Orleans, Louisiana: AAAI, 2018: 2517-2525.

编　辑　蒋　晓