文章编号：1003-0530(2022)06-1202-11

具有遮挡鲁棒性的监控视频人脸再识别算法

张 博[1,2] 赵 巍[1,2] 段鹏松[1,2] 武 琦[1,2]
（1. 郑州大学网络空间安全学院，河南郑州 450002；2. 郑州大学汉威物联网研究院，河南郑州 450002）

摘 要： 传统身份识别技术需要将待识别人员信息预先录入，同时未考虑识别过程中的遮挡问题，不能满足公共场所基于监控视频的再识别需求。现有行人再识别算法多依赖于服饰等外观特征，难以进行长期追踪与再识别。针对以上问题，本文提出了一种对遮挡具有鲁棒性的人脸再识别算法。首先，对监控视频中的人脸进行检测与对齐，并判断人脸中存在的遮挡位置；其次，根据遮挡位置查找掩码字典并选择对应掩码，再用掩码排除遮挡元素；最后，使用注意力机制对多帧图片分配权重以更新特征，再使用分区域匹配方法得到识别结果。为验证该方法的有效性，本文分别在COX数据集和人工合成遮挡的数据集上对所提方法进行了测试。其中，在COX数据集上的rank-1准确率为95.2%，在合成遮挡的数据集上rank-1准确率为73.0%，相比现有方法有明显优势。

关键词： 深度学习；人脸再识别；注意力机制
中图分类号： TP183 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2022.06.007

引用格式： 张博，赵巍，段鹏松，等. 具有遮挡鲁棒性的监控视频人脸再识别算法[J]. 信号处理，2022，38(6)：1202-1212. DOI：10.16798/j.issn.1003-0530.2022.06.007.
Reference format： ZHANG Bo，ZHAO Wei，DUAN Pengsong，et al. Surveillance video re-identification with robustness to occlusion[J]. Journal of Signal Processing，2022，38(6)：1202-1212. DOI: 10.16798/j.issn.1003-0530.2022.06.007.

Surveillance Video Re-Identification with Robustness to Occlusion

ZHANG Bo[1,2] ZHAO Wei[1,2] DUAN Pengsong[1,2] WU Qi[1,2]
（1. Zhengzhou University School of Cyber Science and Engineering，Zhengzhou，Henan 450002，China；
2. Zhengzhou University & Hanwei IoT Research Institute，Zhengzhou，Henan 450002，China）

Abstract: Traditional identification technologies require pre-recorded information from target personals，while failing to consider any visual obstructions in the identification process，resulting in its unsatisfactory performance in surveillance-video-based re-identification scenarios，especially for public spaces. Most existing person re-identification approaches examine appearance features such as clothing and decoration，which are prone to change in time and space，and thus are unreliable for long-term tracking. An effective and reliable approach for long-term re-identification is to utilize stable biometric features such as facial features. However，with occlusion，low resolution，lack of illumination，and perspective gestures exhibited in surveillance videos，traditional facial recognition methods that are excellent for image recognition cannot perform well. To address these issues，this paper proposed a deep-learning-based face re-identification algorithm. The algorithm combined an attention mechanism with a mask dictionary to dynamically and appropriately assign weights to video frame features，thereby reducing the effect of occlusion and effectively improving the re-identification accuracy. Extensive experiments demonstrated that the proposed method was able to achieve a rank-1 accuracy of up to 95.2% on the cox dataset，and 73.0% on the same dataset with synthetic occlusion. These results comfirm the superior performance of the proposed algorithm compared to state-of-the-art re-identification algorithms.

Key words: deep learning；face re-identification；attention mechanism

收稿日期：2022-01-13；修回日期：2022-04-14
基金项目：国家自然科学基金面上项目(61972092)；郑州市协同创新重大专项(20XTZX06013)；河南省高等学校重点科研项目计划(21A520043)

1　引言

随着监控系统应用场景数量和种类的快速攀升，利用多个摄像机对相同个体进行跨时空身份识别的需求愈发强烈，并在公共安全、执法等领域展现出巨大潜力[1]。现有身份识别技术一般依赖于外观特征（如衣服、装饰等）或生物特征（如面容、步态等）。外观特征的波动性较大，导致基于其的身份识别技术稳定性较差[2]。生物特征的复杂性较高，但稳定性更好，利用其进行身份识别更为可靠[3]。在生物特征中，面部特征作为便捷的非侵入性视觉特征，可以更有效的进行人体跟踪与识别[4]。

近些年来，随着基准数据库的增大、高级网络结构[5-6]和各种损失函数[7-8]的广泛使用，基于深度学习的人脸识别技术取得了显著的进步，在某些基准数据库上的识别能力已经超越了人类。虽然基于深度学习的识别模型在无限制的人脸识别场景下取得了巨大的成功，但是仍然无法满足监控视频的人脸再识别任务。主要原因是监控摄像机捕获的面部图像存在分辨率和角度的差异、信息冗余以及面部遮挡等问题。此外，监控摄像机拍摄的目标对象往往没有在数据库中记录其完整特征，因此需在首次发现目标时快速提取有效特征，这给传统面部识别技术带来巨大挑战。

为解决上述问题，需对现有面部识别方法进行改进，以满足在监控视频下再识别的现实需求。首先，由于监控视频帧中目标面部存在姿势、表情、光照和遮挡等诸多不同，相同目标面部在不同帧中的特征存在较大差异。因此，可以使用注意力机制对质量好的图片分配更多权重，得到更完善的面部特征，以解决各帧图像分辨率和角度存在差异的问题。其次，监控视频相邻帧之间的面部图像往往非常接近，存在较大的信息冗余，直接使用连续的帧提取特征会导致识别方法效率低下。本文采取等距随机取样的方法选取合适的视频帧进行特征提取，不仅保留视频整体的面部特征，还减少了数据冗余，提升了算法的整体效率。最后，对于监控视频中存在的面部遮挡问题，本文使用了 PDSN 网络[9]来训练人脸各区块遮挡的掩码来弱化遮挡对特征的影响，并通过分区域匹配的方法减少识别误差。本文的主要贡献总结如下：

通过注意力机制与掩码字典的联合使用，先将视频帧中受遮挡影响的特征元素舍弃，再对剩余的特征动态分配权重，降低了监控视频下人脸遮挡对再识别的影响。

针对掩码字典在再识别场景下准确度下降的问题，提出了分区域匹配的方法，降低了掩码字典的误差，提高了再识别的准确度。实验结果表明，本文的方法在 COX 监控视频数据上 rank-1 准确度达到了 95.2%，并在合成面部遮挡的监控视频数据上 rank-1 准确度达到了 73.0%。

本文章节安排如下：第 1 节为绪论。第 2 节分析了目前人脸识别及再识别的研究现状。第 3 节介绍了本文算法的技术细节。第 4 节展示并分析了本文算法的实验结果。第 5 节是本文总结和未来展望。

2　研究现状

2.1　基于图像的人脸识别

早期的人脸识别方法没有足够的数据来进行强大的模型训练，也没有可靠的测试基准，集中应用在小规模的受限场景。直 LFW[10]数据集的出现，研究人员开始转向无限制的人脸识别。随着 CASIA[11]、CelebFaces[12]、MS-Celeb-1M[13]、MegaFace[14]等数据集的创建，人脸识别技术得到了快速的发展，如 SCHROFF 等人的研究[7]在 LFW 基准上的识别准确率超越了人类。

传统人脸识别算法在较为清晰的图像数据集上已经取得了很大成功。然而监控视频数据集展现出光照、倾角以及遮挡等不利因素，使得这些算法难以获得令人满意的效果。鉴于此，一些研究人员开始考虑通过对特征进行选择来去除多余嘈杂的特征，保留对识别有用的特征，如 LI 等人在[15]提出了一种半监督的局部特征选择方法，通过学习每类特征的重要性来筛选出针对不同类别的特征子集，以此来选择最具鉴别力的特征。

一些文献针对面部遮挡问题进行了研究。WAN 等人[16]提出了在 CNN 模型的中间层增加一个 MaskNet 分支，为被遮挡的面部区域激活的隐藏单元分配较低的权重。Trigueros 等人[17]通过用合成遮挡的人脸图片来增加训练数据以此解决遮挡问题。YU 等人[18]使用 SIFT 和 SVM 算法来进行遮挡

的面部识别，它将图像划分为四个局部区域，使用加权平均方法来确定最终的分类结果。DUAN等人[19]提出了用GAN来生成无遮挡的正面人脸来降低遮挡的影响。

这些方法虽然对面部的遮挡有着不同程度的鲁棒性，但需要一个较好的正面图像作为基准用于识别，并且无法利用视频中的时间信息，无法完成监控视频下的人脸再识别任务。

2.2 监控视频下的人脸再识别

目前，虽然基于监控视频的人脸再识别研究处于起步阶段，但也出现一些研究成果。DANTCHEVA等人[20]通过结合头发、皮肤和装饰等特征，进行视频监控系统中正面到侧面的人脸匹配。FARINELLA等人[21]对人脸进行预处理，去除几何和光度的变化，并表示为三元模式的空间直方图，以此进行人脸的再识别。QIU等人[22]构建了一个领域自适应字典来处理两张人脸图像的匹配。LI等人[23]探讨了面部信息在人员再识别中的作用，证明面部是一种更可靠的生物识别特征，可以作为长期跟踪目标的依据。LI等人[24]构建了一个人脸再识别数据集，并采用改进的DNN架构和区块匹配技术，并使用完全卷积结构和空间金字塔池化(SPP)来进一步提高性能。WANG等人[25]则采用了深度模型进行特征学习和聚类来识别身份。WANG等人[26]针对实际监控场景中经常遇到的人脸图像分辨率较低的问题，提出了一种利用松弛耦合非负矩阵分解的低分辨率人脸识别算法。CHENG等人[4]为了解决真实监控视频场景下的再识别问题，制作了一个大规模的人脸再识别的数据集，并对监控面部图像固有的低分辨率问题进行了研究。

整体来说，对监控视频下人脸再识别的研究仍然存在诸多不足。虽然这些方法尝试解决监控摄像头下的人脸再识别问题，但它们未考虑监控视频中容易出现的遮挡问题，尤其无法满足当前疫情防控中的人脸再识别需求。

2.3 注意力机制

注意力机制已经成为深度学习领域一个重要的概念，被广泛应用于不同领域。LI等人[27]通过空间注意力解决了图像之间的对齐问题，有效解决了特征被遮挡区域破坏的问题，并使神经网络关注更有鉴别力的物体特征。LI等人[28]提出了一个自注意力模型，通过探索像素和类别之间的相关性来建立全局空间依赖性模型，以在保证性能的同时降低计算复杂性。但是，空间注意力机制主要关注单一图像上的特定信息，不能完成视频中连续图像的信息捕获。在本文中，我们将使用时序注意力模型，为信息更丰富的视频帧分配更高的权重以提高识别准确率。

3 具有遮挡鲁棒性的人脸再识别算法

人脸再识别需要在监控摄像头首次拍摄到一个的人面部时，迅速记录下其有效特征，之后可在多个摄像机下跨空间与时间进行匹配。根据文献[29]，再识别任务可以分为两类：开集再识别问题和闭集再识别问题。这两类问题的区别主要在于画廊集的不同，画廊集(gallery set)是被查询的集合，而探针集(probe set)是查询集合。每个需要识别的对象都是一个探针，需要在画廊集中检索出与其最相似的目标。开集再识别问题没有预先固定的画廊集，画廊集随着时间变化。闭集再识别问题画廊集的大小是固定的，是一个一对多的匹配问题。

本文提出了一个具有遮挡鲁棒性的人脸再识别算法，其整体结构如图1所示。首先，需要对探针视频中的人脸进行检测与对齐，并通过选取合适的视频帧来进行特征提取；其次，通过特征提取网络提取不受遮挡影响的特征元素；最后，根据所提取的特征在画廊集中进行匹配以实现再识别。如果匹配失败，则将探针扩充入画廊集。

3.1 预处理

在预处理阶段，我们把所有的视频帧图片经过RetinaFace网络[30]进行检测。在检测出人脸框与5个面部关键点后，使用仿射变换对视频帧中人脸进行对齐并调整为固定大小。通过这种方式，人脸图片的五个关键点会出现在图片的固定位置上。同时，由于监控视频相邻帧之间的人脸图片十分相似，存在信息的冗余。为了能够充分利用整个视频的视觉信息，避免连续视频帧的特征冗余，本文采取了一个等距随机取样方法：对于一个输入视频，我们把它分成T个时间相等的片段，并从每个片段中随机抽取一帧图像。后续的操作将对抽取的T帧

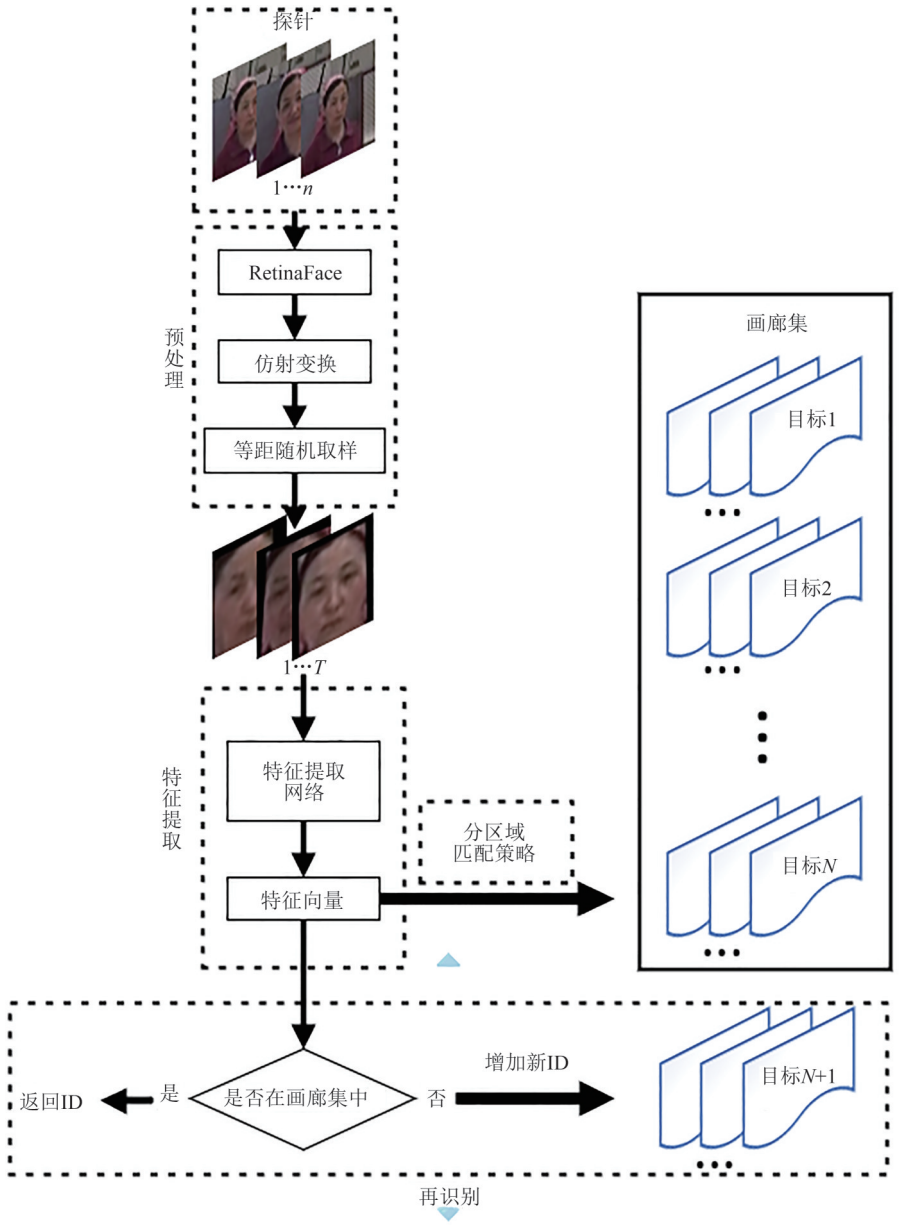


图 1　基于人脸特征的再识别算法流程图

Fig. 1　Flow chart of face feature-based re-identification algorithm

图片进行特征提取，以代表整段视频的特征。

3.2　特征提取

本文设计的特征提取网络由四部分构成，分别为 PSPNet[31]、掩码字典、主干网络和时间注意力机制，如图 2 所示。首先，对预处理后的视频图片使用 PSPNet 判断遮挡的区块集合；其次，按照遮挡区块集合从掩码字典中选取掩码，并将掩码和主干网络提取的特征图相乘；最后，通过注意力机制对各帧提取的特征图分配权重，并通过 FC 层得到最终的特征向量。

3.2.1　遮挡位置检测

本文使用了 PSPNet 语义分割模型对面部的遮挡区域进行分割。在对面部遮挡数据集 MAFA[32]的图片进行处理和标注后，我们对 PSPNet 进行了训练，训练效果如图 3 所示。

分割出遮挡区域后，把人脸图片划分为 5×5 个等大小的区块，以使眼睛、鼻子和嘴巴等面部器官可以被某一区块所覆盖。之后，通过计算遮挡的面

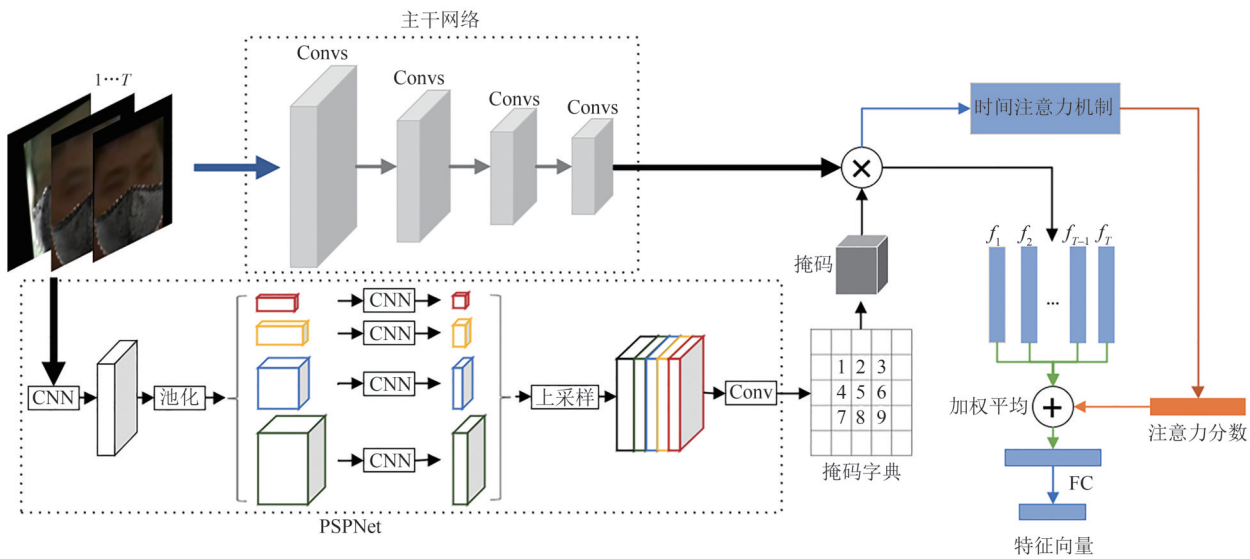


图2 特征提取网络

Fig. 2 Feature extraction net

图3 PSPNet语义分割效果

Fig. 3 PSPNet semantic segmentation effect

部区域与各个区块的交并比来确定哪些区块存在遮挡。当该交并比大于预设阈值时,则判定该区块存在遮挡。

3.2.2 主干网络

本文使用改良的Resnet50模型[33]作为主干网络提取图片的特征,并使用大边缘余弦损失函数[34]在CASIA-WebFace[11]数据集上进行训练,在LFW[10]测试基准上的准确率达到了99.0%。

3.2.3 掩码训练

对于面部的遮挡问题,最直接的方法是用受遮挡影响较小的面部特征进行比较,以降低遮挡物体对特征的影响。而PDSN网络[9]可以学习遮挡的面部区块和被破坏的图像特征之间的关系,能够准确地定位损坏的特征元素,其结构如图4所示。

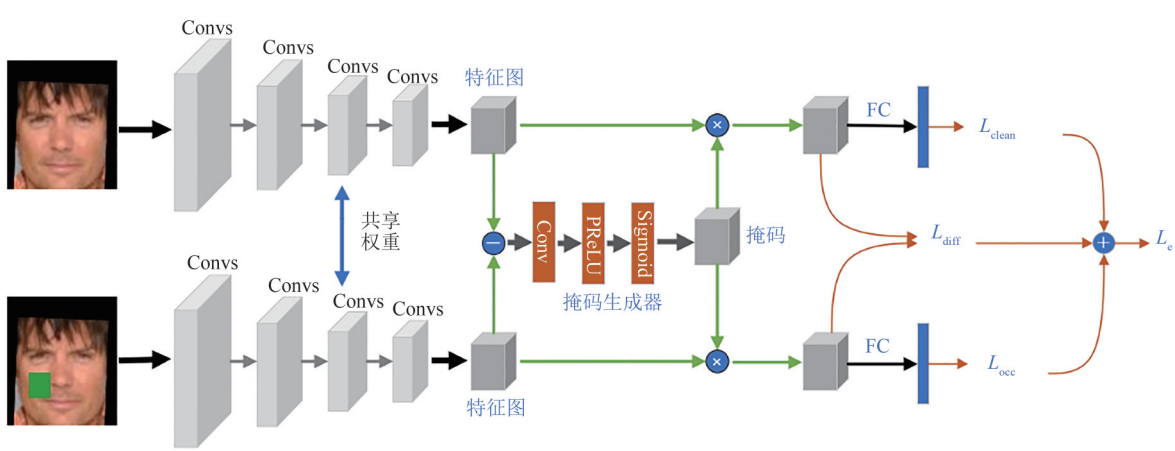


图4 PDSN网络结构图

Fig. 4 PDSN network structure diagram

PDSN 网络由主干卷积神经网络和掩码生成器分支组成,主干网络负责提取成对的人脸图片特征,掩码生成器则生成一个掩码并与主干网络提取的特征图相乘,从而降低遮挡对特征图的影响。该网络的训练数据为成对图片,分别为无遮挡的原图以及该图某区块被遮挡的副本。它的损失 L_θ 由 L_{occ}、L_{clean}、L_{diff} 组合而成,如公式(1)所示。

$$L_\theta = \sum_i L_{\text{occ}}(\theta;\ \tilde{f}\left(x_j^i\right), y^i) + L_{\text{clean}}\left(\theta;\ \tilde{f}\left(x^i\right), y^i\right) + \lambda L_{\text{diff}}\left(\theta;\ \tilde{f}\left(x_j^i\right), \tilde{f}\left(x^i\right)\right) \tag{1}$$

其中,$\tilde{f}(.) = \mu_\theta(.) f(.)$,$f$ 为卷积层输出的特征图,μ_θ 表示输出的掩码。x^i 表示第 i 对人脸图像,x_j^i 表示第 i 对第 j 区块被遮挡的人脸图像。分类损失 L_{occ} 与 L_{clean} 分别为有遮挡图片的分类损失和无遮挡图片的分类损失,其目的是使图像的特征图乘以掩码后仍能被分类器正确分类。相对损失 L_{diff} 则使掩码生成器关注由于遮挡而偏离其真实值的特征元素。三个函数的定义分别如公式(2)、(3)、(4)所示。

$$L_{\text{clean}}\left(\theta;\ \tilde{f}\left(x^i\right), y^i\right) = -\log\left(p_{y^i}\left(F\left(\tilde{f}\left(x^i\right)\right)\right)\right) \tag{2}$$

$$L_{\text{occ}}(\theta;\ \tilde{f}\left(x_j^i\right), y^i) = -\log\left(p_{y^i}\left(F\left(\tilde{f}\left(x_j^i\right)\right)\right)\right) \tag{3}$$

$$L_{\text{diff}}\left(\theta;\ \tilde{f}\left(x_j^i\right), \tilde{f}\left(x^i\right)\right) = \left\| \mu_{\theta(.)} f(x^i) - \mu_{\theta(.)} f(x_j^i) \right\|_1 \tag{4}$$

其中,F 表示全连接层的输出,$\mu_\theta(.) = \mu_\theta\left(\left| f\left(x_j^i\right) - f\left(x^i\right)\right|\right)$,$\|.\|_1$ 为 L1 范式。

训练完成后,从每个遮挡小块的掩码生成器中提取一个固定的掩码,并进行二值化的操作,用来抛弃受遮挡严重影响的元素。二值化操作如公式(5)所示,其中 m 为求均值后的掩码,$\left\{\tilde{m}[1], \cdots, \tilde{m}\left[\mu \times K\right]\right\}$ 表示 m 中 $\mu \times K$ 个最小元素,μ 为丢弃阈值(本文中设置为 0.25),$K = C \times W \times H$,为特征图中元素总量。

$$M[k] = \begin{cases} 0 & \text{if } \tilde{m}[k] \in \left\{\tilde{m}[1], \cdots, \tilde{m}\left[\mu \times K\right]\right\} \\ 1 & \text{else} \end{cases} \tag{5}$$

完成二值化的操作后,将各个区块的掩码构建一个字典。当两个人脸图像进行匹配时,对 PSPNet 检测出的遮挡区块进行字典匹配和掩码操作(遮挡区块求并集),即可去除相应区块所影响的特征元素。另外,多个区块遮挡时只需同时乘以多个区块对应的掩码。

3.2.4　时序注意力机制

由于监控视频不同帧的面部图像存在姿势、表情、光照和遮挡的差异,因此视频帧之间可提取的特征不尽相同。因此,应该考虑为不同视频帧的图片分配不同的权重。然而基本的时间聚合技术,如平均池化或最大池化,通常会削弱或过度强调有代表性特征的贡献[27]。

与基本的时间聚合技术不同,注意力机制可以轻松建立长时间的依赖关系,因此被广泛地应用于计算机视觉中[28]。本文使用改进的时间注意力机制,判断不同帧特征重要性并赋予相应特征权重值。注意力机制把乘以掩码后的特征作为输入,输出 T 个注意力分数,并与 T 帧图像的特征计算加权平均。如公式(6)所示,其中 $t \in [1, T]$,a^t 为对应的注意力分数,f^t 为对应的视频帧特征。

$$f = \frac{1}{T} \sum_{t=1}^{n} a^t f^t \tag{6}$$

在本文的实验中,我们将时间注意机制与多帧特征向量求均值、平均池化和最大池化的方法进行了比较,发现时间注意模型可以得到最高的准确度。

3.3　分区域匹配

由于我们对掩码执行了二值化操作,因此掩码与特征图相乘后,受遮挡影响严重的元素将会被设置为零。当两个目标使用相同的掩码时,它们特征图中对应位置的零元素会增加,导致二者特征向量相似度增大。对于一个无遮挡的探针视频,它与画廊集中无遮挡的视频计算相似度时(二者 ID 相同),因为二者不存在遮挡,PSPNet 检测的遮挡区块的数量为零,则二者在提取特征过程中不使用掩码。而该探针在与画廊集中面部存在遮挡的视频计算相似度时(二者 ID 不同),PSPNet 会检测出遮挡的区块,并得到相应区块的掩码。虽然使用掩码可以排除部分受遮挡影响的特征元素,但也会造成二者相似度的提升,使后者计算出的相似度超越前者,最终造成匹配的错误。

为此,我们提出了一种分区域匹配的方法。当一个探针在与画廊集中各个目标进行匹配时,首先,通过 PSPNet 检测匹配双方的遮挡区块,并根据遮挡区块选择掩码计算双方特征向量的相似度;其次,根据遮挡的区块进行区域类别匹配,即相同区块遮挡而计算出的相似度放进同一区域类别;最

后，对各个区域内的相似度进行区域内排序，并对排序后的结果进行区域间比较。区域间比较时，首先选取各个区域相同排名的相似度，再检索这些相似度所代表的画廊集视频，最后对这些画廊集视频进行两两比较，比较的双方再次与探针视频计算相似度。此时，需要选择相同的掩码（对比较双方的遮挡区块求并集），相似度小的舍弃，直至剩余一个。

例如，当探针视频面部图像中区块 {1} 存在遮挡时，将其在画廊集中进行匹配。假设画廊集有墨镜和口罩两类遮挡区域以及无遮挡区域；其中墨镜遮挡的区块为{3,5}；口罩遮挡的区块为{7,8}。此时，结合探针视频，可以划分出三类区域：{1}、{1,3,5}以及{1,7,8}。对于每一类区域选择对应掩码，并计算探针与画廊集中该类目标的相似度，再进行类内纵向排序与跨类区域横向比较：首先，对{1,3,5}区域以及{1,7,8}区域中排名最高的相似度所代表的画廊集视频进行比较，遮挡区块设置为{1,3,5,7,8}；其次，查找字典选择掩码后再次与探针计算相似度，相似度高的留下与{1}区域进行相同步骤的比较，得到最相似的画廊集视频；最后，对区域内排序结果第二的目标进行相同步骤的跨类区域横向排序。

3.4 网络训练及再识别

本文所提出的监控视频人脸再识别算法的完整训练过程包括以下三步：

1）使用 CASIA-WebFace[11]数据集训练主干网络，损失函数采用大边缘余弦损失函数[34]；

2）构建 PDSN 网络，载入主干网络的模型的权重，并用成对图片进行训练并构建掩码字典；

3）固定主干网络权重，用视频帧图像序列训练时间注意力机制网络。

在大多数再识别应用场景中，用户的 ID 无法事先获得。因此画廊集需要随着摄像头拍摄时间而不断扩充，即为再识别问题中的开集再识别。当目标出现在摄像头下时，应首先判断是否为新目标。我们用 $G=\{G_1, G_2, \cdots, G_N\}$ 表示画廊集，用 $P=\{P_1, P_2, \cdots, P_N\}$ 表示探针集。对于一个探针 $P_x \in P$，其采取分区域匹配策略得到的最终排序结果可用公式（7）进行判断是否为新目标。

$$\max_{i=1,2\cdots N}\left(\mathrm{dist}\left(P_x, G_i\right)\right) \geqslant Y \tag{7}$$

其中，dist 是 P_x 和 G_i 的距离，Y 是阈值。当最大的余弦距离低于阈值，则判断探针不在画廊集中，需要为其注册一个新的 ID 并添加到画廊集。若超过阈值，则用公式（8）获得与 $P_x \subseteq P$ 相匹配的 ID。

$$\mathrm{ID}\left(P_x\right)=\mathrm{ID}\left(G_{i^*}\right), i^*=\underset{i=1,2\cdots N}{\operatorname{argmax}}\left(\mathrm{dist}\left(P_x, G_i\right)\right) \tag{8}$$

4 实验及结果分析

4.1 实验设置

4.1.1 数据集

COX 人脸监控视频数据集[35]是一个由 3 台摄像机拍摄，拥有 1000 个不同的 ID 以及 3000 段视频序列的数据集。和传统人脸视频数据集相比，COX 包含更多在姿势、表情、光照、模糊和面部分辨率等方面有自然变化的帧。由于 COX 数据集中人的面部没有遮挡，为了验证本文算法，我们在此数据集基础上使用文献[36]的方法补充了常见的面部遮挡。

4.1.2 场景设置

实验中，我们共选取了三种不同场景，如下：

（1）画廊集与探针集均不存在遮挡。

（2）画廊集不存在遮挡，探针集均存在遮挡。

（3）探针集与画廊集同时存在遮挡与非遮挡的视频。

4.2 实验结果

4.2.1 时间聚合技术方法的对比

在本文实验中，Baseline 对应的是文献[33]提出的改良的 Resnet50 模型，并在基于图像的人脸数据集 CASIA-WebFace 上用大边缘余弦损失函数进行训练。Baseline 会在匹配的两段视频中随机抽取一帧图像计算相似度。为验证不同时间聚合方式的识别效果，我们将 Baseline 与多帧特征向量求均值（Avg）、平均池化（AvgPool）、最大池化（MaxPool）和时间注意力机制（TA）动态分配权重等不同时间聚合方法分别结合测试效果。为了评估实验结果，我们使用了 rank-n 和 mAP 作为评价指标。rank-n 表示的是搜索结果中前 n 项里存在正确结果的概率。mAP 表示平均准确率，用于衡量算法的搜索能力，测试结果如表 1 所示。从表 1 中可以发现：通过注意力机制（TA）动态分配权重有着最好的效果；而平均池化（AvgPool）或最大池化（MaxPool）操作会削弱或过度强调有代表性特征的贡献，导致准确度不如

表1　画廊集与探针集人脸均无遮挡情况下准确率(%)

Tab. 1　Accuracy(%) with unobstructed faces in both gallery set and probe set

方法	rank-1	rank-5	rank-10	rank-20	mAP
Baseline	61.3	74.6	79.9	84.4	67.7
Baseline+Avg	94.1	97.5	98.4	98.9	95.8
Baseline+MaxPool	94.2	97.9	98.5	98.9	95.9
Baseline+AvgPool	94.6	97.8	98.4	98.8	96.0
Baseline+TA	95.2	98.6	99.1	99.3	96.7

TA。

4.2.2　掩码字典的有效性验证

我们在画廊集无遮挡而探针集均有遮挡的情况下测试了掩码字典(MD)的效果,实验结果如表2所示。可以看出,“TA+MD”的组合方式对再识别的准确度有明显提升。这是因为掩码字典方法会在识别中对遮挡元素进行定位并排除影响。然而实际情况中,监控摄像头首次拍入的面部图像就可能带有遮挡,从而造成画廊集中的数据特征缺失。

我们进一步测试了画廊集与探针集同时存在遮挡与非遮挡情况下的人脸再识别,测试结果如表3所示。可以看出,相比于表2,随着暴露的面部特征变多,不使用掩码字典的准确度有所提升,但使用掩码字典后的准确度却显著的下降,特别是在“Baseline+MD”的情况下,准确率甚至低于只使用Baseline的情况。

此种异常情况的可能原因是由于掩码会把遮挡的元素设置为零,随着丢弃的元素越多,最终得到的特征向量间的相似度越高(即使是不同ID)。这就会导致遮挡视频间的相似度(不同ID)高于未遮挡视频间的相似度(相同ID),且随着丢弃阈值μ的提高,这个问题会愈发严重。为了验证猜测,本文设置了不同的丢弃阈值来测量rank-1的准确率,测试结果验证了我们的猜想,结果如表4所示。可以看出,随着丢弃阈值的增大,匹配的准确率越低。在画廊集中全是清晰无遮挡的视频帧时,这个现象并不会造成匹配准确度大幅度下降。这是因为探针与画廊集进行匹配时丢弃了相同区域的特征,探针视频与画廊集每段视频相似度都会提高但不会改变排序,但在画廊集与探针集都存在遮挡与非遮挡时,会改变相似度的排序导致再识别出现误差。

表2　画廊集人脸无遮挡,探针集人脸均存在遮挡情况下准确率(%)

Tab. 2　Accuracy(%) of gallery set with unobstructed faces and probe set with obscured faces

方法	rank-1	rank-5	rank-10	rank-20	mAP
Baseline	25.5	43.3	51.8	63.1	34.5
Baseline+TA	58.9	78.1	84.8	89.7	67.8
Baseline+MD	29.2	48.0	56.2	65.3	38.4
Baseline+TA+MD	65.1	82.5	88.8	92.5	73.2

表3　画廊集、探针集人脸同时存在遮挡与非遮挡情况下准确率(%)

Tab. 3　Accuracy(%) in the presence of both occlusion and non-occlusion for gallery set and probe set faces

方法	rank-1	rank-5	rank-10	rank-20	mAP
Baseline	33.6	48.9	55.8	65.0	41.3
Baseline+TA	59.3	74.1	80.3	86.9	66.3
Baseline+MD	27.6	44.8	54.1	63.5	36.4
Baseline+TA+MD	62.2	78.9	85.0	89.7	69.8

表4　不同丢弃阈值μ对再识别rank-1准确率(%)的影响

Tab. 4　The effect of feature discarding threshold to rank-1 accuracy(%)

μ	0	0.05	0.15	0.25	0.35	0.45
画廊集有遮挡	59.3	69.6	70.9	62.2	52.4	43.8
画廊集无遮挡	58.9	63.8	65.0	65.1	59.3	53.3

4.2.3　分区域匹配的有效性验证

为了解决掩码字典在再识别中出现的问题(即随着遮挡区块的增多,丢弃的特征会越多,两段视频的相似度也会越高),我们提出了分区域匹配(SRM)的方法来降低掩码字典造成的误差。测试的结果如表5所示,可以看出依靠分区域匹配的方法可以显著降低掩码字典在再识别下的误差。特别是当丢弃阈值越大、或遮挡越严重时,该方法的准确率越高。这是由于我们的匹配方法在进行相似度排序时,考虑到了特征元素丢弃后造成的误差,并额外使用相同掩码进行了一次判断。当丢弃阈值达到0.25时,分区域匹配方法达到了73.0%的准确率,而原方法准确率只能达到62.2%。

4.2.4　与经典方法的对比

本文的方法与传统的人脸识别方法在各个数

表5 不同的丢弃阈值μ在画廊集、探针集人脸同时存在遮挡与非遮挡情况下rank-1准确率(%)

Tab. 5 Different discard thresholds μ in gallery set, probe set faces with both occlusion and non-occlusion case rank-1 accuracy(%)

μ	0.05	0.15	0.25	0.35	0.45
Baseline+TA+MD	69.6	70.9	62.2	52.4	43.8
Baseline+TA+MD+SRM	69.3	72.1	73.0	66.6	61.0

据集上进行了对比,我们复现了几个经典的人脸识别模型,并使用相同的数据集CASIA-WebFace进行了训练,最后在3种不同的数据集上进行了测试,测试结果如表6所示。其中LFW[10]是一个标准的人脸测试基准数据集,拥有6000对测试图像。COX-Masked为探针集与画廊集同时存在遮挡与非遮挡面部图像的视频数据集。对于COX与COX-Masked数据集,我们的方法使用时序注意力机制为多张图片分配不同权重;其他几种方法则选取相同的图片,并对得到的特征向量求取均值,以此代表整段视频的特征。

表6 不同方法在各个数据集上的准确率(%)

Tab. 6 Accuracy(%) of different methods on each dataset

方法	LFW	COX	COX-Masked
PDSN[9]	99.00	94.1	66.5
CosFace[34]	99.20	91.4	49.8
ArcFace[33]	99.10	92.6	51.2
SphereFace[37]	99.22	90.8	43.8
Baseline	99.00	94.1	55.3
Ours	99.00	95.2	73.0

可以看出,本文方法在三个不同的数据集上均展现出较好的识别结果。另外,虽然在LFW上未能有最好效果,但在监控视频数据集下,特别是当面部出现严重的遮挡时,传统方法的准确率大幅度下降,而我们的方法依然能保持较好的识别效果。

5 结论

本文提出了一种基于深度学习的人脸再识别算法,该方法通过结合注意力机制和掩码字典,并依靠提出的分区域匹配方法,降低了掩码字典在再识别场景下的误差,有效提升了监控视频下人脸再识别的准确率。该方法解决了基于全身特征的再识别方法无法长期进行再识别的缺陷,并通过对面部遮挡进行处理,提高了面部存在遮挡时的再识别准确率。在合成遮挡的COX数据上的实验结果表明,本文所提方法可以充分利用面部的有效特征提升深度模型的面部遮挡鲁棒性,进而实现长期可靠的再识别。

本文研究专注于监控视频中面部特征的再识别。实际应用中,监控视频的清晰度、分辨率、光照等因素难免对识别准确率造成影响。下一步,我们将研究对衣着服饰及姿态等信息的可靠性评估,并将上述特征与面部特征融合匹配,以进一步提高再识别准确率。

参考文献

[1] GONG S, CRISTANI M, LOY C C, et al. The re-identification challenge[M]//Person re-identification. Springer, London, 2014: 1-20.

[2] LI Minxian, ZHU Xiatian, GONG Shaogang. Unsupervised tracklet person re-identification[J] IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(7): 1770-1782.

[3] KOIDE K, MENEGATTI E, CARRARO M, et al. People tracking and re-identification by face recognition for rgb-d camera networks[C]//2017 European Conference on Mobile Robots (ECMR). Paris, France. IEEE, 2017: 1-7.

[4] CHENG Zhiyi, ZHU Xiatian, GONG Shaogang. Face re-identification challenge: are face recognition models good enough?[J]. Pattern Recognition, 2020, 107: 107422.

[5] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 7132-7141.

[6] WANG Yitong, GONG Dihong, ZHOU Zheng, et al. Orthogonal deep features decomposition for age-invariant face recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 738-753.

[7] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: a unified embedding for face recognition and clustering[J]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 815-823.

[8] WEN Yandong, ZHANG Kaipeng, LI Zhifeng, et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision, Springer, 2016: 499-515.

[9] SONG Lingxue, GONG Dihong, LI Zhifeng, et al. Occlu-

sion robust face recognition based on mask learning with pairwise differential siamese network [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 773-782.

[10] HUANG G B, MATTAR M A, BERG T L, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments [C]//Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition, 2008.

[11] YI Dong, LEI Zhen, LIAO Shengcai, et al. Learning face representation from scratch. ArXiv Preprint ArXiv: 1411.7923, 2014.

[12] SUN Yi, WANG Xiaogang, TANG Xiaoou. Deep learning face representation from predicting 10, 000 classes [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. IEEE, 2014: 1891-1898.

[13] GUO Yandong, ZHANG Lei, HU Yuxiao, et al. MS-celeb-1M: A dataset and benchmark for large-scale face recognition [C]//European Conference on Computer Vision, Springer, 2016: 87-102.

[14] KEMELMACHER-SHLIZERMAN I, SEITZ S M, MILLER D, et al. The megaface benchmark: 1 million faces for recognition at scale[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. IEEE, 2016: 4873-4882.

[15] LI Zechao, TANG Jinhui. Semi-supervised local feature selection for data classification [J]. Science China Information Sciences, 2021, 64(9): 1-12.

[16] WAN Weitao, CHEN Jiansheng. Occlusion robust face recognition based on mask learning [C]//2017 IEEE International Conference on Image Processing. Beijing, China. IEEE, 2017: 3795-3799.

[17] SÁEZ TRIGUEROS D, MENG Li, HARTNETT M. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss [J]. Image and Vision Computing, 2018, 79: 99-108.

[18] YU Guanwen, ZHANG Zhaogong. Face and occlusion recognition algorithm based on global and local [J]. Journal of Physics: Conference Series, 2020, 1453(1): 012019.

[19] DUAN Qingyan, ZHANG Lei. Look more into occlusion: realistic face frontalization and recognition with boostgan [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 214-228.

[20] DANTCHEVA A, DUGELAY J L. Frontal-to-side face re-identification based on hair, skin and clothes patches [C]//2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance. Klagenfurt, Austria. IEEE, 2011: 309-313.

[21] FARINELLA G M, FARIOLI G, BATTIATO S, et al. Face re-identification for digital signage applications [C]//International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage. Springer, Cham, 2014: 40-52.

[22] QIU Qiang, NI Jie, CHELLAPPA R. Dictionary-based domain adaptation methods for the re-identification of faces [M]//Person Re-Identification. Springer, London, 2014: 269-285.

[23] LI Pei, BROGAN J, FLYNN P J. Toward facial re-identification: experiments with data from an operational surveillance camera plant [C]//2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems. Niagara Falls, NY, USA. IEEE, 2016: 1-8.

[24] LI Pei, PRIETO M L, FLYNN P J, et al. Learning face similarity for re-identification from real surveillance video: A deep metric solution [C]//2017 IEEE International Joint Conference on Biometrics. Denver, CO, USA. IEEE, 2017: 243-252.

[25] WANG Yujiang, SHEN Jie, PETRIDIS S, et al. A real-time and unsupervised face re-identification system for human-robot interaction [J]. Pattern Recognition Letters, 2019, 128: 559-568.

[26] 王超，赵阳，裴继红. 松弛耦合非负矩阵分解的低分辨率人脸识别算法[J]. 信号处理，2020，36(7)：1127-1135.
WANG Chao, ZHAO Yang, PEI Jihong. Low resolution face recognition algorithm based on relaxed coupled non-negative matrix factorization [J]. Journal of Signal Processing, 2020, 36(7): 1127-1135.(in Chinese)

[27] LI Shuang, BAK S, CARR P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 369-378.

[28] LI Zechao, SUN Yanpeng, ZHANG Liyan, et al. Ctnet: context-based tandem network for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[29] BEDAGKAR-GALA A, SHAH S K. A survey of approaches and trends in person re-identification [J]. Image and Vision Computing, 2014, 32(4): 270-286.

[30] DENG Jiankang, GUO Jia, VERVERAS E, et al. Retinaface: single-shot multi-level face localisation in the wild[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 5202-5211.

[31] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017: 6230-6239.

[32] GE Shiming, LI Jia, YE Qiting, et al. Detecting masked faces in the wild with lle-cnns[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017: 426-434.

[33] DENG Jiankang, GUO Jia, XUE Niannan, et al. Arcface: additive angular margin loss for deep face recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 4685-4694.

[34] WANG Hao, WANG Yitong, ZHOU Zheng, et al. Cosface: large margin cosine loss for deep face recognition [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 5265-5274.

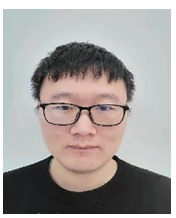
[35] HUANG Zhiwu, SHAN Shiguang, WANG Ruiping, et al. A benchmark and comparative study of video-based face recognition on cox face database[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5967-5981.

[36] ANWAR A, RAYCHOWDHURY A. Masked face recognition forsecure authentication. ArXiv Preprint ArXiv: 2008.11104, 2020.

[37] LIU Weiyang, WEN Yandong, YU Zhiding, et al. Sphereface: deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 212-220.

作者简介

张 博 男,1982年生,河南郑州人。郑州大学网络空间安全学院讲师,博士,主要研究方向为视觉物联网、视频识别、无线自组织网、人工智能等。
E-mail: zhangbo2050@zzu.edu.cn



赵 巍 男,1997年生,河南驻马店人。郑州大学网络空间安全学院在读硕士研究生,主要研究方向为视频识别。
E-mail: 202012332015248@gs.zzu.edu.cn

段鹏松(通信作者) 男,1983年生,山西运城人。郑州大学网络空间安全学院讲师,郑州大学信息工程学院在读博士,主要研究方向为无线感知、视频识别等。
E-mail: duanps@zzu.edu.cn

武 琦 男,1996年生,河南洛阳人。郑州大学网络空间安全学院在读硕士研究生,主要研究方向为视频识别。
E-mail: wq1996@gs.zzu.edu.cn