文章编号：2095-2163(2023)05-0181-06 中图分类号：TP391 文献标志码：A

基于改进纹理特征与迁移学习的人脸表情识别

焦阳阳，黄润才
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：针对小样本表情识别无法有效提取表情特征，以及单一特征提取方法提取的信息不够丰富等问题，提出一种融合视觉注意力（VIT）与改进局部图结构特征的人脸表情识别算法。首先对局部图结构进行改进，在计算特征时采样更多的邻域像素，重新优化权重分配机制，对表情图像使用纹理描述符提取局部特征。同时将表情图像送入视觉注意力模型中，通过迁移学习的方法得到全局特征。最后将局部纹理特征与全局特征进行融合，得到融合特征并使用 Softmax 对表情进行分类。通过在 CK+与 Oulu-CASIA 数据集上进行实验，分别取得了 97.4%与 87.6%的识别准确率。结果表明，本文方法能准确识别出人脸的基本面部表情，与其他方法相比能得到更高的识别准确率。

关键词：图像处理；表情识别；注意力机制；特征融合

Facial expression recognition based on improved texture feature and transfer learning

JIAO Yangyang, HUANG Runcai
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 Aiming at the problem that small sample expression recognition cannot effectively extract expression features and the information extracted by a single feature extraction method is not rich enough, a facial expression recognition algorithm that combines visual attention (VIT) and improved local graph structure features is proposed. Firstly, the local graph structure is improved, more neighborhood pixels are sampled when calculating features, the weight allocation mechanism is re-optimized, and local features are extracted by using texture descriptors for expression images. At the same time, the facial expression images are sent into the visual attention model, and the global features are obtained through the transfer learning method. Finally, local texture features and the global features are fused to obtain fused features and use Softmax to classify expressions. Through experiments on the CK+ and Oulu-CASIA datasets, the recognition accuracy rates of 97.4% and 87.6% were obtained, respectively. The results show that the method in this paper can accurately identify the basic facial expressions of human faces, and can obtain higher recognition accuracy compared with other methods.

【Key words】 image processing; facial expression recognition; attention mechanism; feature fusion

0 引 言

在情感计算领域中，面部表情可以最有效地表达出一个人的情感信息。随着计算机的飞速发展，研究人员开始尝试使用机器自动识别人脸表情。人类的面部表情按照 Ekman 和 Friesen[1] 的定义可分为愤怒、厌恶、恐惧、兴奋、悲伤与惊讶等 6 类。伴随着研究的不断深入，表情识别准确率不断提高，人脸表情识别也开始广泛应用于人们的日常生产生活中，如医疗诊断、驾驶员疲劳检测以及学生课堂管理等。

人脸表情识别按技术路线可分为 3 个步骤：表情图像预处理、特征提取与表情识别。目前用于提取人脸表情特征的方法有两种：基于传统特征的方法与基于深度学习的方法。传统特征提取方法使用不同的编码算子来提取图像的局部信息，其中编码算子主要有局部二值模式（LBP）[2]、方向梯度直方图（HOG）[3] 和局部图结构（LGS）等。Abusham[4] 于 2011 年提出使用局部图结构来提取图像特征，随后 Mohd[5] 等针对 LGS 的非对称性提出了改进后的

作者简介：焦阳阳（1998-），男，硕士研究生，主要研究方向：计算机视觉、表情识别；黄润才（1966-），男，博士，副教授，主要研究方向：智能计算、计算机网络与应用。
通讯作者：黄润才 Email：hrc@ sues.edu.cn

收稿日期：2022-06-08

SLGS。但传统方法只能提取图像的局部信息,无法获取全局语义信息,因此深度学习技术被不断运用在表情识别领域。冯杨[6]等设计了一个小尺度卷积核的神经网络来提取面部表情特征,提高了表情识别的准确率。Jiang[7]等设计了一种混合深度分离残差网络,用于表情特征提取。Transformer 在最近几年中被广泛运用在自然语言处理领域中,研究人员也开始尝试将其运用在计算机视觉任务。Alexey[8]等人提出了 VIT,将 Transformer 移植到了计算机视觉领域中,在图像分类及目标检测等不同任务中均获得了巨大的成功。由于单一方法提取的特征信息不够丰富与小样本表情数据集的表情样本不足等问题,不少研究人员开始尝试融合不同特征以提高特征的表征能力,并尝试使用迁移学习的方法训练神经网络。Wang[9]等使用 MO-HOG 与卷积神经网络提取的特征进行融合。Yang[10]等将几何特征、纹理特征与 ResNet 网络提取的特征进行结合,构成复合面部特征进行表情识别。

传统的 LGS 算子表征了图像的局部信息,但存在一些不足,如采样范围小,左右两边权重分配不均匀。本文首先对 LGS 进行了改进,提出了完全局部图结构(Complete Local Graph Structure,CLGS)特征描述符,通过 CLGS 对表情图像提取局部图像特征。随后通过迁移学习的方法,使用视觉注意力机制对面部表情进行全局特征提取,通过将局部特征与全局特征进行级联融合,使表情特征能够表达更加丰富的信息,最后使用 Softmax 对表情进行分类。

1 人脸表情识别模型

本文提出的人脸表情识别模型如图 1 所示。首先将原始图像经过预处理操作,把背景等无关因素去除掉,然后对图像进行分块处理,分别送入两条支路提取特征向量。将分块后的图像展平成一维向量,在每个向量位置上添加一个位置嵌入,得到 Transformer 编码器的输入向量,经过 Transformer 处理后得到全局特征向量。与此同时将分块后的图像通过 CLGS 特征描述符提取特征,并转换为直方图形式,得到局部纹理特征。最后将全局特征与局部纹理特征拼接融合,经过全连接层后送入 Softmax 中得到最终预测结果。

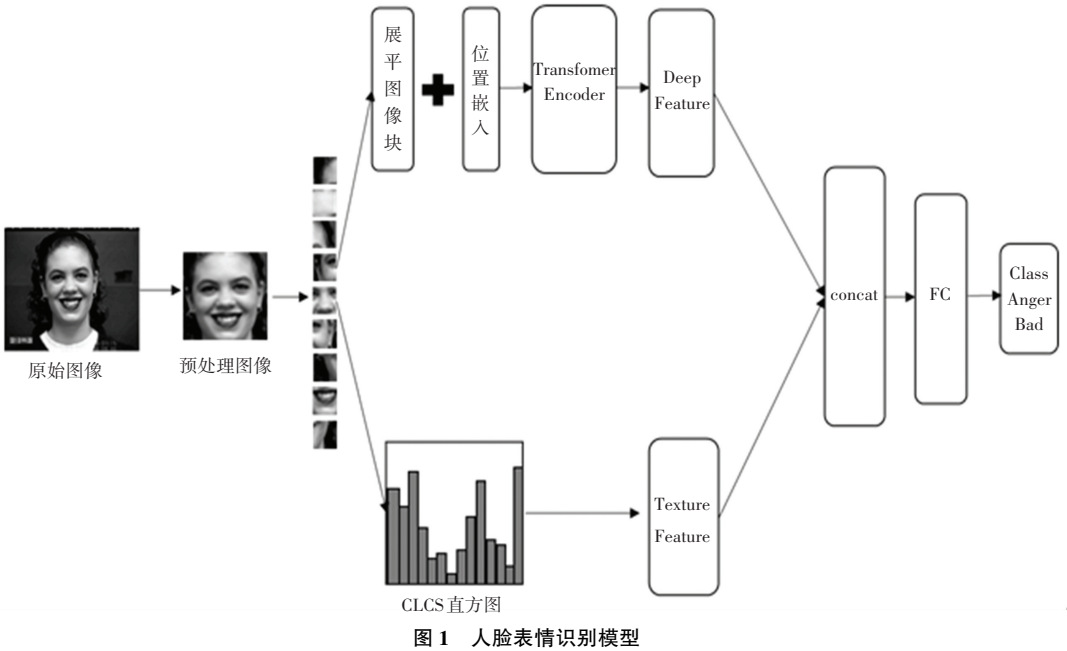


图 1 人脸表情识别模型

Fig. 1 Facial expression recognition model

1.1 图像预处理

在原始图像中存在着许多对表情特征提取无关的信息,如果直接将原图像送入模型进行处理,对表情识别的准确性有一定的影响。因此需要对表情图像进行预处理,包括人脸检测、灰度及尺寸归一化等。检测并裁剪出人脸部位,然后将三通道 RGB 图像转换为灰度图,统一缩放成相同规格的尺寸大小,得到模型所需的输入图像。

1.2 CLGS 特征

Abusham 等提出的局部图结构(LGS)算子用于描述局部纹理特征,其通过构建一个图结构来描述中心像素与周边像素的关系,其计算过程如图 2 所示。在中心像素周边选取 5 个像素点,从左侧开

始按照图中标记的箭头位置，依次比较箭头首端与箭头末端所指像素的大小，大于则将连接两像素的边取 0，小于取 1。沿着图中标记的顺序，依次得到 8 个二进制数，将其按照排列顺序赋予一定的权重转换为十进制数，这个数值即为中心像素的 LGS 特征值。在图 2 中，得到的 8 个二进制数为 00110100，转换为十进制数 52。

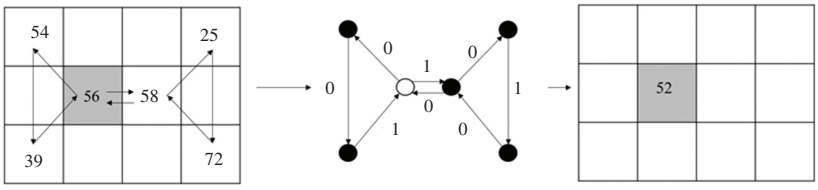


图 2　LGS 编码示意图

Fig. 2　Diagram of LGS coding

从图 2 可知，LGS 存在着对左右两边像素点利用不均匀与权重赋予左重右轻等特点，左边采样两个像素点，权重赋予了 128、64 与 32，而右边虽然多一个采样点，但权重明显比左边低。于是 SLGS 通过在中心像素两侧提取相同数量的像素，解决了采样不平衡的问题。但 SLGS 的采样范围仅利用了周边的一半像素点，另一半像素点则没有利用，并且权重赋予的问题也没有解决。因此，本文提出了 CLGS 描述符对上述问题进行了改进。

如图 3 所示，CLGS 描述符通过构建两个不同的图结构分别计算出特征值，描述符采样范围扩大了一倍。然后根据图 4 所示的权重赋予机制，分别对计算完成后的二进制值乘以相应的权重得到两个特征值，中心像素的两边权重分配更加合理。局部最大特征值能够更好地反映周边像素的变化情况，保存更多的纹理信息，因此取最大的那个值作为中心像素的特征值。将图像分为 16×16 大小的图像块，分别处理为 CLGS 图像，再对这些图像内的特征值进行统计得到特征直方图，作为图像的局部特征信息。

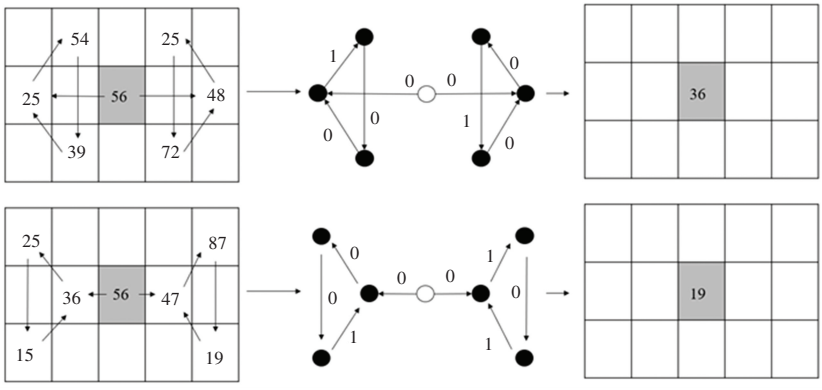


图 3　CLGS 示意图

Fig. 3　Diagram of CLGS

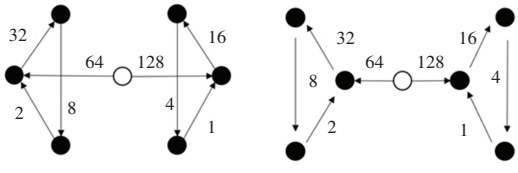


图 4　权重赋予示意图

Fig. 4　Weight assignment diagram

1.3　Vision Transformer（VIT）

VIT 通过计算图像块之间的关系获得图像的全局特征，视觉注意力机制的结构如图 5 所示。

首先将图像进行裁切，分为 16×16 大小的图像块，再将图像块进行展平处理，转换为一维向量。与自然语言处理中的 BERT 类似，在输入序列的最前端添加一个可学习的向量，用于表示分类信息；然后使用可学习的一维向量作为位置向量，与图像向量进行相加，保留位置信息。式（1）展示了输入向量的计算过程：

$$Z_0 = [\boldsymbol{Xclass}; X_p^1 E; X_p^2 E; \cdots; X_p^N E] + \boldsymbol{E}_{pos} \quad (1)$$

式中：$\boldsymbol{Xclass}$ 表示分类向量，$X_p^1 E$ 表示图像块，$\boldsymbol{E}_{pos}$ 为位置向量。

将处理得到的向量序列作为编码器的输入。在整个编码器中，最重要的部件是多头注意力机制，由 z 个自注意力构成。自注意力通过计算单个向量

与其他向量之间的关系得到全局注意力表示,其通过可学习的3个参数Q、K、V来进行表示。具体的计算如式(2):

$$head_j = Attention(Q,K,V) = \text{softmax}(QK^T/\sqrt{d})V \quad (2)$$

其中,d为输入向量的维度,输入向量与3个权重矩阵相乘得到Q、K、V。

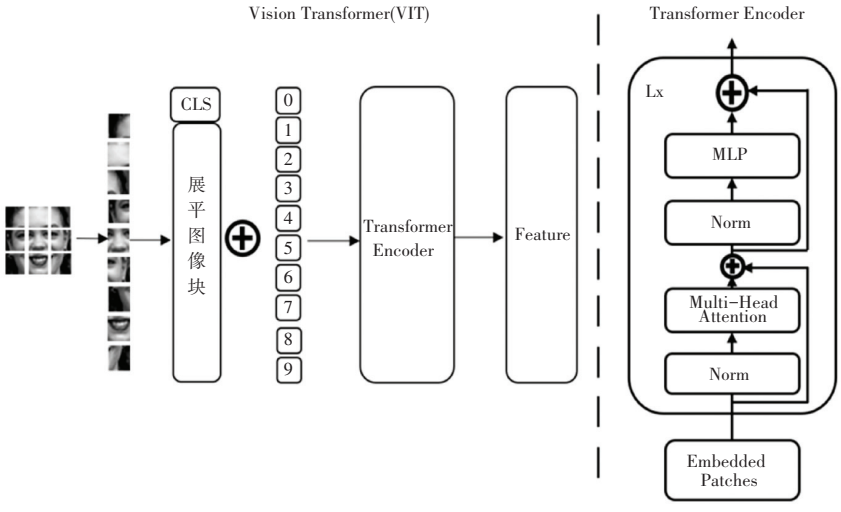


图 5 VIT 示意图

Fig. 5 VIT diagram

在得到单个自注意力矩阵后,使用多个自注意力头,生成多头注意力。计算公式如式(3):

$$MHA(Q,K,V) = concat(head_1,\cdots,head_Z)\boldsymbol{w}^o \quad (3)$$

其中,$\boldsymbol{w}^o$为权重矩阵。输入的向量矩阵首先进行归一化处理,然后送入多头注意力机制中,再次经过归一化以及感知机层,得到输出向量矩阵。在编码过程中,编码器还使用了两个残差连接,其目的是为了防止梯度消失。经过多个编码器的计算后,输出最终的特征向量。

由于表情识别数据集样本小,而深度学习的训练需要大量样本的支撑,因此迁移学习被广泛运用于小样本任务中。迁移学习指的是首先在一个与目标任务类似的任务上进行模型的训练,该任务的样本量非常大,常用的有 ImageNet 数据集。在该任务上训练好模型,然后将模型参数迁移到目标任务上。迁移学习可以解决小样本数据集图像特征不足的问题,本文使用在 ImageNet 上预先训练过的模型,通过微调得到全局特征向量。

1.4 特征融合

如前所述,通过 CLGS 描述符得到了图像的局部特征信息后;再通过迁移学习的方式,使用 VIT 得到了图像的全局信息。在此基础上可通过将局部信息与全局信息相融合,得到最终的表情特征向量。融合方法使用串联拼接的方式:

$$\boldsymbol{X} = concat[\boldsymbol{Xdeep},\boldsymbol{Xclgs}] \quad (4)$$

其中,$\boldsymbol{Xdeep}$为全局特征向量;$\boldsymbol{Xclgs}$为局部特征向量;$\boldsymbol{X}$为模型最终得到的特征向量。

经过全连接层后,通过 Softmax 进行表情分类。在整个模型训练过程中,选择交叉熵损失函数作为梯度下降的优化函数:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y^i \ln\hat{y}^i + (1-y^i)\ln(1-\hat{y}^i)] \quad (5)$$

其中,y^i为真实值;$\hat{y}^i$为预测值;N为样本数。使用梯度下降的方法不断缩小预测值与真实值之间的差异。

2 实验结果与分析

本文模型使用 Pytorch 深度学习框架搭建,操作系统为 Ubuntu, GPU 为 NVIDIA Tesla K80, Python 版本为 3.8。其中神经网络参数设置为:优化器使用 Adam, Batchsize 为 64, Epoch 设置为 20。

2.1 数据集

本文所进行的所有实验均使用 CK+[11]与Oulu-CASIA[12]数据集。CK+数据集包含了 123 个被采集对象的 593 个表情图像序列,其中包含愤怒、厌恶、恐惧、中性、兴奋、悲伤与惊讶 7 种表情,本文选用 CK+数据集最后三帧的图像进行实验。Oulu-CASIA 数据集共采样 80 个人的 6 种基本表情,对比 CK+数据集缺少中性表情,采样方式分为两种:可见光与近红外光,本文使用可见光数据进行实验。将两个数据集均按照 8∶2 的比例划分出训练集和测

试集，模型验证使用五折交叉方法。

2.2　消融实验

为了验证本文提出的模型有效性，进行了模型消融实验，共设置了两组对照实验。实验一仅使用 VIT 视觉注意力机制，实验二使用未经过预训练的 VIT 模型，使模型从头开始训练。实验分别在 CK+与 Oulu-CASIA 数据集上进行，并与本文模型进行对比。实验结果见表 1。

表 1　消融实验结果

Tab. 1　Results of ablation experiments

实验名称	CK+ 准确率/%	Oulu-CASIA 准确率/%
实验一	86.2	80.9
实验二	78.9	78.0
本文模型	97.4	87.6

从表 1 中可见，本文模型比单一特征提高了 8%的准确率，证明本文将全局特征与局部特征进行融合，能够表达更加丰富的信息。实验二的结果证明了经过预训练的模型比重新开始训练的模型识别效果更好，能够提升表情识别的准确率。

2.3　纹理特征对比实验

本文提出了 CLGS 特征描述符，为了验证描述符的有效性，选用了不同类型的描述符进行对比实验。实验选用 LGS、SLGS 作为参照，最终实验结果如图 6 所示。

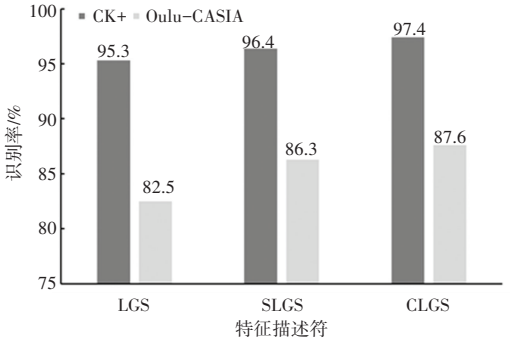


图 6　纹理特征实验结果

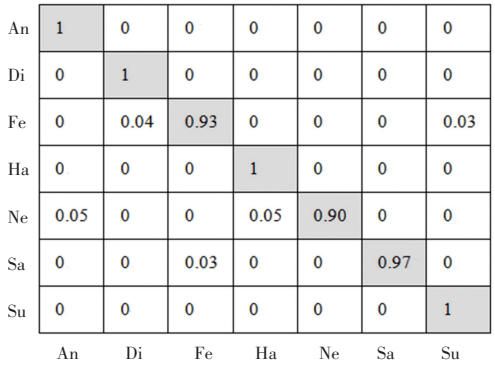
Fig. 6　Experimental results of texture features

从图 6 可知，本文提出的 CLGS 在 CK+数据集上取得了 97.4%的准确率，在 Oulu-CASIA 上得到了 87.6%的准确率，均高于其他特征描述符。CLGS 对目标周围像素的利用率均高于其他纹理特征，而改进后的权重赋予机制对左右两边像素的利用更加合理。

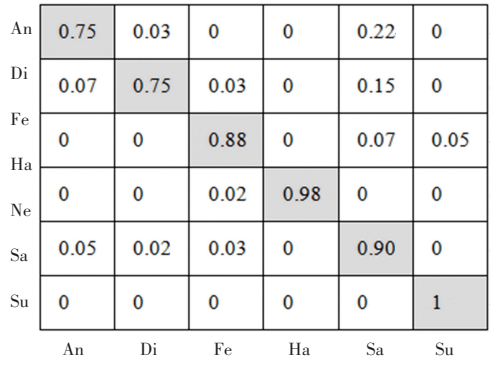
2.4　混淆矩阵

为了反映本文模型对不同表情的识别率，绘制了 CK+和 Oulu-CASIA 数据集的表情分类混淆矩阵。

从图 7(a)可以得知，模型对 CK+数据集中部分恐惧、自然及悲伤图像识别效果较差；在图 7(b)中，悲伤、厌恶及愤怒的识别率明显低于其他表情，模型对于兴奋和惊讶两种表情的识别率较高。分析可知，识别率较低的表情其面部动作相似度高，动作幅度较小，识别率高的表情面部动作幅度较大，特征比较明显。



(a) CK+数据集



(b) Oulu-CASIA 数据集

图 7　混淆矩阵

Fig. 7　Confusion matrix

2.5　算法对比

表 2 展示了本文提出的方法与其它主流表情识别算法的识别率对比。经过对比，无论是传统方法还是深度学习方法，本文模型在 CK+与 Oulu-CASIA 数据集上的识别准确率均高于表中所列其它方法，证明了本文方法的有效性。

表 2　与其他算法识别率对比

Tab. 2　Comparison of recognition rate with other algorithms

算法名称	CK+准确率/%	Oulu-CASIA 准确率/%
STM-ExpLet [13]	94.1	74.6
LDN[14]	90.7	85.2
LBP+CNN[15]	92.8	86.1
DTAGN [16]	97.2	81.4
LOMo[17]	95.1	82.1
本文模型	97.4	87.6

3 结束语

本文对传统 LGS 算子进行了改进,提出的 CLGS 算子能更加合理地表示出图像的局部纹理信息,并通过迁移学习的方式提取了表情的全局特征,将局部特征与全局特征进行融合与表情分类。在 CK+和 Oulu-CASIA 上的实验证明了本文表情识别模型的有效性。由于模型还存在对特定表情识别准确率不高的问题,下一步将继续研究如何提高难区分表情的识别率。

参考文献

[1] EKMAN P, FRIESEN W V. Constants across cultures in the face and emotion [J]. Journal Personality and Social Psychology, 1971, 17(2): 124-129.

[2] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[3] 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.

[4] ABUSHAM E E A, BASHIR H K. Face recognition using local graph structure (LGS) [C] // Proc of International Conference Human-Computer Interaction, Springer, 2011: 169-175.

[5] MOHD F A, SAYEED M S, MUTLIU K S, et al. Face recognition with symmetric local graph structure(SLGS)[J]. Expert Systems with Applications, 2014, 41(14): 6131-6137.

[6] 冯杨,刘蓉,鲁甜. 基于小尺度核卷积的人脸表情识别[J]. 计算机工程, 2021, 47(4): 262-267.

[7] ZHANG S, JIANG D, YU C. A mixed depthwise separation residual network for image feature extraction [J]. Wireless Networks, 2021: 1-12.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[9] WANG H, WEI S, FANG B. Facial expression recognition using iterative fusion of MO-HOG and deep features[J]. The Journal of Supercomputing, 2020, 76(5): 3211-3221.

[10] YANG J, ADU J, CHEN H, et al. A facial expression recongnition method based on dlib, ri-lbp and resnet[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1634(1): 012080.

[11] LUCEY P, COHN J F, KANADE T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]//2010 IEEE computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010: 94-101.

[12] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with gabor wavelets[C]//Proceedings Third IEEE international conference on automatic face and gesture recognition, IEEE, 1998: 200-205.

[13] LIU M, SHAN S, WANG R, et al. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, 2014: 1749-1756.

[14] RIVERA A R, CASTILLO J R, CHAE O O. Local directional number pattern for face analysis: Face and expression recognition [J]. IEEE transactions on image processing, 2012, 22(5): 1740-1752.

[15] LOPES A T, DE AGUIAR E, DE SOUZA A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order [J]. Pattern recognition, 2017, 61: 610-628.

[16] JUNG H, LEE S, YIM J, et al. Joint fine-tuning in deep neural networks for facial expression recognition[C]//Proceedings of the IEEE international conference on computer vision, IEEE, 2015: 2983-2991.

[17] SIKKA K, SHARMA G, BARTLETT M. Lomo: Latent ordinal model for facial analysis in videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, 2016: 5580-5589.

(上接第 180 页)

[5] 于意. 基于改进 YOLOv3 的奶山羊目标检测方法研究[D]. 咸阳: 西北农林科技大学,2021.

[6] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[7] MA N, ZHANG X, ZHENG H T, et al. ShuffleNet V2: Practical guidelines for efficient CNN architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[8] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.

[9] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module [C]//Proceedings of the European Conference on Computer vision (ECCV). 2018: 3-19.

[10] ZHENG Z, WANG P, LIU W, et al. Distance-IoU Loss: Faster and better learning for bounding box regression [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 12993-13000.

[11] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[12] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[13] REDMON J, FARHADI A. Yolov3: An incremental improvement [J]. arXiv preprint arXiv:1804.02767, 2018.

[14] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778-2788.