文章编号 1004-924X(2019)04-0963-08

多模深度卷积神经网络应用于视频表情识别

潘仙张*,张石清,郭文平
(台州学院 智能信息处理研究所,浙江 台州 318000)

摘要:由于视频中的手工特征和主观情感之间的直接相关性很小,识别视频序列中的面部表情是一项很有挑战性的任务,为了克服这个缺陷,有效提高视频中的人脸表情识别性能。本方法采用两个深度卷积神经网络,即空间卷积神经网络和时间卷积神经网络,用于视频中的时空表情特征学习。其中,空间卷积神经网络用于提取视频中每一帧静态的表情图像的空间信息特征,而时间卷积神经网络用于从视频中多帧表情图像的光流信息中提取动态信息特征。然后,将这两个深度卷积神经网络学习到的时空特征进行基于深度信念网络(DBN)的特征层融合,输入到支持向量机实现视频中的人脸表情分类任务。在公共的 RML 和 BAUM-1s 视频情感数据集的测试结果表明,该方法分别取得了 71.06%和 52.18%的正确识别率,明显优于现有文献报导的结果。多模深度卷积神经网络的人脸表情识别方法能提高视频中人脸表情的识别性能。

关 键 词:深度卷积神经网络;多模深度学习;表情识别;时空特征;深度信念神经网络
中图分类号:TP391 **文献标识码**:A **doi**:10.3788/OPE.20192704.0963

Video-based facial expression recognition using multimodal deep convolutional neural networks

PAN Xian-zhang*, ZHANG Shi-qing, GUO Wen-ping

(*Institute of Intelligent Information Processing, Taizhou college, Taizhou* 318000, *China*)
* *Corresponding author, E-mail: pxz@tzc.edu.cn*

Abstract: Recognizing facial expressions in video sequences is challenging because of the difficulty in distinguishing between hand-crafted features and subjective emotions. To solve this problem, we aim to improve the performance of facial expression recognition in videos. Our method used two deep convolutional neural networks (DCNNs) (i.e., spatial and temporal convolutional neural networks) to learn the temporal-spatial expression features in videos. The spatial convolutional neural network was used to extract the spatial features of static expression images from each video frame, where as the temporal convolutional neural network was used to extract dynamic features from optical flow information hidden in multi-frame expression images of a video. The temporal-spatial features were then fused using a deep belief network. Finally, support vector machines were employed to perform facial expression classification. Based on experimental results on public RML and BAUM-1s video-based emotional datasets, our method achieved an accuracy of 71.06% and 52.18%, respectively,

收稿日期:2018-08-06;**修订日期**:2018-10-13.
基金项目:浙江省公益技术研究计划基金资助项目(No. LGF19F020009);浙江省自然科学基金资助项目(No. LY14F020036,No. LY16F020011);国家自然科学基金资助项目(No. 61203257)

which is clearly better than the results of existing studies. This study thus showed that our multimodal DCNN can improve the performance of facial expression recognition in videos.

Key words: deep convolutional neural network; multimodal deep learning; facial expression recognition; temporal-spatial features; Deep Belief Network(DBN)

1 引　言

情感交流是人与人之间最自然的一种交流方式。视频中的人脸表情识别是指通过计算机自动识别出视频中人脸的情感状态。近年来，越来越多研究者致力于研究人脸表情的自动识别技术，因为该研究在人机交互系统[1]、智能视频监控系统[2]等领域具有重要的应用价值。

根据输入图像形式的不同，人脸表情识别系统可以分为：基于视频的表情识别系统和基于静态图像的表情识别系统。基于视频的表情识别系统的技术难度更大，因为动态序列图像比单张图像更难处理。本文只关注基于视频的表情识别方法。一个基本的视频表情识别系统主要有3个步骤：视频预处理、表情特征提取和表情分类。视频预处理主要是从视频中的序列图像中检测并提取出人脸；特征提取是指从视频中的人脸图像中提取能够刻画表情的特征；表情分类是指把这些提取的特征输入到分类器就可以实现表情的识别。

近年来，随着深度学习技术[3-6]在图像领域取得的突破，也有研究者尝试将深度卷积神经网络(Deep Convolutional Neural Networks, DCNNs)[7]用于静态图像的人脸表情识别。然而，这些基于静态图像的空间特征并不能很好地实现视频的人脸表情分类，因为这种方式没有考虑视频中表情的时序动态信息。为此，本文提出一种能直接提取视频中的人脸图像的空间特征以及时序动态信息特征的多模深度卷积神经网络模型。

本文的主要贡献如下：

提出了多模深度卷积神经网络模型，提取视频中的时空表情特征，并且采用DBN(Deep Belief Network)网络对时空特征进行融合，从而进一步改善人脸表情识别性能。在RML和BAUM-1s数据集上的实验结果表明，本文方法比现有方法取得的正确识别率要分别高出至少2%。

在视频表情识别的3个步骤中，视频的表情特征提取是其中最重要的一个环节。目前，已有的文献在视频表情识别过程中普遍采用手工特征。例如，文献[8]提取视频中的眼睛、鼻子、嘴唇运动的光流信息，将其作为视频的表情分类特征。脸部动画参数(Facial Animation Parameters, FAPs)是另一种被普遍用来识别视频中人脸表情的特征，FAPs主要是描述外嘴唇轮廓和眉毛的运动信息[9]。文献[10]提取序列图像的人脸动作单元的变化信息的haar-like特征，再把这些特征编码成二进制形式用于视频中的人脸表情分类。近年来，从视频中提取局部二值模式(Local Binary Pattern, LBP)特征[11]，以及它的变种，如LBP-TOP[12]，LPQ(Local Phase Quantizatio)[13]都被应用于视频的表情识别[14]。Zhao等人[15]使用LBP-TOP编码视频中的脸部序列图像，并用它作为人脸表情识别的特征。尽管这些手工特征已被成功地用于视频中的人脸表情识别，但是它们可靠性不够，因此不能有效地区分视频中的人脸表情。

为此，深度学习技术[3]可能提供一个线索。目前，一些代表性的深度学习方法，如深度卷积神经网络[16]和深度DBN[6]网络因其具有很强的特征提取能力已经被应用到表情识别中。文献[17]采用一个包含3个卷积层和2个池化层的卷积神经网络结构用于提取人脸表情特征。文献[18]提出一种基于帧变换(Frame Transformer)的深度学习框架。其采用5个卷积层和3个全连接层的网络提取视频中每帧的特征，并把这些特征输入到一个基于帧变换的情感分类网络，以便得到该视频中人脸表情的分类结果。文献[19]提出使用DBN(Deep Bolzmann Machines)提取视频中的高级特征，能够有效地进行视频表情识别。

然而，上述文献将DCNNs和DBM用于视频表情识别时，只考虑到提取视频表情中的图像空间特征，并没有考虑到对表情识别有帮助的视频序列中的动态变化特征，如光流信息。为了充分利用视频表情中的光流信息，本文提出了一种基于多模深度卷积神经网络的视频表情识别方法，其包含空间卷积神经网络和时间卷积神经网络，

分别用于从视频中提取高层次的空间特征和时间特征。在公共情感数据集 RML[20] 和 BAUM-1s[21] 上实验测试表明，本文的方法在视频人脸表情识别上取得了良好的效果。

2　多模深度卷积神经网络的视频表情识别模型

本文的基于多模深度卷积神经网络的视频表情识别模型框架，如图 1 所示。它主要包含两个独立的深度卷积神经网络，分别是时间卷积神经网络和空间卷积神经网络。时间卷积神经网络主要是处理视频的光流信号，提取出高层次的时间特征。空间卷积神经网络主要是处理视频中每帧的人脸图像，并提取出高层次的空间特征。然后，把提取出的时间特征和空间特征分别做平均池化(Average-pooling)，在特征层上进行基于 DBN 的时空特征融合。最后，把融合后的时空特征用支持向量机(Support Vector Machines，SVM)完成视频表情的分类任务。具体来说，该模型主要包含 3 个步骤：视频预处理，深度时空表情特征的提取和融合，表情的分类。

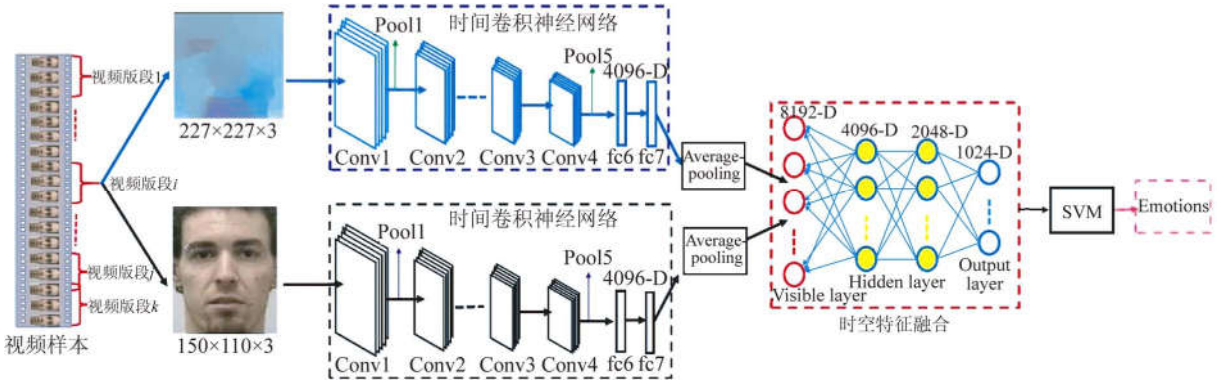

图 1　基于多模深度卷积神经网络的视频表情识别模型

Fig. 1　Video-based facial expression model of using multimodal deep convolutional neural network

2.1　视频预处理

每段视频的时长不同，而 DCNNs 需要固定大小的数据输入。因此，本文把视频分割成很多固定时长的片段，作为时间卷积神经网络和空间卷积神经网络的输入。通过这种方式，也可以在一定程度上扩大用于 DCNNs 训练的数据集。

假设 L 代表该视频片段包含的帧数。选择合适大小的 L 对该视频的时间信息特征的提取也很重要。如果 L 太小，该视频段包含的动态变化信息不足。相反，如果 L 太大，该视频片段可能会包含太多噪声，从而影响识别性能[22]。

本文通过对 L 在[2，20]范围内进行依次搜索，发现当 $L=16$ 时，时间卷积神经网络取得的效果最好，因此，本文将每段视频分割成 16 帧大小的片段。当 $L>16$，本文丢弃该视频的前面和后面的$(L-16)/2$ 帧；当 $L<16$，本文复制该视频的前面和后面的$(16-L)/2$ 帧。对于 1 段 $L=16$ 的视频片段，包含 15 帧的光流图像，因为每相邻两帧的空间图像会生成一帧的光流图像。该光流图像代表相邻两帧的相应位置的位移信息，具体计算过程如下：设视频中的相邻两帧 t 和 $t+1$，位移向量 d_t 代表该视频的位移信息。光流图像 I_t 由 $d_t x$ 和 $d_t y$ 组成，$d_t x$ 和 $d_t y$ 是 I_t 的两个通道，分别代表视频中相邻两帧图像位置的水平位移分量和垂直位移分量。考虑到 DCNNs 的输入是 3 个通道的 RGB 图像，因此，本文拟采用 Horn-Schunck[23]方法计算出光流图像 I_t 的幅度分量 $d_t z$，作为 I_t 的第 3 个通道，即：

$$d_t z=\sqrt{d_t^2 x+d_t^2 y}. \tag{1}$$

对于空间卷积神经网络输入图像的预处理，本文采用文献[24]中的人脸检测算法实时提取出视频片段中每帧包含的人脸图像。对于一幅人脸图像，其宽度一般是两只眼睛距离的 2 倍，而人脸图像的高度一般是两只眼睛距离的 3 倍。因此，根据两只眼睛的标准距离(55 pixel)，本文拟采用文献[25]的方法从原始人脸图像中裁剪出包含嘴巴、鼻子、额头等关键表情部位的区域大小为 150×110×3 的图像，作为空间卷积神经网络的

输入。为了符合 DCNNs 的输入大小，本文分别把提取的光流图像和人脸表情图像都缩放到 227×227×3 的大小。

2.2 深度时空特征提取

当完成视频预处理之后，就可以采用 DCNNs 提取视频片段中高层次的时空表情特征。由于视频表情数据集一般都比较小，因此本文不直接训练自己的 DCNNs。而是采用已经预训练好的 DCNN 模型在目标视频情感数据集上进行微调训练，从而实现迁移学习的目的。

为此，本文首先采用在 ImageNet 数据集上已经预训练好的 AlexNet[7] 参数将图 1 中的时间卷积神经网络和空间卷积神经网络初始化。这两个卷积神经网络的结构与 AlexNet 模型相同，包含 5 个卷积层，3 个池化层和 3 个全连接层（fc6，fc7，fc8）。其中，fc6 和 fc7 包含 4 096 个结点，而 fc8 是分类层（如 Softmax），对应目标数据的类别数目，如表情的种类数量。本文采用这两个卷积神经网络 fc7 输出的 4096-D 特征，分别作为 DCNNs 学习到的时间特征和空间特征。

2.3 深度时空特征融合

在 DBN 中的每个 RBM 都可以用于学习时空特征表示，因此采用由多层 RBM 构成的 DBN 作为深度特征融合模型。本文使用具有 4 096 个隐藏节点的高斯-贝努利 RBM 作为 DBN 的第 1 层，使用具有 2 048 个隐藏节点的贝努利-贝努利 RBM 作为 DBN 的第 2 层，使用具有 1 024 个隐藏节点的贝努利-贝努利 RBM 作为 DBN 的第 3 层，它的 1024-D 维输出特征作为本文最终的情感识别特征。这样本文采用的 DBN 结构如下 8192-4096-2048-1024-C，C 为情感类别。

为了获得视频的全局特征，拟对 DCNNs 在视频片段上学习到的时间特征和空间特征分别采用平均池化方法，计算出视频的全局特征，然后在特征层水平上进行时空特征融合，从而得到一个 8192-D 的特征。再用 DBN 模型对 8192-D 的特征进行特征融合，最终选择出 1024-D 的特征作为该视频的时空特征。

2.4 表情的分类

当获得 1024-D 的视频全局特征之后，采用线性支持向量机（SVM）实现最终的视频表情识别任务，输出视频的表情类别。

为了方便微调本框架的神经网络，本文分别将光流图像和人脸图像的大小调整为 227×227×3 作为 CNNs 的输入。因为多模深度卷积神经网络直接提取出了该视频的高级时空特征，这些特征直接被 SVM 作为输入，从而得到该人脸表情的分类结果。

2.5 网络训练

图 1 中的时间卷积神经网络和空间卷积神经网络的训练分别采用微调方式实现。为此，拟将原始 AlexNet 模型中的 fc8 进行更新，即采用目标视频情感数据集的情感类别数（如 6 类）代替原来 ImageNet 的 1 000 类图像类别数目。然后采用反向传播（Back Propagation，BP）算法进行网络的训练，以便更新两个卷积神经网络的网络权重值。具体训练过程如下：

假设数据集 $X=\{(a_i,b_i,y_i)\}(i=1,2,\cdots,N)$，$i$ 代表该视频片段的第 i 帧，a_i 代表提取出的光流图像，b_i 代表静态的人脸图像，y_i 代表该视频片段的情感标签。

对于空间卷积神经网络 B 的训练，目标是使得负的对数似然损失函数 H 最小化，即：

$$\min_{W^B,\lambda^B}\sum_{i=1}^{N}H(\mathrm{softmax}(W^B v^B(b_i;\lambda^B)),y_i), \quad (2)$$

$$H(B,y)=-\sum_{j=1}^{k}y_i\log(y_j^B), \quad (3)$$

其中：W^B 代表 Softmax 层的权值，$v^B(b_i;\lambda^B)$ 代表 fc7 层输出的特征 4096-D，λ^B 代表网络 B 的参数。Softmax 层对数损失的计算见公式(3)。其中，y_j^B 代表空间卷积神经网络 B 的 Softmax 层的第 j 个输出值，k 代表数据集的视频表情类别的数目。同理，时间卷积神经网络的训练方法和上述空间卷积神经网络的训练方法一样。

图 1 中基于 DBN 的特征融合网络训练主要分为如下 2 个步骤：

(1)采用一种贪婪分层训练算法，以自底向上的方式实现无监督的预训练[26]。这一步主要是通过公式(4)使 DBN 中每个 RBM 的输入数据 v_i 和重构数据 v_i' 的误差最小：

$$\min_{W^f,\theta^f}\sum_{i=1}^{N}C(v_i,v_i'), \quad (4)$$

其中：N 是训练的样本总数，$C(v_i,v_i')$ 是交叉熵损失函数：

$$C(v_i,v_i')=$$

$$\sum_{d}^{D}(-v_{i,j}\log v'_{i,d}+(1-v_{i,d})\log(1-v'_{i,d})).$$

(5)

(2)在第一步无监督学习完成后，每个 RBM 多获得了初始的网络参数，再进行有监督地微调训练本网络，优化 DBN 网络参数。

本文以 DBN 的最后一个隐含层输出作为 SVM 分类器的输入。整个 DBN 网络的训练以反向传播算法来调整网络参数。

3　实验结果与分析

为了评价本文方法的视频表情识别性能，拟在公共的 RML 和 BAUM-1s 视频情感数据集上进行实验测试。DCNNs 训练时，将批处理 batch 大小设为 30，学习速率设为 0.001，最大循环(epoch)次数设为 2 000，采用 MatConvNet toolbox 工具箱。DBN 由 3 个 RBM 组成，每个 RBM 循环次数为 100，DBN 采用 DeeBNet toolbox 工具箱。实验平台为显存 25 GB 的 NVIDIA GPU。SVM 采用线性核函数，即 linear kernel 核函数，采用 LIB-SVM package 包实现。实验时，通过训练与测试对象无关的 Leave-One-Subject-Out (LOSO)交叉验证方法测试本文方法的性能，最后取平均识别率作为最终报导的结果。

3.1　实验的数据集

RML[20]数据集有 720 段视频，由 8 个人的表情组成。该数据集上有 6 种表情，即生气、厌恶、害怕、高兴、悲伤和惊讶(angry，disgust，fear，joy，sadness and surprise)。每段视频样本的平均时长约为 5 s。视频中每帧图像的大小为720×480×3。图 2 是从 RML 数据集中提取的部分人脸图像。

BAUM-1s 数据集[21]有 1 222 段视频，由 31 个人的表情组成，该数据集上有 8 种表情，本文只研究其中的 6 种表情，即生气、厌恶、害怕、高兴、悲伤和惊讶(angry，disgust，fear，joy，sadness and surprise)，它们来自 521 个视频样本。视频中每帧图像的大小为 720×576×3。图 3 是从 BAUM-1s 数据集的视频中提取的部分人脸图像。

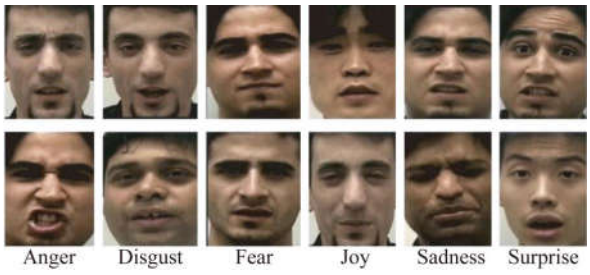

图 2　从 RML 数据集中提取的人脸图像

Fig. 2　Samples of cropped facial images on RML database

图 3　从 BAUM-1s 数据集中提取的人脸图像

Fig. 3　Samples of cropped facial images on BAUM-1s database

因为每个 RML 和 BAUM-1s 数据集中的每个视频样本被分割多个视频片段作为 DCNNs 的输入，这样就增大了本文训练数据集的数据量。在本实验中，通过这种方法把 RML 中的 720 个视频样本变成了 12 000 个视频片段，把 BAUM-1s 的 521 个视频样本变成了 7 000 个视频片段。

3.2　实验结果分析

为了评估本文的时空卷积神经网络提取特征的有效性，表 1 给出了不同 DCNNs 学习到的特征及其融合的表情识别结果。从表 1 可以看出，时空特征对视频的人脸表情识别都是非常有用的。例如，空间 DCNN 特征和时间 DCNN 特征在 RML 数据集上的正确识别率分别达到了 64.40%和 49.94%。与时间 DCNN 特征相比，空间 DCNN 特征更有用。这说明时间 DCNN 特征对视频的表情识别也起到非常关键的作用。此外，融合时空 CNN 特征取得的正确识别率达到了 71.06%。这证明了将时空 DCNN 特征用于视频表情识别的优势，因为它能提供更好的视频表情的区分度信息。

表 1　不同 DCNN 特征取得的识别结果

Tab. 1　Recognition performance of different DCNN features (%)

特征类型	BAUM-1s	RML
空间 DCNN 特征	50.48	64.40
时间 DCNN 特征	48.01	49.94
时空特征融合	52.18	71.06

为了说明本文方法对每种表情的识别情况，图 4 和图 5 分别给出了时空特征融合在 RML 和 BAUM-1s 数据集上获得最好性能时的模糊矩阵。从图 4 和图 5 可以看出，生气(anger)是最难识别的。其正确识别率分别是 50% 和 44.44%，分析其原因可能是这种表情的特征与其它表情的特征比较相似，容易混淆。

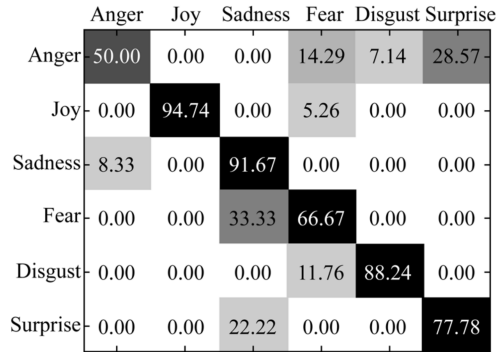

图 4　RML 数据集中时空特征融合取得识别结果的模糊矩阵

Fig. 4　Confusion matrix of the best recognition performance on the RML database

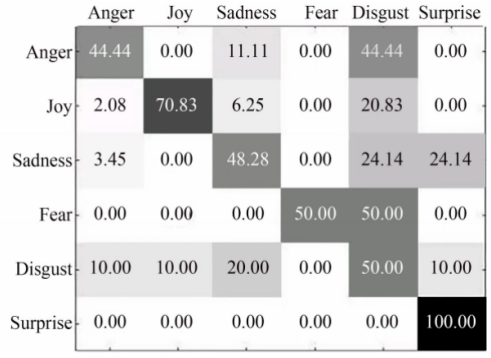

图 5　BAUM-1s 数据集中时空特征融合取得识别结果的模糊矩阵

Fig. 5　Confusion matrix of the best recognition performance on the BAUM-1s database

表 2 列出了本文方法与现有文献报导的实验结果的比较。值得注意的是，所有比较的工作都是采用 LOGO 交叉验证进行的，即保证训练对象是与在测试对象无关的条件下进行测试的最好识别算法的识别结果来比较的。由表 2 可见，本文方法比已有文献报导的识别性能都要好，这说明了本文方法的优势。比如文献[27]采用 3D-CNN 方法用于 RML 中视频表情识别，取得了 68.09% 的正确识别率，而本文方法取得了 71.06% 的正确识别率。这充分说明了本文采用的多模深度卷积神经网络的识别性能要优于 3D-CNN。其中单个时间 DCNN 特征和空间 DCNN 特征的视频表请识别性能在大多数情况下也比 Gabor 小波、LBP 之类的手工特征都要好，这也证明了 DCNN 用于视频表情特征学习的有效性。

表 2　与现有文献报导的结果比较

Tab. 2　Comparisons with the state of the arts

数据集	文献	特征	正确识别率/%
RML	NED Elmadany, et al.,[28]	Gabor 小波	64.58
	Shiqing Zhang, et al.,[27]	3D-CNN	68.09
	ShiqingZhang, et al.,[29]	LBP	56.90
	本文方法	深度时空特征	71.06
BAUM-1s	S Zhalehpour, et al.,[21]	LPQ	45.04
	Shiqing Zhang, et al.,[30]	3D-CNN	50.11
	本文方法	深度时空特征	52.18

4　结　论

本文提出了一种基于多模深度卷积神经网络的视频表情识别模型。本文方法在训练过程主要分为 2 个阶段。首先，在目标视频表情数据集上分别对时空卷积神经网络进行微调，以便学习出具有判别力的时空表情特征。其次，对于学习到的时空特征进行基于 DBN 的特征层融合，并输入到支持向量机实现视频表情的分类。在 RML 和 BAUM-1s 数据集上的实验结果表明了本文方法的有效性。不过，深度学习网络一般具有非常复杂的网络参数，需要消耗很多的计算资源[31]。因此接下来一个研究方向，就是研究如何减少深度模型参数，以便在保证性能的前提下进一步加速深度学习的运算速度。

参考文献：

[1] VENI S, THUSHARA S. Multimodal approach to emotion recognition for enhancing human machine interaction-a survey [J]. *International Journal on Advanced Science, Engineering and Information Technology*, 2017, 7(4): 1428-1433.

[2] ZHAO X, ZHANG S. A review on facial expression recognition: feature extraction and classification [J]. *Iete Technical Review*, 2016, 33(5): 505-517.

[3] HINTON GE, SALAKHUTDINOV RR. Reducing the dimensionality of data with neural networks [J]. *Science*, 2006, 313(5786): 504-507.

[4] 熊昌镇，单艳梅，郭芬红. 结合主体检测的图像检索方法 [J]. 光学 精密工程, 2017, 25(3): 792-798.
XIONG CH ZH, SHAN Y M, GUO F H. Image retrieval method based on image principal part detection [J]. *Opt. Precision Eng.*, 2017, 25(3): 792-798. (in Chinese)

[5] 刘智，黄江涛，冯欣. 构建多尺度深度卷积神经网络行为识别模型 [J]. 光学 精密工程, 2017, 25(3): 799-805.
LIU ZH, HUANG J T, FENG X. Action recognition model construction based on multi-scale deep convolution neural network [J]. *Opt. Precision Eng.*, 2017, 25(3): 799-805. (in Chinese)

[6] ZHAO Z, JIAO L, ZHAO J, *et al.*. Discriminant deep belief network for high-resolution SAR image classification [J]. *Pattern Recognition*, 2017, 61: 686-701.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON GE. ImageNet classification with deep convolutional neural networks[C]. *International Conference on Neural Information Processing Systems*, 2012: 1097-1105.

[8] ESSA IA, PENTLAND AP. Coding, analysis, interpretation, and recognition of facial expressions [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2002, 19(7): 757-763.

[9] ALEKSIC PS, KATSAGGELOS AK. Automatic facial expression recognition using facial animation parameters and multistream HMMs [J]. *IEEE Transactions on Information Forensics & Security*, 2006, 1(1): 3-11.

[10] YANG P, LIU Q, METAXAS DN. Boosting coded dynamic features for facial action units and facial expression recognition [C]. 2007 *CVPR '07 IEEE Conference on Computer Vision and Pattern Recognition*, 2007: 1-6.

[11] NANNI L, LUMINI A, BRAHNAM S. Survey on LBP based texture descriptors for image classification [J]. *Expert Systems with Applications*, 2012, 39(3): 3634-3641.

[12] ZHAO G, AHONEN T, MATAS J, *et al.*. Rotation-invariant image and video description with local binary pattern features [J]. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 2012, 21(4): 1465-1477.

[13] NANNI L, BRAHNAM S, LUMINI A. Local phase quantization descriptor for improving shape retrieval/classification [J]. *Pattern Recognition Letters*, 2012, 33(16): 2254-2260.

[14] KAYAOGLU M, EROGLU ERDEM C. Affect recognition using key frame selection based on minimum sparse reconstruction [J]. *Proceedings of the* 2015 *ACM on International Conference on Multimodal Interaction-ICMI'15*, 2015: 519-524.

[15] ZHAO G, PIETIKÄINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2007, 29(6): 915-928.

[16] KIM BK, ROH J, DONG SY, *et al.*. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition [J]. *Journal on Multimodal User Interfaces*, 2016, 10(2): 1-17.

[17] ACAR E, HOPFGARTNER F, ALBAYRAK S. A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material [J]. *Multimedia Tools & Applications*, 2017, 76(9): 11809-11837.

[18] GAO J, FU Y, JIANG Y G, *et al.*. Frame-transformer emotion classification network [C]. ACM *on International Conference on Multimedia Retrieval*, *Bucharest*, *Romania*, 2017: 78-83.

[19] PANG L, NGO CW. Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos[C]. *the Proceedings of the* 5*th* ACM *on International Conference on Multimedia Retrieval* (ICMR), *Shanghai*, *China*, 2015: 619-622.

[20] WANG Y, GUAN L, VENETSANOPOULOS AN. Kernel cross-modal factor analysis for infor-

mation fusion with application to bimodal emotion recognition [J]. *IEEE Transactions on Multimedia*, 2012,14(3): 597-607.

[21] ZHALEHPOUR S, ONDER O, AKHTAR Z, *et al.*. BAUM-1: a spontaneous audio-visual face database of affective and mental states [J]. *IEEE Transactions on Affective Computing*, 2017, 8(3): 300-313.

[22] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C]. *Nternational Conference on Neural Information Processing Systems*, *Montreal*, *Canada*, 2014: 568-576.

[23] BRUHN A, WEICKERT J, SCHNöRR C. Lucas/kanade meets horn/schunck: combining local and global optic flow methods [J]. *International Journal of Computer Vision*, 2005,61(3): 211-231.

[24] RANJAN R, PATEL VM, CHELLAPPA R. Hyper face: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2016,PP(99): 1-1.

[25] ZHANG S, ZHAO X, CHUANG Y, *et al.*. Learning discriminative dictionary for facial expression recognition [J]. *IETE Technical Review*, 2017,33(5): 1-7.

[26] HINTON GE, OSINDERO S, TEH YW. A fast learning algorithm for deep belief nets [J]. *Neural Computation*, 2006,18(7): 1527-1554.

[27] ZHANG S, ZHANG S, HUANG T, *et al.*. Learning affective features with a hybrid deep model for audio-visual emotion recognition [J]. *IEEE Transactions on Circuits & Systems for Video Technology*, 2017,(99): 1-1.

[28] ELMADANY NED, HE Y, GUAN L. Multiview emotion recognition via multi-set locality preserving canonical correlation analysis[C]. *IEEE International Symposium on Circuits and Systems*, *Montral*, QC, *Canada*, 2016:590-593.

[29] ZHANG S, ZHANG S, HUANG T, *et al.*. Multimodal deep convolutional neural network for audio-visual emotion recognition[C]. *in Proceedings of the 6th ACM on International Conference on Multimedia Retrieval(ICMR)*, *New York*, USA, 2016:281-284.

[30] ZHANG S, ZHANG S, HUANG T, *et al.*. Learning affective features with a hybrid deep model for audio-visual emotion recognition [J]. *IEEE Transactions on Circuits & Systems for Video Technology*, 2018,28(10),1-14.

[31] 李宇,刘学莹,张洪群,等.基于卷积神经网络的光学遥感图像检索 [J].光学 精密工程, 2018,26(1): 200-207.
LI Y, LIU X Y, ZHANG H Q, *et al.*. Optical remote sensing image retrieval based on convolutional neural networks [J]. *Opt. Precision Eng.*, 2018, 26(1):200-207. (in Chinese)

作者简介:

潘仙张(1981—),男,浙江临海人,硕士,高级工程师,2008 年于兰州交通大学获得硕士学位,主要从事图像处理、情感计算、机器学习方面的研究。E-mail: pxz@tzc.edu.cn

张石清(1980—),男,湖南衡阳人,博士,副教授,2012 年于电子科技大学获得博士学位主要从事图像处理、模式识别、情感计算方面的研究。E-mail: tzczsq@163.com