DOI: 10.13245/j.hust.15S1109

基于改进的卷积神经网络的负面表情识别方法

唐 浩[1] 黄伟鹏[2] 李哲媛[1] 刘 宏[1]
(1 北京大学深圳研究生院,广东 深圳 518055;
2 华南理工大学电子与信息学院,广东 广州 510641)

摘要 为了解决传统的基于人工特征的负面表情识别方法在面部无遮挡、姿态非倾斜的人脸表情图像上表现良好,但是在复杂场景下的识别效果较差的问题,提出了一种基于改进的卷积神经网络的负面表情识别方法.首先利用卷积神经网络的无监督特征学习的特性,预训练两个不同拓扑结构的卷积神经网络,用以提取表情特征;然后融合这些特征,训练分类性能更强的支持向量机.改进后的卷积神经网络算法具有较好的鲁棒性和泛化能力,在训练数据库 ICML-fer2013 上取得了 86.2%的识别率,在测试数据库 CK+,GENKI 和 JAFFE 上分别取得了 81.6%,87.0%和 80.8%的识别率.

关键词 负面表情识别;卷积神经网络;无监督特征学习;特征融合;支持向量机
中图分类号 TP391 **文献标志码** A **文章编号** 1671-4512(2015)S1-0457-04

Negative facial expression recognition based on improved convolutional neural networks

Tang Hao[1] *Huang Weipeng*[2] *Li Zheyuan*[1] *Liu Hong*[1]
(1 Shenzhen Graduate School, Peking University, Shenzhen 518055, Guangdong China; 2 School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract The traditional negative facial expression recognition methods based on manual feature perform well on frontal face without facial occlusion, but their performance gets worse in complex condition. To solve this problem, we propose an improved method based on convolutional neural networks (CNN). Firstly, two different architectures of CNN were pre-trained as feature extractors due to CNN's capability of unsupervised feature learning. The feature CNN extracted was then used to train a more powerful classifier: support vector machine. This improved CNN algorithm has better robustness and generalization, achieving recognition accuracy of 86.2% on the training set ICML-fer2013, and 81.6%,87.0%, 80.8% on the testing sets CK+, GENKI, JAFFE respectively.

Key words negative facial expression recognition; convolutional neural network; unsupervised feature learning; feature fusion; support vector machine

人脸表情识别是一个富有挑战性的交叉课题,涉及生理学、心理学、图像处理和计算机视觉等领域.Ekman 和 Friesen 定义了 6 种基本表情:高兴、生气、惊讶、恐惧、厌恶和悲伤[1],并提出了面部动作编码系统(FACS).Suwa 等首先实现了图像序列的自动表情分析[2],标志着人脸表情识别的研究正式进入到计算机视觉领域.

本文着重于负面表情识别的研究.根据效价度和唤醒度理论[3],可将悲伤、生气、恐惧、厌恶归类为负面表情,高兴、惊讶、中性归类为非负面表情.在表情识别中,传统的方法通常基于人工特征,比如局部二值模式(LBP)和方向梯度直方图(HOG).人工特征的设计非常繁琐,且在面部受遮挡、姿态倾斜等复杂情况下的效果较差,无法满

收稿日期 2015-06-30.
作者简介 唐 浩(1989-),男,硕士研究生, E-mail: haotang@sz.pku.edu.cn.
基金项目 国家自然科学基金资助项目(60875050,60675025).

足人机交互中的鲁棒性要求. 近几年,深度学习快速发展,特别是卷积神经网络在图像识别领域取得了巨大的成功,它是为识别二维图像而特殊设计的,具有无监督特征学习的能力.

受此启发,在此提出一种基于卷积神经网络的改进方法:首先,预训练两个不同拓扑结构的卷积神经网络;然后,分别用预训练的两个卷积神经网络提取人脸表情图像的特征;最后,融合表情特征,训练分类性能更强的分类器,取得了优于传统方法的结果.

1 卷积神经网络

卷积神经网络(CNN)是近年发展起来的,在图像识别等领域得到广泛应用的一种深度学习方法,它直接以二维图像为输入,且具有自动学习特征的能力. 图 1 是本研究采用的一个含有 3 个卷积层的 CNN 结构,C_1,C_2,C_3 表示卷积层,分别采用 16,32,32 个卷积核,卷积核大小分别为 5×5,3×3,3×3,每个卷积核与上一层的所有特征图

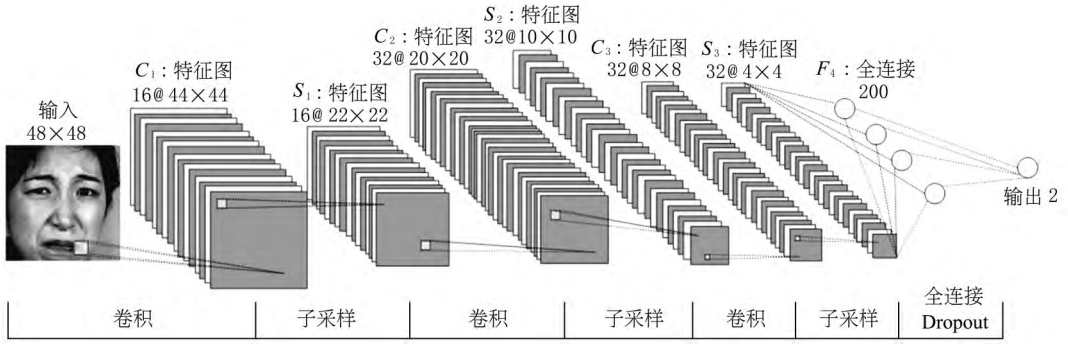

图 1 基于卷积神经网络的负面表情识别-拓扑结构

进行连接. 下面以该结构为例,对卷积层、子采样层和全连接层进行说明.

1.1 卷积层

自然图像有其固有特性:在从某一图像子块上学习到一些特征后,可将这些特征作为探测器,应用到所有子块中去,获得不同子块的激活值. CNN 中的卷积也是利用图像的这种固有特性,具体做法是:卷积层中一个可训练的卷积核与上一层中不同组合的特征图进行卷积,加上偏置得到当前层的特征图. 该过程可以用下式表示

$$x_j^l = \sum_{i\in M_j} y_i^{l-1} \otimes k_{ij}^l + b_j^l, \tag{1}$$

式中:x_j^l 为第 l 层第 j 个特征图的输入;y_i^{l-1} 为第 $l-1$ 层第 i 个特征图的输出;k_{ij}^l 为前一层第 i 个特征图与当前层第 j 个特征图之间的卷积核;b_j^l 为第 l 层第 j 个特征图的偏置;$i\in M_j$ 为前一层中与当前层第 j 个特征图有连接的所有特征图.

1.2 子采样层

通过卷积层后,特征图的个数增加,使得特征维数快速上升,为了避免陷入维数灾难,可在卷积层后加入子采样层. 子采样层可以在保留原始特征信息的条件下,极大地降低特征维数,并且具有平移不变性等优点,其过程可以用下式表示

$$x_j^l = f(\beta_j^l \mathrm{down}(x_j^{l-1}) + b_j^l), \tag{2}$$

式中:$\mathrm{down}(x_j^{l-1})$ 为对第 $l-1$ 层第 j 个特征图进行子采样;β_j^l 为乘性偏置;b_j^l 为加性偏置;$f(*)$ 为激活函数;x_j^l 为第 l 层第 j 个特征图.

图 1 中 S_1,S_2 和 S_3 表示子采样层,采样单元大小均为 2×2,采样方式为最大子采样.

1.3 全连接层

全连接层上的每一个神经单元,均与上一层特征图中的所有神经单元互相连接. 每一个神经单元的输出可以用下式表示

$$\boldsymbol{h}_{W,b}(\boldsymbol{x}) = f(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x} + b), \tag{3}$$

式中:$\boldsymbol{x}$ 为神经元的输入;$\boldsymbol{h}_{W,b}(\boldsymbol{x})$ 为神经元的输出;$\boldsymbol{W}$ 为连接权重;b 为偏置;$f(*)$ 为非线性激活函数.

常用的非线性激活函数有 Sigmoid 和 Tanh,它们有时会导致梯度消失的问题. 为了克服该问题,采用修正线性单元 ReLU (Rectified linear unit)[4],ReLU 在很多深度网络结构中被采用,其表现通常优于其他激活函数.

图 1 中 F_4 采用 Softmax 全连接,含 200 个神经单元,激活函数为 ReLU,使用 dropout[5],其作用是防止过拟合,提高网络的泛化能力.

2 改进的卷积神经网络

图 1 中设计的 CNN 结构在负面表情识别的性能上已经明显优于其他传统方法(见下文实验结果),但是考虑到其分类层 Softmax 的分类性能不够理想,对其做了一些改进.

改进的CNN如图2所示，预训练了两个不同的CNN模型，左边是一个包含3个卷积层的CNN（下文称3-CNN），右边是一个包含4个卷积层的CNN（下文称4-CNN）. 3-CNN的网络参数选取训练误差最小时的参数，4-CNN的网络参数选取测试误差最小时的参数. 因为训练集和测试集来自不同的表情数据库，所以这样的拓扑结构可以兼顾更多的复杂情况，使得模型的泛化能力更好. 另外，两个网络的拓扑结构不同，可以学习并提取到更多样的表情特征.

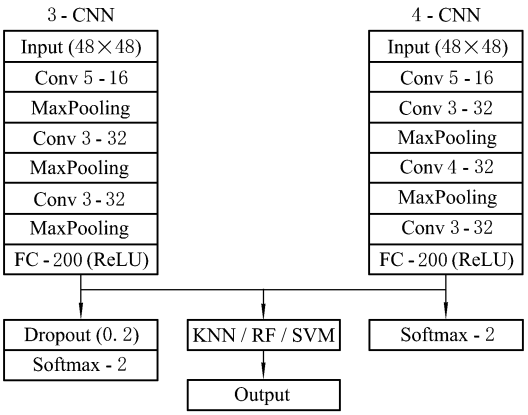

图2　改进的卷积神经网络

预训练完毕后，将训练图像分别输入到3-CNN和4-CNN，在3-CNN和4-CNN的全连接层后分别输出一个200维的特征向量，将这两个200维的特征向量拼接得到一个400维的特征向量，作为其他分类器的输入，训练得到一个分类性能更强的模型. 在实验中对比了KNN，Random Forest和SVM三种分类器，发现SVM是三者之中分类性能最好的（见下文实验结果）.

3　实验及分析

3.1　实验平台与数据库

实验的硬件平台为戴尔的OPTIPLEX 9020：Core i7处理器，8 GHz内存. 软件平台为深度学习框架Theano，采用Python语言.

本实验研究二分类问题：负面表情和非负面表情，负面表情包含生气、悲伤、恐惧和厌恶，非负面表情包含高兴、惊讶和中性. 实验所采用的训练数据库为ICML-fer2013[6]，含有上述7种表情，共35 886张图片，图片来自互联网，场景各异，部分图片面部受遮挡、姿态倾斜. 在该数据库上进行表情识别具有挑战性，但是更加接近真实场景，因此本实验采用该数据库. 从中随机选取30 000张图片作为训练集，剩下的5 886张作为验证集. 另外，为了评估模型的泛化能力，用训练好的模型对CK+[7]，GENKI[8]和JAFFE[9]这三个人脸表情数据库进行测试. CK+含有123个人的7种表情序列，从中选取123个人的2 095张表情图片. GENKI含有4 000张表情图片，分为smile和non-smile，从中选取1 997张smile图片作为正面表情图片. JAFFE含有10位女性的213张表情图片.

3.2　对比实验说明

在上述四个数据库上，基于CNN（图1）和改进CNN（图2）做了多次实验. 此外，为了评估CNN和改进CNN的性能，设置了两个对比实验，采用的方法分别是基于LBP特征和基于多层感知机. 下面对这两个对比实验进行简单说明.

局部二值模式（LBP）是一种用于提取图像局部纹理特征的算法，具有旋转不变性和灰度不变性等显著的优点，已被广泛应用于人脸分析. 本实验采用LBP等价模式，尝试了不同的算子（LBP_8^1，LBP_8^2，LBP_{16}^2），分类器采用线性SVM.

多层感知机（MLP）是一种前向结构的人工神经网络，除了输入输出层，中间可以有多个隐藏层，且层与层之间全连接. 本实验尝试了隐藏层分别为2层、3层、4层的MLP结构.

3.3　结果分析

用上述四种方法在数据库上进行测试. 表1给出不同LBP算子的识别率，分类器采用线性SVM. 从表1可以得出：当采用LBP_{16}^2算子时，在训练集ICML-fer2013上取得最高的识别率73.7%，但在三个测试集上的识别率未达到最高；三种LBP算子在测试集上的泛化性能都很一般.

表1　不同LBP算子的识别率　　%

算法	数据库			
	ICML-fer2013	JAFFE	GENKI	CK+
LBP_8^1-SVM	61.6	61.5	73.0	66.6
LBP_8^2-SVM	70.1	59.6	77.3	62.0
LBP_{16}^2-SVM	73.7	61.1	76.9	62.4

表2给出不同MLP结构的识别率，从表2可以得出：当网络的隐藏层层数增加时，在训练集

表2　不同MLP结构的识别率　　%

隐藏层数（各层节点数）	数据库			
	ICML-fer2013	JAFFE	GENKI	CK+
2（600-300）	71.7	66.7	54.0	79.5
3（1 200-600-300）	68.3	67.2	54.1	76.7
4（600-1 200-600-300）	69.1	69.5	55.9	81.1

上的识别率变化不大,但是在测试集上的识别率有所提高.

CNN 与改进 CNN 的识别率见表 3,其中改进 CNN 的分类层采用以下三种:K 近邻,$k=100$;随机森林,决策树个数为 500,采用 Gini 指数;支持向量机,采用径向基核函数. 从表 3 可得出:当分类器采用支持向量机时,在训练集上取得最高的识别率(86.2%),并且在 JAFFE 和 GENKI 数据库上的识别率也达到最高(80.8%和 87.0%),在 CK+上的识别率(81.6%)与采用其他两种分类器时的识别率非常接近,因此支持向量机的综合性能优于 K 近邻和随机森林.

表 3 CNN 及改进 CNN 的识别率 %

算法	数据库			
	ICML-fer2013	JAFFE	GENKI	CK+
CNN	75.0	70.7	75.3	81.5
CNN-KNN	82.8	76.6	86.4	82.7
CNN-RF	82.0	77.0	84.2	82.6
CNN-SVM	86.2	80.8	87.0	81.6

图 3 对比了表 1～表 3 的结果,可明显看出:CNN 在各数据库上的识别率高于传统的基于 LBP 特征、MLP 算法的识别率,体现出 CNN 优越的性能和特征学习能力. 而改进的 CNN 在保留 CNN 特征学习能力的基础上,进一步增强了分类器的分类性能,得到更高的识别率. 当分类器采用支持向量机时,在四个数据库上的总体识别性能最好.

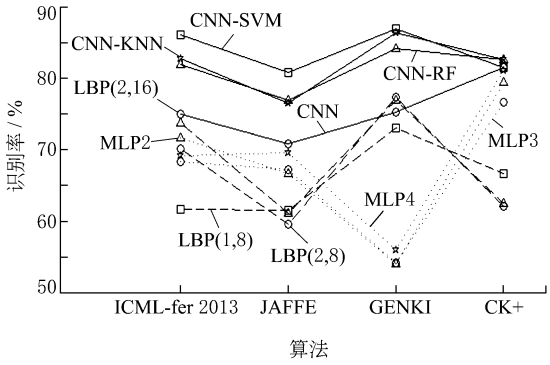

图 3 不同算法的识别率对比

4 结论

本文提出了一种基于改进的卷积神经网络的负面表情识别方法,在包含复杂场景、姿态、光照和面部遮挡的训练数据库ICML-fer2013上取得了较高的识别率,并在 JAFFE,GENKI 和 CK+数据库上取得了较好的泛化性和鲁棒性. 与传统的特征提取方法相比,本文提出的利用 CNN 提取特征的方法具有一定的优势,在不同的数据库上均取得了高于传统方法的识别率. 本研究后续的工作是优化 CNN 的拓扑结构,在提高识别率的同时,降低网络复杂度.

参 考 文 献

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

[2] Suwa M, Sugie N, Fujimora K. A preliminary note on pattern recognition of human emotional expression [C]// Proc of International Joint Conference on Pattern Recognition. New York: IEEE, 1978: 408-410.

[3] Sun K, Yu J, Huang Y, et al. An improved valence-arousal emotion space for video affective content representation and recognition[C]// Proc of IEEE International Conference on Multimedia and Expo. New York: IEEE, 2009: 566-569.

[4] Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2013: 8609-8613.

[5] Hinton G E, Srivastava N. Improving neural networks by preventing co-adaptation of feature detectors [DB/OL]. [2015-04-30]. http://arxiv.org/abs/1207.0580.

[6] Goodfellow I J, Erhan D, Carrier P L, et al. Challenges in representation learning: a report on three machine learning contests[J]. Neural Information Processing, 2013, 23(1): 117-124.

[7] Lucey, Patrick, Cohn J F, et al. The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression[C]// Proc of IEEE Computer Society Conference on CVPR Workshops. New York: IEEE, 2010: 94-101.

[8] Movellan J R. The MPLab GENKI database, GENKI-4K subset [DB/OL]. [2015-04-30]. http://mplab.ucsd-edu.

[9] Lyons M J, Shigeru Akemastu, Miyuki Kamachi. Coding facial expressions with gabor wavelets[C]// Proc of 3rd IEEE International Conference on Automatic Face and Gesture Recognition. New York: IEEE, 1998: 200-205.