用于人脸表情识别的多分辨率特征融合卷积神经网络

何志超, 赵龙章, 陈闯

南京工业大学电气工程与控制科学学院, 江苏 南京 211816

摘要 在人脸表情识别任务中,传统的机器学习方法是基于人工来提取特征,其特征提取过程时间复杂度高且稳健性差,而现有依赖单通道卷积核的卷积神经网络提取特征不够充分,进而导致识别率不高。针对这些问题,提出一种多分辨率特征融合的卷积神经网络。利用两个相互独立且深度不同的通道对图片进行特征提取,使卷积神经网络自主学习同一图像下不同分辨率的特征,然后将不同分辨率的特征送入全连接层并进行特征融合,最后经过softmax分类器进行表情分类。在JAFFE和CK+表情数据库上进行了多次实验,结果表明,与传统的机器学习方法和现有的卷积神经网络结构相比,所提卷积神经网络结构模型具有稳健性好、泛化能力强、收敛速度快的优点。

关键词 机器视觉; 人脸表情识别; 特征提取; 卷积神经网络; 多分辨率特征融合

中图分类号 TP391.4 **文献标识码** A **doi**: 10.3788/LOP55.071503

Convolution Neural Network with Multi-Resolution Feature Fusion for Facial Expression Recognition

He Zhichao, Zhao Longzhang, Chen Chuang

College of Electrical Engineering and Control Science, Nanjing Tech University, Nanjing, Jiangsu 211816, *China*

Abstract In facial expression recognition, the traditional machine learning methods based on the manual feature extraction are time-consuming and less robust. The current convolution neural networks relying on single channel convolution kernel are not sufficient to extract feature, which makes the recognition rates low. We propose a multi-resolution feature fusion convolution neural network, which is combined with two uncorrelated and channels with different depths to extract multi-resolution features. After fusing the two channels feature, a softmax classification is used to classify the facial expression. The experiments on JAFFE and CK+ facial expression databases show that compared with traditional machine learning methods and existing convolution neural networks, the proposed convolution neural network structure model has the advantages of good robustness, strong generalization ability, and fast convergence speed.

Key words machine vision; facial expression recognition; feature extraction; convolution neural network; multi-resolution feature fusion

OCIS codes 150.1135; 100.4996

1 引 言

作为一个涉及到心理学、计算机视觉、机器学习等领域的交叉性课题,人脸表情识别一直以来都是研究的热点,并且在人机交互[1]、虚拟现实[2]、安防监控[3]、安全驾驶[4]等领域有着广泛的应用前景。由于实时场景中光照、非正面、遮挡等种种复杂因素给表情识别带来了很大的困难[5],所以人脸表情识别是一项富有挑战性的任务。

近年来随着深度学习的快速发展,作为一种典型的深度学习算法,卷积神经网络(CNN)[6]已被广泛应用于图像超分辨率重建[7]、行人检测[8]、遥感影

收稿日期: 2017-12-11; **收到修改稿日期**: 2018-01-02

基金项目: 国家自然科学基金(51277028)

作者简介: 何志超(1992—),男,硕士研究生,主要从事图像处理、机器学习等方面的研究。E-mail: 453122670@qq.com

导师简介: 赵龙章(1961—),男,博士,教授,主要从事模式识别与机器学习、等离子体等方面的研究。E-mail: 3402594645@qq.com

像分类[9]、大尺度测量[10]等领域。张昭旭等[11]利用AlexNet模型对表情图像进行特征提取,并使用支持向量机进行分类,实验结果表明CNN在提取图像本质特征方面有其他方法不可比拟的效果。陈向震等[12]采用小尺度的卷积核更加细致地提取局部特征,并借助两个连续的卷积层增加模型的非线性表达能力,实验结果表明该方法在人脸表情分析与识别上具有良好的性能和实用性。唐浩等[13]利用CNN进行无监督特征学习,在不同的拓扑结构上提取表情特征,将特征融合后使用支持向量机进行分类,该方法解决了在面部遮挡、姿态倾斜等复杂场景下识别效果较差的问题。Wang等[14]提出从卷积的中间层提取不同尺度的特征,以结合局部信息与全局信息,同时采用三元组损失学习方法对该结构进行训练,该方法提高了模型的辨识能力。Cui等[15]使用多重CNN集成的方法进行表情识别,首先通过脸部图像中的关键点将面部对齐,然后裁剪出眼睛和嘴巴区域,并分别对这三个区域训练CNN,最后将这三个CNN集成为一个面部表情识别的CNN,实验表明在表情识别任务中该方法具有更好的性能。

本文提出一种多分辨率特征融合的CNN,通过两个相互独立且卷积层深度不同的网络对图片进行特征提取,然后将提取到的不同分辨率特征送入全连接层进行特征融合,最后将融合后的特征经过softmax分类器进行表情分类。多分辨率融合旨在使模型同时具备不同分辨率特征的学习能力,进而增加模型的识别率并提高模型泛化能力。实验证明,该方法能够降低表情识别的误差,与其他CNN的优化方法相比具有更好的泛化能力和更快的收敛速度。

2 CNN介绍

CNN主要包括卷积层、池化层、激活函数、全连接层。

一般与输入图像直接连接的是卷积层,使用不同的卷积核将输入转变为特征图并传给下一层。卷积层的计算过程可以表示为

$$\boldsymbol{I}_j^l=\sum_i \boldsymbol{I}_i^{l-1}\otimes \boldsymbol{k}_{ij}^{l-1}+\boldsymbol{b}_j^l, \tag{1}$$

式中:$\boldsymbol{I}_j^l$ 和 $\boldsymbol{I}_i^{l-1}$ 分别为第 l 层第 j 个特征图和第 $l-1$ 层第 i 个特征图;$\boldsymbol{k}_{ij}^{l-1}$ 为从 $\boldsymbol{x}_i^{l-1}$ 到 $\boldsymbol{x}_j^l$ 的卷积核;$\otimes$ 为卷积运算;$\boldsymbol{b}_j^l$ 为相应的偏置。

通常激活函数和池化层会连接在卷积层后面。在层数比较深的神经网络中,一般使用线性激活函数(ReLU)[16]作为激活函数,本研究中用到的激活函数也全部是ReLU函数。ReLU激活函数表达式为

$$\mathrm{ReLU}(x)=\begin{cases}x, & x>0\\ 0, & x\leqslant 0\end{cases}, \tag{2}$$

式中:x 为卷积层输出特征图的像素值。

池化层的一般表达式为

$$\boldsymbol{Z}_j^l=\mathrm{down}(\boldsymbol{Y}_j^l), \tag{3}$$

式中:$\boldsymbol{Y}_j^l$ 和 $\boldsymbol{Z}_j^l$ 分别为池化层的输入与输出;down(·)表示池化函数,常用的池化函数有最大池化函数和均值池化函数,本研究中的池化函数均为最大池化函数。

全连接层一般在最后将前面提取到的特征图进行全连接,其中每一个神经节点的输出可以表示为

$$h(\boldsymbol{I}^L)=f(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{I}^L+\boldsymbol{b}), \tag{4}$$

式中:$h(\cdot)$ 表示全连接层的输出;L 为CNN中卷积层的总深度;$\boldsymbol{I}^L$ 为第 L 层卷积层的输出;$\boldsymbol{W}$ 为全连接权重;$\boldsymbol{b}$ 为偏置;$f(\cdot)$ 表示激活函数。

3 多分辨率特征融合CNN

3.1 传统CNN的不足

由于不同卷积层输出的特征图分辨率不一样,较浅层的特征图分辨率较高,对于较小的特征信息比较敏感,但对物体整体性特征的表达欠缺;较深层的特征图分辨率较低,这类特征比较抽象,对于物体的轮廓和外观等方面表达准确,但是对于物体纹理等细节方面的特征不敏感。传统CNN只使用最后一层特征图作为全连接层的输入,而在网络结构较深时,最后一层特征图分辨率往往很低,所以提取到的主要是一些比较抽象的特征,难以提取到一些细小的特征。而在人脸表情识别任务中,不仅需要外观和整体轮廓等比较抽象的特征,也需要一些类似于纹理等细节方面的特征,因而使用传统的CNN结构会使得特征提取不充分,进而影响结果的准确性。

3.2 多分辨率特征融合的具体方法

针对3.1节提出的传统CNN的问题,提出一种可以对不同分辨率的特征进行融合的CNN网络结构,该结构可以同时利用不同分辨率的特征信息进行表情分类。具体的结构如图1所示。

该结构由两个相互独立且卷积层深度不同的卷积网络组成,这两个网络分别有2层和5层卷积层。图1中输入为256 pixel×256 pixel的图像,左边部分经过5层卷积层得到128个分辨率为4 pixel×

4 pixel的特征图，右边部分经过 2 层卷积层得到 128 个分辨率为 32 pixel×32 pixel 的特征图，将不同分辨率的特征图与下一层进行全连接特征融合，最后使用 softmax 分类器将融合的特征用于分类。

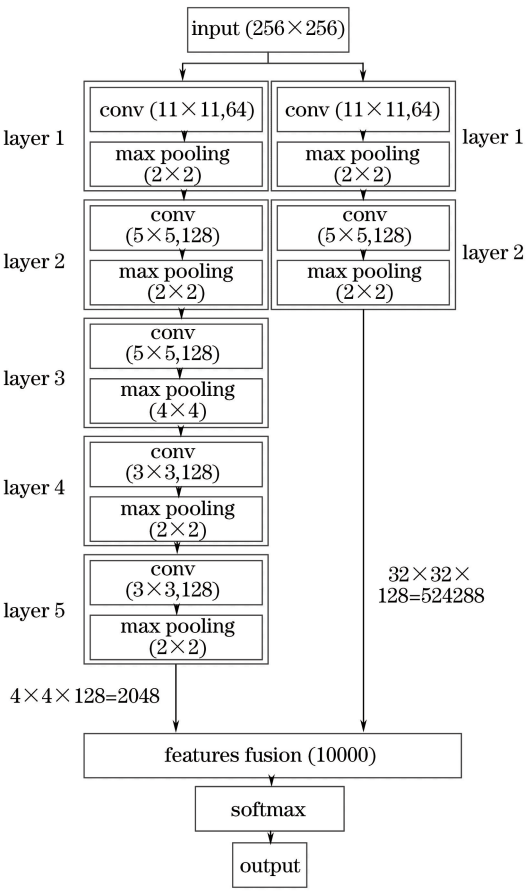

图 1 CNN 结构框图

Fig. 1 Architecture of CNN

3.3 多分辨率特征融合 CNN 的优势

提出的 CNN 结构具有以下优势：一是通过两个相互独立且卷积层深度不同的网络结构得到不同分辨率的特征图，该结构同时具备提取较抽象和较细节特征的学习能力，使其对于特征的提取更加充分，因而能够有效地降低表情识别的误差；二是由于所提 CNN 结构具有更强的特征提取能力，因而学习到的特征能够更好地进行表情识别任务，在训练阶段也会更快达到收敛。

4 实验结果及分析

为了验证在表情识别中多分辨率特征融合 CNN 模型的应用效果，在 CK＋和 JAFFE 表情数据库上进行了多组实验。实验是在基于 Python 的深度学习框架 TensorFlow 上进行的，使用的操作系统为 Ubuntu 14.04，硬件配置如下：中央处理器（CPU）为 Intel XeonE5-2643 v2，主频 3.50 GHz；内存为 16 GB；图形处理器（GPU）为 NVIDIA GeForce GTX980Ti，显存为6 GB。

4.1 数据库选择和数据预处理

CK＋和 JAFFE 人脸表情数据库在人脸表情识别领域应用广泛，许多人脸表情识别研究成果均在该两个数据库上进行验证。因此实验选用 CK＋与 JAFFE 人脸表情数据库，两个数据库都有 7 种表情：高兴、悲伤、愤怒、恐惧、惊讶、厌恶、中性。CK＋数据库是卡内基梅隆大学于 2010 年发布的，由 10730 张 640 pixel×490 pixel 的 png 格式的灰度图组成。它包含了不同类型人的不同表情，图 2 所示为 CK＋数据库的部分图片。JAFFE 数据库由 216 张256 pixel×256 pixel 的 tiff 格式灰度图组成，它的样本来自于日本女性。图 3 为 JAFFE 数据库的部分图片。

图 2 CK＋数据库的部分图片

Fig. 2 Partial pictures of CK＋ database

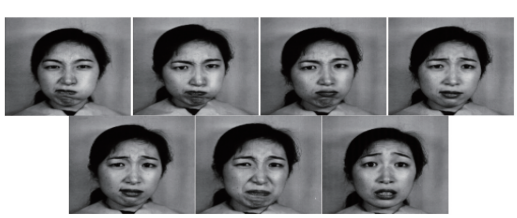

图 3 JAFFE 数据库的部分图片

Fig. 3 Partial pictures of JAFFE database

实验中采用 5 折交叉验证方法，将 CK＋人脸表情数据集随机平均分成 5 份，取其中 4 份数据集作为训练样本，1 份作为测试样本，共进行 5 次训练和测试。选取 JAFFE 数据库中所有的 216 张图片作为测试集用于验证模型的泛化性能。训练样本和测试样本设计见表 1。

表 1 训练样本和测试样本设计

Table 1 Design of training samples and test samples

Database	Number of training samples	Number of test samples
CK＋	8584	2146
JAFFE	—	216

在训练分类器之前，先对数据库图片进行预处

理。图 1 设计的 CNN 其输入图片分辨率为 256 pixel×256 pixel，由于 CK＋和 JAFFE 数据库的分辨率不同，因此使用最近邻插值[17]的方法将数据库的所有图片分辨率统一为 256 pixel × 256 pixel。

4.2 实验过程与结果

实验中对 CK＋数据库采用 5 折交叉验证方法，取 5 次测试识别率的平均值作为测评算法优劣的指标，并使用该指标对各个算法的性能进行对比与分析，以此说明实验结果的可靠性。

首先通过实验验证了多分辨率特征融合 CNN 与传统的 CNN 结构相比具有更好的特征提取能力与识别能力；然后将文献[11，14-15]的方法与多分辨率特征融合 CNN 进行对比，以说明了多分辨率特征融合 CNN 算法的优越性。

使用不同卷积层深度的传统 CNN 进行实验，其中卷积层数从 1 到 7 层。在数据集上进行交叉验证，识别率见表 2。从表 2 中可知，随着卷积层数的增加，网络的识别率和泛化能力在提升；特别地，当卷积层数达到 6 和 7 时，识别率提升幅度明显减小，说明 5 层是一个较佳的卷积层数，这时当前网络结构对人脸表情识别模型的表征已经达到了瓶颈，因而再增加卷积层数，对识别结果的影响不大，反而会增加网络的复杂度，影响计算机运行效率。

表 2 不同卷积层数 CNN 的识别率

Table 2 Recognition rate of CNN with different convolution layers

Number of convolution layers	Recognition rate /%	
	CK＋	JAFFE
1	67.3	62.5
2	75.2	70.9
3	79.1	75.8
4	82.3	81.2
5	84.2	83.8
6	84.4	84.1
7	84.5	84.3

采用多分辨率特征融合 CNN 进行实验。根据前面的分析，卷积层数为 5 的 CNN 在人脸表情识别人中具有很强的特征提取与表情分类能力，所以实验中将 5 层卷积层与其他 1～4 层不同卷积层数提取的特征进行全连接融合，全连接融合的节点数都取 10000，识别率见表 3。从表 2 和表 3 可知，采用多分辨率特征融合 CNN 后，人脸表情的识别率相比表 3 中的传统 CNN 的识别结果有了较大的提升，这是由于多分辨率特征融合 CNN 通过不同深度的卷积层得到分辨率不同的特征图，获得比一般 CNN 结构更加丰富的特征信息，进而有效提高 CNN 对人脸表情识别的准确率。其中，卷积层数为 5 层和 2 层的搭配得到的识别效果最好。

表 3 不同分辨率特征融合的 CNN 识别率

Table 3 Recognition rate of CNN with different resolution features

Number of two convolution layers	Recognition rate /%	
	CK＋	JAFFE
5-1	85.9	85.2
5-2	92.1	91.7
5-3	89.6	89.5
5-4	85.1	83.9

将文献[11，14-15]的方法与多分辨率特征融合 CNN 进行对比，其中传统的 CNN 使用卷积层数为 5 层的结构，多分辨率特征融合 CNN 使用卷积层数为 5 层和 2 层搭配的结构，识别率见表 4。由表 4 可知，与文献[11，14-15]的方法相比，多分辨率特征融合 CNN 人脸表情的识别率明显高于其他改进的 CNN 方法。这是因为多分辨率特征融合 CNN 将图片输入到两个互相独立且深度不同的卷积层，通过卷积层的特征提取，得到分辨率不同的特征图，可以获得比一般 CNN 结构更加丰富的特征信息，因而使用全连接层融合这些特征对图片进行分类，能够大大提高 CNN 对人脸表情识别的准确率。而文献[14]提取的是同一个卷积通道的不同中间层特征，由于每一层特征图是由上一层卷积层得到的，所以不同层的特征图之间所包含的信息会有重复，因而其提取的特征之间并不是独立的，这不仅会导致模型的过拟合，而且会使得全连接层的参数量急剧增加，使其识别率远低于多分辨率特征融合 CNN 算法。文献[15]单独提取出眼睛和嘴巴区域进行识别分类，这样会丢失大量其他区域的信息，因而其识别效果无法得到很大的提升。

表 4 不同算法的识别率

Table 4 Recognition rates of different algorithms

Algorithm	Recognition rate /%	
	CK＋	JAFFE
Ref. [11]	79.6	78.2
Ref. [14]	83.0	79.7
Ref. [15]	85.6	85.3
Proposed method	92.1	91.7

为了进一步说明多分辨率特征融合 CNN 的优越性，图 5 给出了上述改进 CNN 方法在训练过程中损失函数的变化过程。由于文献[15]方法是训练

三个 CNN，因而其训练迭代次数明显多于其余几种方法，所以在此只比较文献[11]、文献[14]和多分辨率特征融合 CNN 算法。从图 5 可以看出，随着迭代次数的增加，多分辨率特征融合 CNN 结构能较快达到收敛，且最终的损失函数值更小。

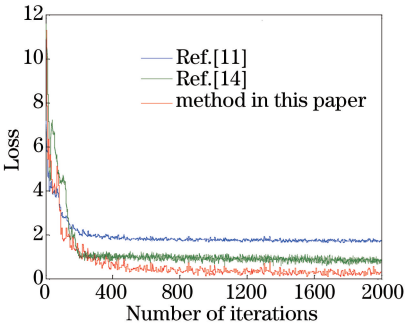

图 4 训练过程中的损失函数

Fig. 4 Loss function during training

5 结 论

提出了一种多分辨率特征融合的 CNN，该网络将图片通过两个相互独立且卷积核分辨率以及卷积层深度不同的通道进行特征提取，使网络同时学习不同分辨率的特征，然后将不同分辨率的特征送入全连接层进行特征融合，最后将融合后的特征经过 softmax 分类器进行表情分类。在不同的数据库上将所提 CNN 与多种改进的 CNN 网络进行对比实验，结果表明多分辨率特征融合 CNN 具有稳健性好、泛化能力强、收敛速度快的优点。

参 考 文 献

[1] Berns K, Hirth J. Control of facial expressions of the humanoid robot head roman [C] // International Conference on Intelligent Robots and Systems, October 9-15, 2006, Beijing, China. IEEE, 2006: 3119-3124.

[2] Bartlett M S, Littlewor T G, Fasel I, *et al*. Real time face detection and facial expression recognition: development and applications to human computer interaction [C] // Computer Vision and Pattern Recognition Workshop, June 16-22, 2003, Madison, Wisconsin, USA. IEEE, 2003: 53.

[3] Ong Z, Ni B, Guo D, *et al*. Learning universal multi-view age estimator using video context [C] // International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. IEEE, 2012: 241-248.

[4] Vural E, Cetin M, Ercil A, *et al*. Automated drowsiness detection for improved driver safety comprehensive databases for facial expression analysis [C] // International Conference on Automotive Technologies, 2008, Istanbul, Turkey. Sanbanci University Research Database, 2008.

[5] Hu J G. Children′s emotional competence evaluate system based on facial expression recognition [D]. Nanjing: Southeast University, 2015.
胡建国. 基于表情识别的儿童情绪能力评测系统[D]. 南京: 东南大学, 2015.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Neural Information Processing Systems 2012. Neural Information Processing Systems Foundation, Inc, 2012: 1097-1105.

[7] Li S M, Lei G Q, Fan R. Depth map super-resolution reconstruction based on convolutional neural networks [J]. Acta Optica Sinica, 2017, 37(12): 1210002.
李素梅, 雷国庆, 范如. 基于卷积神经网络的深度图超分辨率重建[J]. 光学学报, 2017, 37(12): 1210002.

[8] Ye G L, Sun S Y, Gao K J, *et al*. Nighttime pedestrian detection based on faster region convolution neural network [J]. Laser & Optoelectronics Progress, 2017, 54(8): 081003.
叶国林, 孙韶媛, 高凯珺, 等. 基于加速区域卷积神经网络的夜间行人检测研究[J]. 激光与光电子学进展, 2017, 54(8): 081003.

[9] Chen Y, Fan R S, Wang J X, *et al*. High resolution image classification method combining with minimum noise fraction rotation and convolution neural network [J]. Laser & Optoelectronics Progress, 2017, 54(10): 102801.
陈洋, 范荣双, 王竞雪, 等. 结合最小噪声分离变换和卷积神经网络的高分辨影像分类方法[J]. 激光与光电子学进展, 2017, 54(10): 102801.

[10] Li T T, Yang F, Xu X L. Method of large-scale measurement based on multi-vision line structured light sensor[J]. Chinese Journal of Lasers, 2017, 44(11): 1104003.
李涛涛, 杨峰, 许献磊. 基于多视觉线结构光传感器的大尺度测量方法[J]. 中国激光, 2017, 44(11): 1104003.

[11] Zhang Z X. Research on feature extraction based on CNN deep learning model [J]. Morden Computer, 2016 (3): 41-44.
张昭旭. CNN 深度学习模型用于表情特征提取方法

探究[J]. 现代计算机, 2016 (3): 41-44.

[12] Chen X Z. Research of facial expression recognition algorithm based on deep learning [D]. Shenyang: Shenyang University of Technology, 2016.
陈向震. 基于深度学习的人脸表情识别算法研究[D]. 沈阳: 沈阳工业大学, 2016.

[13] Tang H, Huang W P, Li Z Y, *et al*. Negative facial expression recognition based on improved convolutional networks [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2015, 43(s1): 457-460.
唐浩, 黄伟鹏, 李哲媛, 等. 基于改进的卷积神经网络的负面表情识别方法[J]. 华中科技大学学报(自然科学版), 2015, 43(s1): 457-460.

[14] Wang J, Yuan C. Facial expression recognition with multi-scale convolution neural network [C] // Pacific Rim Conference on Multimedia, September 15-16, 2016, Xi′ an, China. New York: Springer International Publishing, 2016: 376-385.

[15] Cui R, Liu M, Liu M. Facial expression recognition based on ensemble of mulitple CNNs [C] // Chinese Conference on Biometric Recognition, October 14-16, 2016, Chengdu, China. New York: Springer International Publishing, 2016: 511-518.

[16] Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVSCR using rectified linear unit and dropout [C] // Acoustics, Speech and Signal Processing, May 26-31, 2013, Vancouver, BC, Canada. IEEE, 2013: 8609-8613.

[17] Jiang N, Wang L. Quantum image scaling using nearest neighbor interpolation [J]. Quantum Information Processing, 2015, 14(5): 1559-1571.