基于可识别身份卷积神经网络的人脸表情识别

摘要：人脸表情识别是计算机视觉中的一个热点课题，但由于个体性差异造成的面部表情的不一致性，因此该课题仍颇具挑战性。为解决这一挑战，同时受最近深度身份网络（DeepID-Net）人脸识别成功的启发，本文提出了一种新颖的深度学习框架来实现对人脸图像的表情识别。相比于现有的深度学习方法，我们提出的基于多尺度全局图像和局部面部斑块的框架显著提高了面部表情识别性能。最后，我们在公开的基准数据集 JAFFE 和拓展 Cohn-Kanade（CK+）上进行实验，并验证了其有效性。

关键词：面部表情识别、深度学习、分类问题、身份可识别

DOI: 10.21629/JSEE.2017.04.18

1 引 言

面部表情是情绪计算（识别、解释、处理和模拟人类情感）的重要线索，对人机交互有重要意义。因此，通过面部图像识别人类表情的问题（称为面部表情识别）是计算机视觉领域活跃的研究课题之一，具有学术和社会意义。典型的面部表情识别协议，首先是检测人脸，然后提取图像特征，接着用分类模型将这些特征分类为预定义的表情类。因此，该问题可被归为由一个分类框架从低级图像中学习一个分类函数，进行标签分类（如愤怒、厌恶、恐惧、快乐、悲伤、惊讶、中性）的问题。

在现有的面部表情识别框架中，鲁棒不变特征对识别性能有非常重要的影响；但在过去 3-5 年深度学习的兴起之后，深度卷积神经网络（CNNs[1-3]）的端到端学习方法被认为是自动发现分类/模式识别的最佳图像特征中最先进的方法。在许多视觉识别问题中（如物体分类[1,4]和场景理解[5-7]），其训练过程通常是将整个图像域与其相应的类标签一起输入深度学习框架。本次实验中我们将揭示，先在预处理阶段检测某人的特定局部，再进行 CNN 深度训练，我们就能实现更高的性能。我们将此方法命名为 identity-inspired CNN（I2CNN），它的优势在于当前的面部图像数据库包含的例子太少，传统 CNN 无法学习细微的独特人脸特征。

在应对表情变化（从中性到非中性表情）人脸验证方面[8-12]，CNN 同样有效。直观来看，深度卷积网络对人脸验证的成功表明，深度学习模型可以透过中性表情跨身份捕获人的面部特征。这项工作的概念很简单：受深度身份网络（DeepID-Net）的启发，本文采用类似的设计原则，相较于原来的 CNN 使用整个面部图像来学习，本模型通过大量从检测图像生成的全局和局部图像块来训练出一个判断更精确的模型。这种设置的合理性可以用这样一个事实来解释：当增加训练样本的种类（此例中即为比整个图像更多的全局和局部图像块）时，深度学习模型的泛化能力可以显著提高。换句话说，在我们提出的框架中使用的面部斑块可以提供局部和全局的身份证据，以减轻人的变化对面部表情的负面影响。

本工作的新颖性和贡献总结如下：

(1)我们的工作采用了最先进的方法--CNN，来解决面部表情识别的问题。

(2)与传统的深度 CNN 方法不同，我们提出了 I2CNN 的概念，它通过对局部人脸成分的检测，就能更好地捕获和利用不同人的表情变化。

(3)我们的方法在两个流行的面部表情识别基准上取得了优越的性能:即 JAFFE 和 CK+数据集。

2 相关研究

2.1 人脸表情识别

人脸表情识别通常描述为分类问题，即将人脸图像或帧划分为独立的表情类别[2,3,13-22]。除了在视频中进行表情识别[13-16]之外，还有很多研究集中于从静止图像[2,3,17,19,21,22]中识别面部表情。现有的方法可以分为浅层方法[17,18,21,22]和基于深度架构的方法[2,3,20,23]。

深度学习方法兴起前[1]，人脸表情识别的流程通常如下：图像特征提取（包括基于几何和基于外观的方法）、空间池本地二进制模式（LBP）[22]，接着是利用神经网络 [24-26]、支持向量机[22,27]或贝叶斯网络[28]进行分类学习。[22]该方法通过使用增强 LBP 特征来提高面部表情识别的性能，从而获得更好的性能。近年来，深度信念网络（DBNs）被用于这一问题[2,3,23]，识别性能得以取得显著的改善。

如[18]所示，纹理和形状特征被用于面部表情识别。参考文献[3]通过人脸解析检测器检测面部成分，是通过 dbn 训练的。参考文献[2]提出了将 DBNs 的深度鉴别特征和多层感知机（MLP）分类方法相结合，该方法优于目前最先进的方法。参考文献[19]使用了两种深度网络：一种提取即时表情特征，另一种基于面部特征点获得即时几何特征，然后使用集成方法将这些特征结合起来，以提高面部表情识别的性能。为了解决非额叶面部表情识别问题，[20]首先提取人脸图像的尺度不变特征变换（SIFT）特征，然后将提取的 SIFT 特征输入深度神经网络，学习最优的判别特征向量集。虽然在野外的面部表情仍然是一个难题，但[29]在深度神经网络中取得了更好的表现。除了静态图像外，[14,30,31]还提出了很多方法来理解图像序列，其中许多都用了动态贝叶斯网络。

[2,3,19,20,23]框架和我们提出的 I2CNN 方法的共同点在于基于局部图像斑块的处理，它可以捕获细粒度的细节和人身份之间的差异。然而现有的深度学习模型（DBNs）[2,3,19,20,23]仍然依赖于人工设计的“工程”特性，于是我们引入了一个基于最近卷积神经网络架构的端到端深度学习框架。

2.2 卷积神经网络

卷积神经网络在过去的几年中引起了广泛的关注，其视觉应用如面部识别[8-12]，人脸检测[32-35]，行人检测[36-41]，对象分类[1,4]，场景理解[5-7]和其他任务[42-50]的有效性也得到证实。具体来说，深度模型被设计为端到端学习方式，利用其强大的层次架构和数千或数百万个网络参数来学习特征以及分类函数。而如今，神经网络层数变得越来越深，例如，VGGNets [11]有 16 或 19 层，GoogleNets [51]有 22 层，而 ResNets [52]有 50,101,152 层。这些方法通常可以在许多模式识别任务上取得更好的性能。

受 DeepID-Net 在人脸分类方面的成功的启发，我们提出了一种所谓的 I2CNN 方法来解决面部表情识别问题，这将在下一节中详细阐述。

3 研究方法

面部表情识别的最大挑战在于由不同人表情的独特之处的和光照和视角等外部条件的变化引起的巨大视觉特征变化。

图 1 不同人表露同一表情时的差异

如图 1 所示，“恐惧”这一表情表现出了很大的差异。例如，有些人的嘴周围有明显的皱纹（上面一排），而另一些人的眼睛有明显的扩张（下面一排）。

我们提出的 I2CNN 方法与 CNN 方法之间的区别在于深度模型的图像输入。具体来说，我们提出的 I2CNN 模型不仅采用了整个人脸图像，还采用了局部斑块。与传统方法相比，我们的方案的工作流程如图 2 所示。

我们提出的框架包括以下步骤：

（1）给定训练后的人脸图像及其表情标签，我们首先从图像背景中定位人脸前景区域。然后，我们生成了一些同时包含局部和全局个人身份信息的人脸斑块。

（2）将生成的图像斑块输入 CNN 进行深度模型训练。为了方便计算，所有斑块都共享同一网络。

（3）将第二层到最后一层的输出连接为特征，并用一个强分类器（支持向量机 SVM）利用这些特征（SVM）来预测表情类。

在测试中，我们首先用人脸检测器检测一个新图像/实例，然后将检测到的人脸区域输入训练后的 I2CNN 模型，以产生分类结果。

3.1 面部定位和面部斑块的生成

为了生成由不同表情引起的独特面部变化的斑块，我们需要一种方法来检测面部前景区域和特定标记区域——也就是说，我们采用了基于局部的面部表征的方法。

对于人脸和面部部位的检测问题，我们使用了[42]中基于 CNN 的方法。该方法用于定位面部的前景和特征点（即两只眼睛的中心、鼻尖和嘴角）。具体来说，为了利用一个粗糙-精细的框架，我们在多个尺度上进行检测，并采用了不同的网络结构。

生成一些斑块的动机是为了捕捉局部孤立的和全局相关的面部成分的细微变化。对于图像斑块的生成，我们也采用了[8]中的方法。详细地说，给定一个裁剪过的面部前景和检测到的面部特征点，我们生成了 10 个具有三个尺度的图像斑块（见图 3）。为了一般化，我们水平翻转所有的图像块，从而每个人脸图像都能形成 60 个图像斑块。

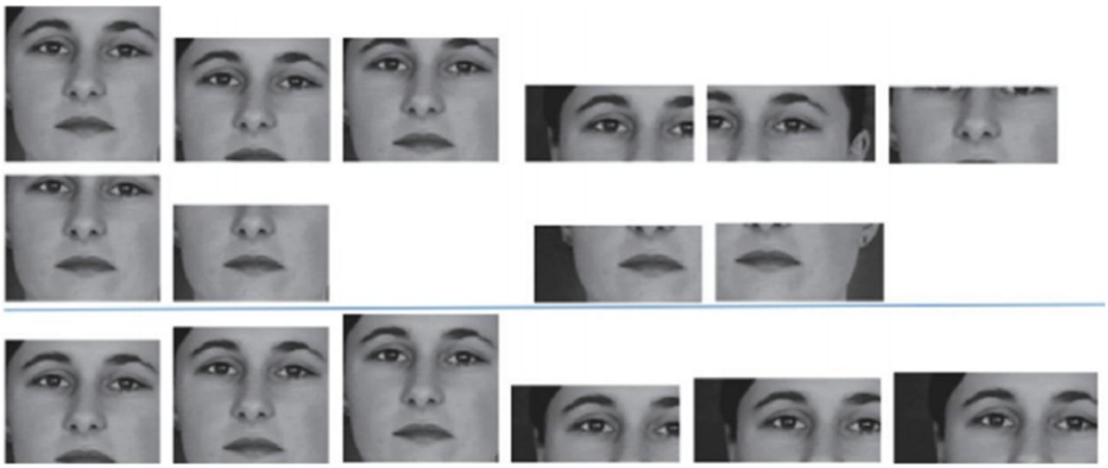

图 3 第一阶段生成的图像斑块（运用[8]中斑块检测的 CNN 方法）

3.2 卷积网络结构

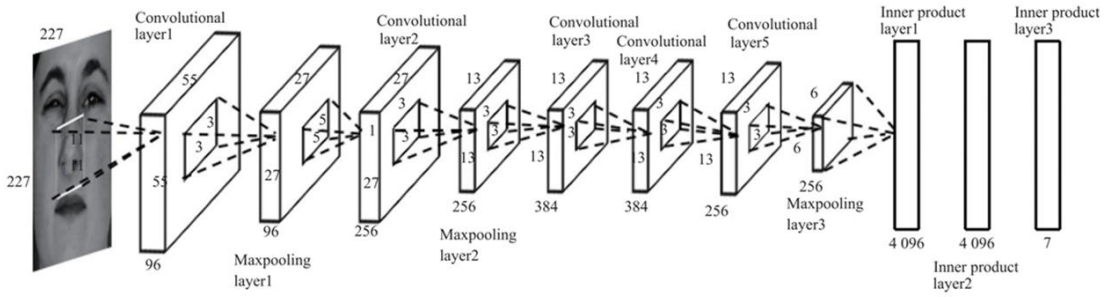

图 4 I2CNN 方法的结构

公平起见，我们采用是开源的 AlexNet 网络[1]的原始结构。如图 4 所示，该网络由 5 个卷积层、3 个全连接层和池化层组成。每个立方体的长度、宽度和高度表示特征图的数量和特征图的尺寸。立方体中的正方形表示内核大小。在最后三个全连接层下标记的数字是每层的神经元数量。最后一层是将分类错误反向传播到网络中的分类层。网络的输入为 227×227，最后一层的维数根据其预测的类数而变化。

CNN 的重要操作是卷积、池化和 ReLU（y = max（0，x））等，输出特征图的形状由(1)式计算

$$h\_o = (h\_i - kernel\_size\_h + 2 \times pad\_h)/\text{stride\_}h + 1$$
$$w\_o = (w\_i - kernel\_size\_w + 2 \times pad\_w)/\text{stride\_}w + 1. \quad (1)$$

式(1)中，h_i，w_i 分别指定输入特征图的高度和宽度，同样地，h_o，w_o 表示输出特征图的高度和宽度，kernel_size_h，kernel_size_w 表示权重过滤器的形状，pad_h，pad_w 指定添加到输入特征图每侧的像素数量。stride_h，stride_w 表示用于输入特征图的滤波器区域间隔。式(2)为 CNN 的卷积运算表达式。

$$y_j = \max(0, \sum_i (w_{i,j} * x_i) + b_j) \quad (2)$$

式(2)中，* 表示卷积，xi 和 yj 分别指定第 i 个输入特征图和第 j 个输出特征图。bj 是第 j 个输出特征图的偏置值。wi，j 表示第 i 个输入特征图与第 j 个输出特征图之间的权

重滤波器。权值滤波器与相同的值局部共享，以学习同一输出特征图中神经元在不同区域的不同特征。

最后一层使用 softmax 函数来表示不同类的概率分布。该函数为式(3)。

$$y_j = \frac{\exp(y'_j)}{\sum_i (\exp(y_i))} \qquad (3)$$

式(3)中 $y'_j = \sum_i^{4\,096} (w_{i,j} * x_i) + b_j$，也就是说，输入是看做一个特征的 4096 维向量的线性组合，输出为最后一层的每个神经元 yj。该样本将被预测为 max（yj）所代表的类。

各层的配置和网络结构对分类性能有很大的影响。在网络中使用 ReLU 层，我们就能利用倒数第二层的输出作为从每个图像斑块提取出的 4096 维深度特征表示。对于 60 个图像斑块（第 3.1 节），面部表情分类的特征向量的最终大小为 60×4 096。特征向量被输入到 SVM 分类器，后者输出最终的分类。我们利用训练数据和验证数据的特征向量来训练一个 SVM 模型，然后每次使用模型对测试数据进行分类。我们使用验证 10 次后取平均值作为最终的精度。

3.3 实现细节

训练过程中，在检测到 60 个图像斑块后，它们被重新调整到 256×256 的大小。CNN 的输入就是从中随机裁剪出的 227×227 斑块。测试过程中也执行相同流程，只不过裁剪的 227×227 斑块取自每个检测到的斑块中心。我们利用训练数据和验证数据，根据上述标准训练一个深度模型。接着，我们使用 Caffe [53]的 Matlab 接口和训练后的深度模型，分别提取训练数据、验证数据、测试数据的深度特征。

对于 SVM 分类器，我们采用 10 倍交叉验证来调整训练参数。在深度模型学习过程中，学习率一开始被设为 0.01，接着每 50000epoc 都会降低十分之一，直到准确度不再增加为止。我们运用权重衰减策略和 dropout 方法（使用前两个全连接层）来避免过拟合。前者最初被设为 5e-4,而 dropout 层的比率最初被设为 0.5

4 实验

4.1 数据集与设置

对于面部表情识别的评估，我们使用了流行的基准数据集： JAFFE [54]和 CK+ [55]。对于 JAFFE 数据集，我们选择了属于 10 个实验对象的的所有图像，一共 213 张。每个序列都有七种基本的情绪（即中性、愤怒、厌恶、恐惧、快乐、悲伤、惊讶）。JAFFE 数据集的说明见图 5。

图 5 JAFFE 数据集中的表情

图 6 CK+数据集中的表情

CK+数据集含有 123 个样貌各异的人在室内的 593 个人脸图像。CK+数据集的说明见图 6。对于 CK+数据集，我们选择了 309 个序列进行实验，选择标准是该序列可以被标记为不含中性的六种基本表情之一（愤怒、厌恶、恐惧、快乐、悲伤、惊讶）。按照[22,23]中的设置，对于每个序列，我们选择第一帧和最后三帧峰值帧进行实验。

4.2 与先进方法的对比

我们在 JAFFE 和 CK+数据集上进行了两个实验，并与目前最先进的面部表情识别方法进行了比较，其结果如表 1 和表 2 所示。一方面，我们在 JAFFE 数据集上的第一个实验中获得了比最先进的技术更好的分类性能。另一方面，我们提出的 I2CNN 方法在 CK+数据集上的 6 类和 7 类表情识别都具有优越的性能，结果如表 2 所示。对于这两种情况，如 Fig.7 所见，我们对每个情绪类的识别准确度都非常高。在图 8 中，我们在混淆矩阵中描述了来自[22]的结果。参考文献[22]使用了图像的 LBP 特征和具有 RBF 内核的 SVM 类化器。相比 Fig.8，除了恐惧和悲伤外，几乎所有的表情我们都有更高的准确度。识别性能的提升情况在混淆矩阵中用不同颜色表达。

表 1 JAFFE 数据集性能对比

Method	Accuracy/%
I^2CNN (our)	75.280 6
SVM (RBF)+Boosted-LBP ([22])	81.0
SVM (linear)+Boosted-LBP ([22])	79.8
SVM (polynomial)+Boosted-LBP ([22])	79.8

表 2 CK+数据集上 6/7 表情类性能对比

Method	6 emo	7 emo
I^2CNN (our)	98.3	96.2
BDBN ([23])	96.7	-
SVM(RBF) + Boosted-LBP ([22])	95.1	91.4
FP + SAE ([3])	-	91.1
SVM(RBF) + LBP ([22])	92.6	88.9
SVM(RBF) + Gabor ([22])	89.8	86.8
CSPL+SVM ([56])	-	89.9
LDP+template matching ([57])	-	86.9
ITBN ([58])	-	86.3
Geometric representation + Gaussian three-augmented navie Bayes classifiers ([13])	73.2	-

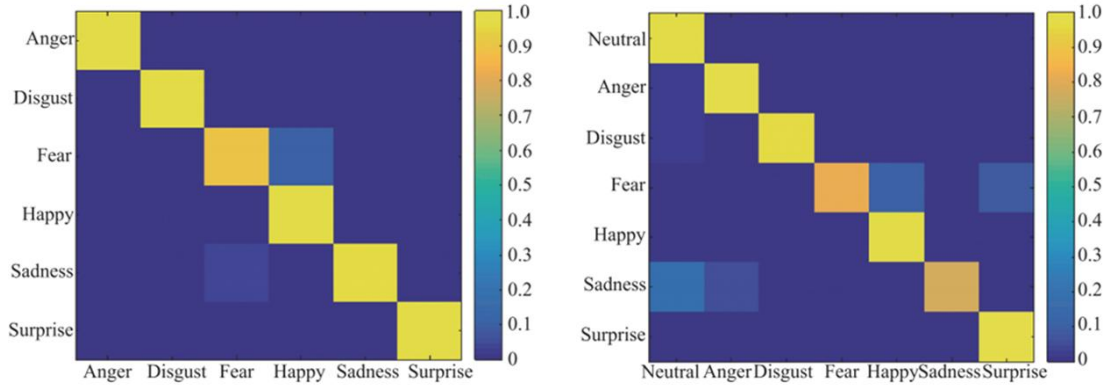

图 7 CK+数据集 6 类（左）和 7 类（右）表情分类的 I2CNN 混淆矩阵

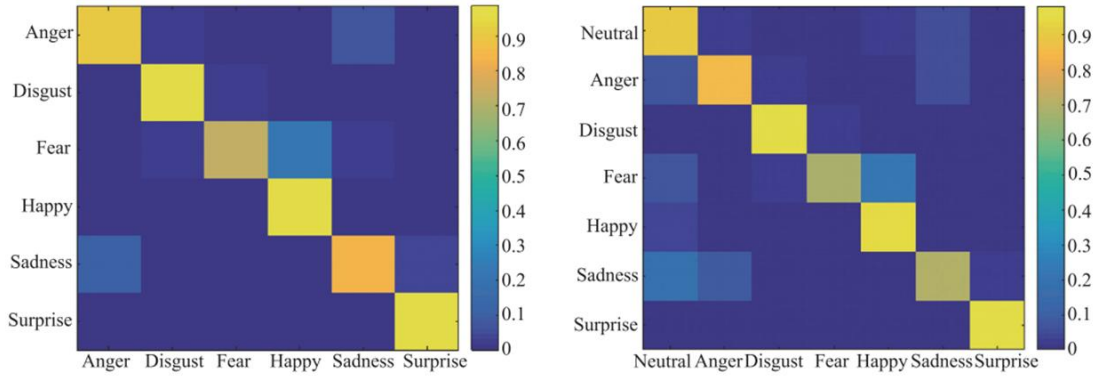

Fig. 8 SVM(RBF) + LBP ([22]) confusion matrices for CK+ six (left) and seven (right) emotion classifications

图 8 CK+数据集 6 类（左）和 7 类（右）表情分类的 SVM(RBF)+LBP([22])混淆矩阵

4.3 身份启发与传统的卷积网络

如表 3 所示，我们分别评估了所提出的 I2CNN 和卷积 CNN [1]。我们可以观察到，所提出的 I2CNN 在两个数据集上都可以获得更好的分类性能，两者的明显边际只能用身份启发部分来解释。在 CK+数据集上的实验结果表明，与传统的 CNN 方法相比，6 类和 7 类情绪识别的分类准确率分别提高到 98.3%和 96.2%。在 JAFFE 上，7 类情绪的准确率可提高 19.76%。在第 4.4 节中，我们评估了图像斑块的数量对识别性能的影响

表 3 I2CNN 与传统 CNN 对比

Method	Accuracy/%		
	JAFFE	CK+ (6)	CK+ (7)
I^2CNN	75.280 6	98.3	96.2
CNN	62.860 2	92.2	89.7

4.4 对图像斑块数量的评估

我们评估了图像斑块对 I2CNN 性能的影响，并尝试了 k = 1,5,60 斑块数的不同组合，结果见表 4。k = 1 意味着我们只使用一个全局图像，k = 5 意味着我们只使用由检测到的五个面部特征点得到的五个局部斑块，而 k = 60 表示我们每个图像都使用所有斑块。相比于单个全局斑块，7 类情绪在 JAFFE 数据集上的性能从 62.8602%提高到 75.2806%。显然，我们可以通过增加每个图像的斑块数来获得更好的面部表情识别性能。同样，与 CK+数据集上的 6 种类和 7 类情绪相比，使用所有图像斑块可以分别提高 6.62%和 7.25%的性能。如表 4 显示，我们可以通过使用更多的面部斑块来提高性能。

表 4 I2CNN 不同图像斑块数的性能对比

Method	Accuracy/%		
	JAFFE	CK+ (6)	CK+ (7)
I^2CNN-60	75.280 6	98.3	96.2
I^2CNN-5	70.982 5	94.7	93.5
I^2CNN-1	62.860 2	92.2	89.7

5 总结

面部表情识别在现实世界中有很多应用，包括情绪分析、人机交互（HCI）以及解决自动驾驶难题，即可以监控司机的情绪/情绪，以避免潜在的事故。面部表情识别的主要问题是如何提取关键的面部特征，以更好地区分不同的情绪。人们可以利用传统的、工程化的特征提取方法，例如 LBP [59]、Gabor [60]和 SIFT [61]。不过，近年来 CNN 已经被证实能够自动提取从低到高层次的特征，并在许多模式识别任务的识别精度上优于行业标杆。

面对这一工作，我们提出了一种新型 I2CNN 方法来减轻不同人表情差异对人脸识别的负面影响。与现有的传统 CNN 方法相比，我们的方案是基于多尺度的全局图像和局部面部斑块，它能在 JAFFE 和 CK+数据集上实现显著的性能提升。

然而，仍有许多公开的挑战。首先，光照条件、人脸的姿势、种族的差异等对面部表情识别有显著影响。其次，调整 CNN 的最佳参数是非常费时的。为了解决这一问题，在未来的工作中，我们将设计一个简单但更有效的 CNN 网络来进行面部表情识别。

致谢

感谢芬兰 CSC-IT 科学中心提供的慷慨的计算资源，以及 NVIDIA 为支持我们的学术研究所捐赠的特斯拉 K40 GPU。

参考文献

[1]A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classifi-cation with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012: 1097–1105.

[2]X. Zhao, X. Shi, S. Zhang. Facial expression recognition via deep learning. IETE Technical Review, 2015: 1–9.

[3]Y. Lv, Z. Feng, C. Xu. Facial expression recognition via deep learning. Proc. of the International Conference on Smart Com puting, 2014: 303–308.

[4]K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. Proc. of the Eu ropean Conference on Computer Vision, 2014: 346–361.

[5]R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierar chies for accurate object detection and semantic segmentation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580–587.

[6]S. Gupta, P. Arbel´aez, R. Girshick, et al. Indoor scene under standing with RGB-D images: bottom-up segmentation, object detection and semantic segmentation. International Journal of Computer Vision, 2015, 112(2): 133–149.

[7]H. Jung, M. K. Choi, K. Soon, et al. End-to-end pedestrian col lision warning system based on a convolutional neural network with semantic segmentation, arXiv: 1612.06, 2016.

[8]Y. Sun, X. Wang, X. Tang. Deep learning face representation from predicting 10, 000 classes. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1891–1898.

[9]Y. Sun, X. Wang. Hybrid deep learning for face verification. Proc. of the IEEE International Conference on Computer Vi sion, 2013: 1489–1496.

[10]F. Schroff, D. Kalenichenko, J. Philbin. Facenet: a unified em bedding for face recognition and clustering. Computer Vision and Pattern Recognition, 2015: 815–823. [11]O. M. Parkhi, A. Vedaldi, A. Zisserman. Deep face recogni tion. Proc. of the British Machine Vision Conference, 2015: 1–12.

[12]W. Ouyang, X. Wang, X. Zeng, et al. Deepid-net: deformable deep convolutional neural networks for object detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2403–2412.

[13]I. Cohen, N. Sebe, A. Garg, et al. Facial expression recog nition from video sequences: temporal and static modeling. Computer Vision and Image Understanding, 2003: 160–187.

[14]R. E. Kaliouby, P. Robinson. Real-time inference of com plex mental states from facial expressions and head gestures. Real-time Vision for Human-Computer Interaction, DOI: 10.10710-387-27890-711.

[15]I. Kotsia, I. Pitas. Facial expression recognition in image se quences using geometric deformation features and support vector machines. IEEE Trans. on Image Processing, 2007: 172–187.

[16]M. Pantic, I. Patras. Dynamics of facial expression: recogni tion of facial actions and their temporal segments from face profile image sequences. IEEE Trans. on Systems, Man, and Cybernetics-Part B, 2006, 36(2): 433–499.

[17]P. K. Manglik, U. Misra, H. B. Maringanti, et al. Facial expres sion recognition. Proc. of the IEEE International Conference on Systems, Man and Cybernetics, 2004: 2220–2224.

[18]W. Zheng, C. Liu. Facial expression recognition based on tex ture and shape. Proc. of the Wireless and Optical Communica tion Conference, 2016: 1–5.

[19]H. Jung, S. Lee, J. Yim, et al. Joint fine-tuning in deep neural networks for facial expression recognition. Proc. of the IEEE International Conference on Computer Vision, 2015: 2983–2991.

[20]T. Zhang, W. Zheng, Z. Cui, et al. A deep neural network driven feature learning method for multi-view facial expres sion recognition. IEEE Trans. on Multimedia, 2016, 18(12): 2528–2536.

[21]D. Ghimire, J. Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors, 2016: 7714–7734.

[22]C. Shan, S. Gong, P. W. McOwan. Facial expression recog nition based on local binary patterns: a comprehensive study. Image and Vision Computing, 2009: 803–816. [23]P. Liu, S. Han, Z. Meng, et al. Facial

expression recognition via a boosted deep belief network. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1805–1812.

[24]C. Padgett, G. W. Cottrell. Representing face images for emo tion classification. Advances in Neural Information Processing Systems, 1997: 894–900. [25]Y. L. Tian. Evaluation of face resolution for expression anal ysis. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2004: 82–82.

[26]Z. Zhang, M. Lyons, M. Schuster, et al. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. Proc. of the IEEE In ternational Conference on Automatic Face and Gesture Recog nition, 1998: 454–459.

[27]M. S. Bartlett, G. Littlewort, M. Frank, et al. Recognizing facial expression: machine learning and application to spon taneous behavior. Proc. of the IEEE Computer Society Con ference on Computer Vision and Pattern Recognition, 2005: 568–573.

[28]I. Cohen, N. Sebe, A. Garg, et al. Facial expression recognition from video sequences. Proc. of the IEEE International Confe rence on Multimedia and Expo, 2002: 121–124.

[29]A. Mollahosseini, B. Hassani, M. J. Salvador, et al. Facial expression recognition from world wild web. Proc. of the IEEE Conference on Computer Vision and Pattern Recogni tion Workshops, 2016: 1509–1516.

[30]J. Hoey, J. J. Little. Value directed learning of gestures and fa cial displays. Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004: 1026–1033.

[31]Y. Zhang, Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(5): 699–714.

[32]M. Szarvas, A. Yoshizawa, M. Yamamoto, et al. Multi-view face detection using deep convolutional neural networks. Proc. of the ACM International Conference on Multimedia Retrieval, 2015: 224–229.

[33]S. Yang, P. Luo, C. L. Chen, et al. Wider face: a face detection benchmark. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5525–5533. [34]H. Li, Z. Lin, X. Shen, et al. A convolutional neural network cascade for face detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015: 5325–5334.

[35]Y. Zheng, C. Zhu, K. Lu, et al. Towards a deep learning frame work for unconstrained face detection. Proc. of the 8 th IEEEInternational Conference on Biometrics: Theory, Applications and Systems, 2016: 1–8.

[36]X. Wang. A discriminative deep model for pedestrian detec tion with occlusion handling. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3258–3265.

[37]W. Ouyang, X. Zeng, X. Wang. Modeling mutual visibility re lationship in pedestrian detection. Computer Vision and Pat tern Recognition, 2013: 3222–3229. [38]X. Zeng, W. Ouyang, X. Wang. Multi-stage contextual deep learning for pedestrian detection. Proc. of the IEEE Confer ence on Gomputer Vision, 2013: 121–128.

[39]P. Luo, X. Wang, X. Tang. Pedestrian parsing via deep decom positional network. Proc. of the IEEE International Confer ence on Computer Vision, 2013: 2648–2655. [40]P. Luo, Y. Tian, X. Wang, et al. Switchable deep network for pedestrian detection. Proc. of the IEEE Conference on Com puter Vision and Pattern Recognition, 2014: 899–906.

[41]X. Zeng, W. Ouyang, M. Wang, et al. Deep learning of scene-specific classifier for pedestrian detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recogni tion, 2014: 472–487.

[42]Y. Sun, X. Wang, X. Tang. Deep convolutional network cas cade for facial point detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3476–3483.

[43]P. Sermanet, D. Eigen, X. Zhang, et al. Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv: 1312.6229, 2014. [44]K. Simonyan, A. Zisserman. Very deep convolutional net works for large-scale image recognition. arXiv: 1409.1556, 2014.

[45]K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition.

IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916.
[46]W. Ouyang, X. Chu, X. Wang. Multi-source deep learning for human pose estimation. Proc. of the IEEE Conference on Com puter Vision and Pattern Recognition, 2014: 2337–2344.
[47]J. Zhang, S. Shan, M. Kan, et al. Coarse-to-fine auto-encoder networks for real-time face alignment. Proc. of the European Conference on Computer Vision, 2014: 1–16.
[48]K. Simonyan, A. Vedaldi, A. Zisserman. Deep inside convo lutional networks: visualising image classification models and saliency maps. arXiv: 1312.6034, 2014. [49]R. C. Malli, M. Aygun, H. K. Ekenel. Apparent age estimation using ensemble of deep learning models. arXiv: 1606.02909, 2016.
[50]X. Tang, X. Wang, P. Luo. Hierarchical face parsing via deep learning. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2012: 2480–2487. [51]C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolu tions. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1–9.
[52]K. He, X. Zhang, S. Ren, et al. Deep residual learning for im age recognition. arXiv: 1512.03385, 2015.
[53]Jia, Yangqing, Shelhamer, et al. Caffe: convolutional archi tecture for fast feature embedding. arXiv, Computer Science, 2014: 675–678.
[54]M. Lyons, S. Akamatsu, M. Kamachi, et al. Coding facial expressions with gabor wavelets. Proc. of the IEEE Interna tional Conference on Automatic Face and Gesture Recogni tion, 1998: 200–205.
[55]P. Lucey, J. F. Cohn, T. Kanade, et al. The extended cohn kanade dataset(ck+): a complete dataset for action unit and emotion-specified expression. Proc. of the IEEE Computer So ciety Conference on Computer Vision and Pattern Recognition Workshops, 2010: 94–101.
[56]L. Zhong, Q. Liu, P. Yang, et al. Learning active facial patches for expression analysis. Proc. of the IEEE Conference on Com puter Vision and Pattern Recognition, 2012: 2562–2569.
[57]T. Jabid, M. H. Kabir, O. Chae. Robust facial expression recognition based on local directional pattern. Proc. of the IEEE Conference on Computer Vision and Pattern Recogni tion, 2014, 51: 784–794.
[58]Z. Wang, S. Wang, Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recogni tion. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3422–3429.
[59]T. Ojala, M. Pietikainen, T. Maenpaa. Multiresolution gray scale and rotation invariant texture classification with local bi nary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002: 971–987.
[60]L. Wiskott, J. M. Fellous, N. Kuiger, et al. Face recognition by elastic bunch graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997, 19: 775–779.
[61]D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60: 91–110.