《媒体与认知》课程实验

人脸表情识别

实验报告

无 42 张　佳 2014011042

无 42 陈誉博 2014011058

无 42 张祎玮 2014011063

1 背景简介

一直以来，人类以拥有情感而自豪，这是人类和机器的一种本质上的区别。随着计算机的发展，人机之间的沟通交流日益受到重视，尤其是带有感情的沟通交流。所以，能够观察和辨别情感是自然、亲切、生动的交互的开始。

表情识别工作的重点在于其中的特征提取阶段。按照提取特征的不同，分为手工设计特征和深度网络特征，两种方法都已进入较为成熟的阶段，部分算法模型已被公司作为商用模型。

2 数据集介绍

2.1 CK+

该数据库是在 Cohn-Kanade Dataset [10] 的基础上扩展来的，发布于 2010 年。序列数增加了 22%，受试者数量增加了 27%。每个序列的目标表达式完全是 FACS 编码的，包含表情标签和 Action Units 标签。表情标签已被修改和验证。

数据库包括 123 个 subjects, 593 个图片序列，每个图片序列的最后一张图片都有 action units 的 label，而在这 593 个图片序列中，有 327 个序列有表情标签。这个数据库是人脸表情识别中比较流行的一个数据库。

2.2 JAFFE

JAFFE 数据库 [11] 专门用于表情识别研究的基本表情，该数据库中包含了 213 幅 256×256 的日本女性的脸相，每幅图像都有原始的表情定义。表情库中共有 10 个人，每个人有 7 种表情（中性、高兴、悲伤、惊奇、愤怒、厌恶、恐惧）。均为正面脸相，且把原始图像进行重新调整和修剪，使得眼睛在数据库图像中的位置大致相同，脸部尺寸基本一致，光照均为正面光源，但光照强度有差异。表情标定很标准，大多数研究表情识别的文章中都使用它来训练与测试。

2.3 TFEID

TFEID 数据库 [3] 由 40 个模型（20 个男性）捕获的 7200 个刺激组成，每个具有八个面部表情：中性，愤怒，蔑视，厌恶，恐惧，幸福，悲伤和惊喜。要求模型注视两个不同的角度（0° 和 45° ）。每个表达式包括两种强度（强和弱），并且由两个 CCD 摄像机同时用不同的视角（0° 和 45°）捕获。

2.4 KDEF

KDEF 数据库 [4] 包含一套完整的 4900 张人脸表情的图片，于 1998 年开发。该材料最初被开发用于心理和医学研究目的。特别适用于感知、注意、情感、记忆和反向掩蔽实验。以多角度创造柔软，均匀的光线拍摄表现形式，使用均匀的 T 恤颜色，以及使用格网在拍摄过程中使参与者的脸部居中，并将眼睛和嘴部定位在固定的扫描时的图像坐标。

该集合包含 70 个人，每个人显示 7 种不同的情感表达，每个表情从 5 个不同的角度拍摄（两次）。

2.5 FER

FER 数据库 [7] 一共有 35887 张人脸图像，分辨率大小 48×48，样本数量大于大多数数据集。图片的主要来源是通过谷歌图片的关键字搜索，样本差异性大，但却是自然条件下人类真实的表情，图片的分辨率较低。

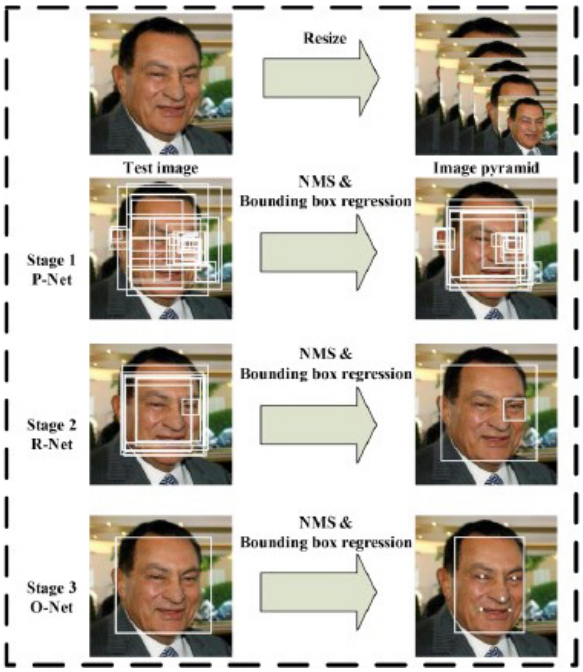

Figure 1: MTCNN 流程图

3 实验原理

3.1 数据预处理：MTCNN

考虑到实验中使用的训练数据来源范围非常广，不仅包含多个学术公开数据库的数据，而且包含自采数据，因此需要对所有数据进行预处理以去除非人脸区域以及图片色调等因素对实验结果的影响。因此，我们采用 MTCNN 方法 [13] 提取人脸区域，并将其转化为灰度图，并统一尺寸，以备使用。下面介绍 MTCNN 的基本原理。

网络模型

MTCNN（Multi-Task Convolutional Neural Network），其总体是一个深层级（24 层）多任务框架的神经网络，采用了三级深卷积网络的级联结构。

首先，我们将其大小调整到不同的尺度以构建 image pyramid，并作为三级级联框架的输入，三级深卷积网络的级联结构如下：

- P-Net（Proposal Network）：生成候选人脸框及其边界框回归向量，然后使用该边界框回归向量来校准候选框，并采用非极大值抑制（NMS）方法合并高度重叠的候选框。
- R-Net（Refine Network）：进一步筛选候选框，使用边界框回归进行校正，并用 NMS 合并候选框。
- O-Net：与 R-Net 类似，选出最优候选框，并输出五个特征点的位置。

网络的整体流程图参见图 1，具体网络结构参见图 2。

网络训练

在训练 MTCNN 的过程中，我们期望网络能够实现下面三个功能：

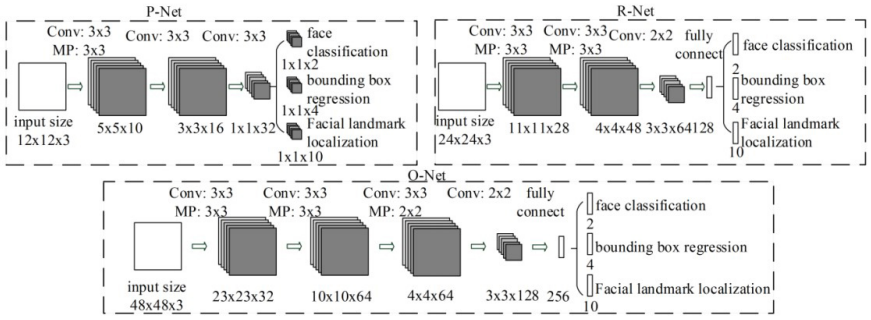

Figure 2: MTCNN 网络结构

- 人脸分类：网络学习的部分目的可以建模为对候选框区域进行是/不是人脸的二分类问题。对于每个训练样本，使用如下的交叉熵损失函数：

$$L_i^{det} = -(y_i^{det}\log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \tag{1}$$

  其中 p_i 是由网络产生的是人脸的概率，y_i^{det} 是真实的标签。

- 边界框回归：对于每个候选框区域，网络要做的是预测该候选框和最近的真实值的偏移。那么网络学习的另一部分目的可以建模为一个回归问题。对于这个回归问题，在训练中使用如下的 Euclidean 损失函数：

$$L_i^{box} = ||\hat{y}_i^{box} - y_i^{box}||_2^2 \tag{2}$$

  其中 $\hat{y}_i^{box}$ 是从网络中得到的回归目标，y_i^{box} 是真实的人脸框。

- 人脸关键点定位：与边界框回归问题类似，人脸特征点定位也可以建模为一个回归问题，那么我们同样使用 Euclidean 损失函数：

$$L_i^{landmark} = ||\hat{y}_i^{landmark} - y_i^{landmark}||_2^2 \tag{3}$$

  其中 $\hat{y}_i^{landmark}$ 是从网络中得到的回归目标，$y_i^{landmark}$ 是真实的人脸关键点坐标。

由于在训练网络的时候，要求卷积网络实现不同的任务，因此需要使用不同种类的训练数据，比如在训练二分类网络的时候会使用人脸图像和非人脸图像。这意味着在训练某些网络的时候，只会使用上面提到的三种损失函数中的一部分，而不是全部使用。基于这种情况，可以用如下的表达式来建模网络的学习目标：

$$\min \sum_{i=1}^{N} \sum_{j \in (det, box, landmark)} \alpha_j \beta_i^j L_i^j \tag{4}$$

其中 N 是训练样本的数目，α_j 是任务权重。对于不同的网络，其不同的功能对应不同的任务权重。

3.2 特征提取

LBP 特征提取

LBP（Local Binary Pattern）特征 [12] 是一种简单，有效的纹理分类的特征提取算法。从纹理分析的角度来看，图像上某个像素点的纹理特征，大多数情况下是指这个点和周围像素点的关系，即这个点和它的邻域内点的关系。从哪个角度对这种关系提取特征，就形成了不同种类的特征。有了特征，就能根据纹理进行分类。LBP 构造了一种衡量一个像素点和它周围像素点的关系。

对图像中的每个像素，通过计算以其为中心的 3×3 邻域内各像素和中心像素的大小关系，把像素的灰度值转化为一个八位二进制序列。具体计算过程如下图所示，对于图像的任意一点 I_c，其 LBP 特征计

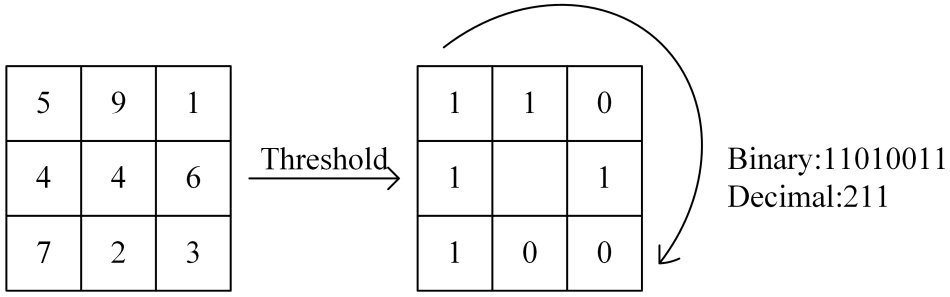

Figure 3: LBP 算子

算为，以 I_c 为中心，取与 I_c 相邻的 8 各点，按照顺时针的方向记为 $I_0, I_1, \ldots, I_7$；以 I_c 点的像素值为阈值，如果 I_i 点的像素值小于 I_c，则 I_i 被二值化为 0，否则为 1；将二值化得到的 0、1 序列看成一个 8 位二进制数，将该二进制数转化为十进制就可得到 I_c 点处的 LBP 算子的值。实验结果表明，LBP 算子对于图片的灰度变化具有很好的鲁棒性。

对于人脸识别问题，一种常用的特征表征方法是 LBP 直方图（LBP Histogram）[1]。对于直方图，可以用如下的表达式来定义：

$$H_i = \sum_{x,y} I\{f_l(x,y) = i\}, i = 0, 1, \ldots, n-1 \tag{5}$$

其中 n 是由 LBP 算子得到的不同的标签数目，而

$$I(A) = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is false} \end{cases} \tag{6}$$

按照以上方式定义的直方图包含了整幅图像中信息，包括边缘、平坦区域等。为了让直方图能够更好的表征人脸信息，可以将图像分割成小的区域 $R_0, R_1, \ldots, R_{m-1}$，而直方图的定义也随之改变：

$$H_i = \sum_{x,y} I\{f_l(x,y) = i\} I\{(x,y) \in R_j\}, i = 0, 1, \ldots, n-1, j = 0, 1, \ldots, m-1 \tag{7}$$

以上直方图可以描述图片的 3 个层次的特征信息：直方图中的每个标签代表了图片的像素级局部性，对每个区域求和代表了区域级的局部性，每个区域的直方图最后的求和代表了整张图片的全局特征。

HOG 特征提取

HOG（Histogram of Oriented Gradient, HOG）[5] 即统计图像局部区域的梯度方向信息来作为该局部图像区域的表征。由其主要思想是用梯度或者边缘的方向密度分布描述局部目标的表象和形状。具体实现方法是：首先，将图像分成小的连通区域，也成为细胞单元（cell），采集各细胞单元内部像素点的梯度或边缘的方向直方图。之后在一个更大的区块（block）内计算各细胞单元的直方图在区块内的密度，并根据这个密度对区块内部的所有细胞单元进行归一化。这样做的好处是可以让 HOG 特征对光照和阴影变化具有更好的鲁棒性。最后将所有区块的直方图串联，成为整幅图像的 HOG 特征。HOG 特征提取的原理图参见图 4。

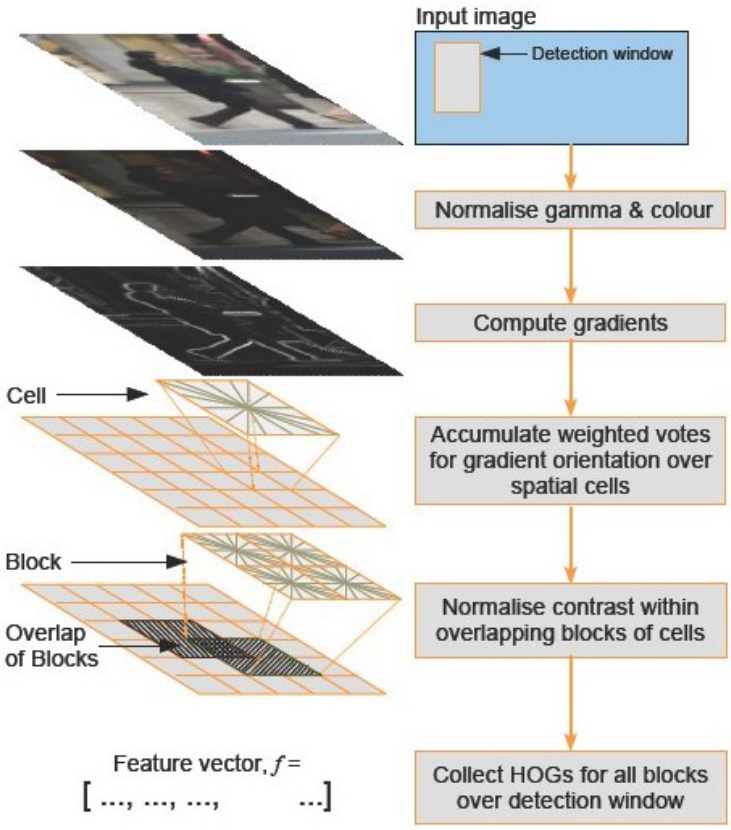

Figure 4: HOG 特征提取

3.3 特征分类

K 近邻分类

K 近邻（K Nearest Neighbour, KNN）[6] 分类是模式识别中的一种经典的分类方法，其思想用一句话来概括就是“物以类聚，人以群分”。KNN 分类器是一种基于距离的分类器，其分类方法是通过计算待分类数据与训练数据特征值之间的距离，然后选取 K 个距离最近的邻居样本进行分类投票，投票胜者作为最终的分类结果。如果 K = 1，那么直接用与待分类样本距离最近的训练样本的类别作为待分类样本的分类结果。在人脸识别问题中，一种常用的方法是用 LBPH 作为人脸图片的特征，并使用 1-近邻方法对人脸进行分类。显而易见的是，KNN 方法不需要训练的过程，因此这个分类方法的关键点就落在了距离函数的选择上。

距离函数的选择有很多种，比如：

- 直方图交叉函数（Histogram intersection）：

$$D(\mathbf{S},\mathbf{M}) = \sum_i \min(S_i, M_i) \tag{8}$$

- 对数似然函数（Log-likehood）：

$$L(\mathbf{S},\mathbf{M}) = -\sum_i S_i \log M_i \tag{9}$$

- χ 方函数（Chi square）：

$$\chi^2(\mathbf{S},\mathbf{M}) = \sum_i \frac{(S_i - M_i)^2}{S_i + M_i} \tag{10}$$

在 LBPH 的特征表征中，图片被切分为很多小的区域。考虑到不同的区域对图片分类结果的影响不完全相同，因此可以对以上的几种距离函数引入权重以代表其对分类结果的不同影响。例如，带权重的 χ 方函数可以写成如下表达式：

$$\chi_{\omega}^{2}(\mathbf{S},\mathbf{M})=\sum_{i,j}\omega_{j}\frac{(S_{i,j}-M_{i,j})^{2}}{S_{i,j}+M_{i,j}} \tag{11}$$

其中 j 为区域标号，ω_j 为对应权重。

SVM 分类

支持向量机（Support Vector Machine，SVM）[2] 是一种常用的小样本集统计学习方法。在传统的线性判别分析中，优化准则是基于经验风险最小化，也就是利用训练样本上的估计风险嗲气样本期望风险。在样本不足的情况下很容易造成分类器的过拟合。在支持向量理论中，为了保证分类器具有较好的推广性能，用结构化风险最小化取代经验风险最小化作为优化准则。

支持向量机的主要设计目的是使得不同类别的样本分类间隔最大。根据分类决策过程，为了实现分类，需要求解分类决策函数 $f(\mathbf{x})=\mathbf{w}^{T}\mathbf{x}+b$，以二分类问题为例，如果 $f(\mathbf{x})>0$ 则被认为属于类别 1，否则属于类别 2。因此，$H:\mathbf{w}^{T}\mathbf{x}+b=0$ 可以视为分类界面。进一步，H_1,H_2 分别为平行于 H 且与 H 距离为 $\frac{1}{||\mathbf{w}||}$ 的两个平面，它们之间的距离称为分类间隔。支持向量机学习的目的是希望找到最优投影向量 $\mathbf{w}$ 和最优分类界面 H，使得分类间隔最大。但是，在实际使用中，并不是所有的问题都是线性可分的。这时就需要引入非线性分类问题。对于非线性分类问题，一种常用的方法是核函数（kernel function）法，通过核函数对样本内积进行非线性操作，以达到提高特征维度、将非线性可分问题转化为线性可分问题的目的。

对于二分类问题，SVM 可以直接使用，并且可以取得很好的效果。但是对于多分类问题，直接使用一般的 SVM 无法解决，这时就需要引入一些组合方法，使得 SVM 能够对多类样本进行分类。常用的组合方法有三种：

- 一对多法（one-versus-rest, OVR）：训练时依次把某个类别的样本归为一类, 其他样本归为另一类，用 k 个类别的样本构造 k 个 SVM。分类时将未知样本分类为具有最大分类函数值的那类。
- 一对多法（one-versus-one, OVO）：在任意两类样本之间设计一个 SVM，用 k 个类别的样本构造 $\frac{k(k-1)}{2}$ 个 SVM。当对一个未知样本进行分类时，最后得票最多的类别即为该未知样本的类别。
- 层次分类法：首先将所有类别分成两个子类，再将子类进一步划分成两个次级子类，如此循环，直到得到一个单独的类别为止。

深度学习分类：ResNet

深度学习方法是今年来机器学习领域一种新兴的方法。由于其思想是模仿人脑的思考机制进行工作，因此在很多之前认为较为困难的任务中都有很好的表现。在计算机视觉中，尤其是图像处理领域，最常用一种的方法是卷积神经网络。

卷积神经网络（Convolutional Neural Network,CNN）是一种前馈神经网络，其本质上是一个多层感知机。CNN 的人工神经元可以响应一部分覆盖范围内的周围单元，其成功的原因也正是在于它所采用的局部连接和共享权值的方式，一方面减少了的权值的数量使得网络易于优化，另一方面降低了过拟合的风险。卷积神经网络能够取得成功的另一个关键原因是其模拟了局部感受野。通过不同形式的卷积神经网络的组合，可以针对不同的任务实现自动特征提取。在网络末端添加分类层后，可以通过优化损失函数来实现分类功能。

在用神经网络进行特征提取与分类决策时，网络的深度对于提取、分类和识别的效果有很大的影响，一般认为较深的网络相比较浅却较广的网络对于特征提取的效果更好。然而，随着网络层数的加深，一方面，梯度消失与梯度爆炸的问题阻碍了收敛，这一问题可以通过 normalized initialization 和

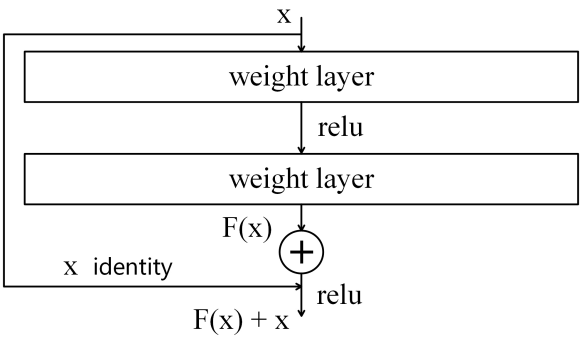

Figure 5: ResNet 中的残差单元

intermediate normalization layers 两种方法进行解决。然而另一方面，当网络开始收敛时，也会出现退化问题：随着网络深度的增加，准确率达到饱和，然后迅速下降，同时，研究表明，这种退化的出现并不是过拟合的结果，并且随着常规网络的堆叠（plain network）深度的增长导致了更高的训练误差。上述问题出现的原因是并非所有系统都是容易优化的，有时利用网络直接逼近期望得到的底层映射函数并非易事。

为了解决上述问题，Kaiming He 等人提出了一种新的网络结构：ResNet（Residual Network）[8]，在加深网络的情况下提高了训练速度，并且克服了梯度的不稳定问题和退化问题。其基本残差单元结构如图 5 所示。ResNet 中残差单元的功能可以用如下表达式描述：

$$\mathbf{y}_l = h(\mathbf{x}_l) + F(\mathbf{x}_l, W_l) \tag{12}$$

$$x_{l+1} = f(\mathbf{y}_l) \tag{13}$$

其中 h 为残差函数，F 为残差单元中权重层的映射函数，f 为激活函数。通常，残差函数选择单位映射函数 $h(\mathbf{x}_l) = \mathbf{x}_l$。ResNet 的核心思想是在网络结构中引入残差结构，使得网络输出的逼近目标变为 0，而不是某一个底层映射函数，其实现难度自然大大降低。通过引入残差函数，避免了网络层数堆叠的问题，降低了退化的影响。而且，单位映射函数不需要引入额外的参数，也没有增加复杂度，使得网络训练更加容易，而且也使得网络的学习速度更快，因为逼近一个单位函数比逼近另外一个形式的映射函数更加容易。另外，研究表明 [9]，相比于相比于其他形式的残差函数（$h(\mathbf{x}_l) = 0.5x_l, h(\mathbf{x}_l) = F_{\mathbf{CNN}}(x_l, W_l')$ 等），使用单位映射函数构建的残差单元分类误差最小。

4 实现细节

4.1 MTCNN

在此次实验中，采用 caffe 对 MTCNN 进行了实现。由 (4) 可知，对于三个阶段不同的网络结构，应该赋予不同的任务权重。在训练过程中，对 P-Net 和 R-Net 使用 $\alpha_{det} = 1, \alpha_{box} = 0.5, \alpha_{landmark} = 0.5$，对 O-Net 使用 $\alpha_{det} = 1, \alpha_{box} = 0.5, \alpha_{landmark} = 1$。另外，由于三个网络对于图片全局信息和局部信息的要求不同，其输入图片的大小也不同。这就是在将图片输入网络之前需要创建 image pyramid 的原因。在网络实现中，P-Net 的输入图片大小为 12×12 的彩色图片，R-Net 的输入图片大小为 24×24 的彩色图片，O-Net 的输入图片大小为 48×48 的彩色图片。

使用以上参数对网络进行训练，就可以实现对提取人脸并定位特征点的功能，部分输出结果参见图 6。由于此次实验选取的训练数据均为正脸，用来测试的数据集也全部为近似正脸，因此人脸转正的步骤可以直接跳过，也就是说不需要用到网络输出的 5 个特征点的坐标。使用网络输出的人脸框将图片中

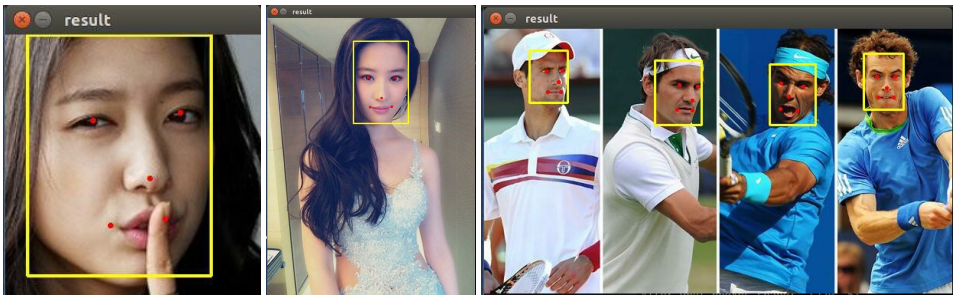

Figure 6: MTCNN 人脸提取效果

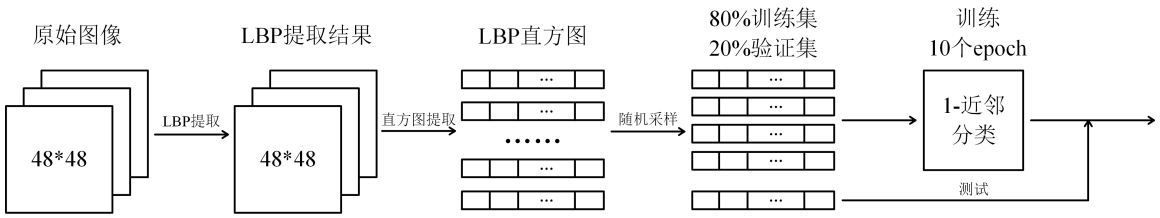

Figure 7: LBPH 分类流程图

的非人脸区域去掉，并将人脸区域图片保存为固定大小（预训练数据大小为 48×48，微调数据大小为 256×256）的灰度图像以备训练使用。

4.2 LBPH 识别

LBPH 分类方法是使用 LBPH 作为图片的特征算子，并使用 1-近邻方法对人脸进行表情分类。由于在对此方法进行测试时，真正的测试数据集尚未公开，因此需要将全部训练集分为训练集和验证集。此处，从训练集中随机采样，得到 80% 的数据作为真正的训练集，剩余 20% 的数据作为验证集，验证模型的分类准确率和泛化能力。使用上述训练集训练 10 个 epoch。LBPH 分类方法的流程图参见图 7。

4.3 SVM 识别

SVM 分类方法是使用 HOG 特征作为图片的特征算子，并使用 OVR SVM 对人脸进行表情分类。在进行 HOG 特征提取的步骤中，需要选择一些提取算子的参数，如窗口大小、区块大小、细胞单元大小、高斯平滑参数等。经过实验测试，最终选择的参数为：窗口大小 48×48，窗口最大方法倍数为 64，区块大小 16×16，区块跨距 8×8，细胞单元大小 8×8，高斯平滑参数为 4。

另外，SVM 还有一些参数需要用户定义。考虑到样本可能存在线性不可分的情况，在实现中采用 RBF 核的 OVR-SVM。RBF 核的定义如下：

$$K(x,z) = \exp(\gamma||x-z||^2) \tag{14}$$

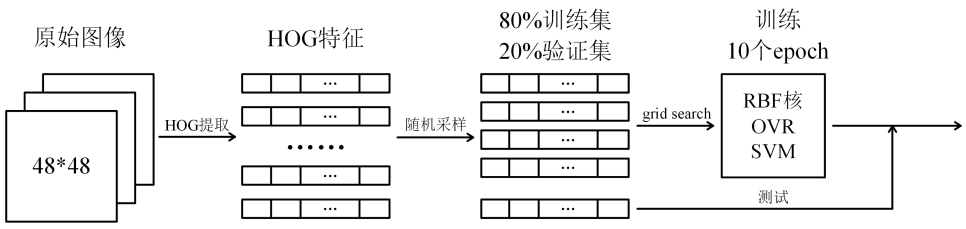

Figure 8: SVM 分类流程图

它包含两个超参数：γ，核函数自身的参数；C，惩罚系数，也就是松弛变量的系数。确定两个超参数的最优值的方法是 grid search：将两个参数的取值范围网格化后自动进行网格搜索，并选取训练效果最好的一个。在训练过程中，为了保证模型的泛化能力，使用交叉验证（Cross-Validation）的方法，每训练一个 epoch 之前都会将所有的训练样本打乱顺序并随机采样，80% 作为训练集、20% 作为测试集，保证所有的数据都能够被充分使用。训练 10 个 epoch，得到 SVM 模型参数。SVM 分类流程图参见图 8。

4.4 ResNet 识别

由于 ResNet 不是某一种固定的网络架构，而是更为零活的残差单元，因此实际意义上的 ResNet 可以有很多种方法。ResNet 的分类效果与其层数有密切的关系，层数越多效果越好。目前比较成熟的 ResNet 分类网络有三种，分别为 ResNet-50，ResNet-101 和 ResNet-152 模型。综合考虑模型分类能力、计算设备限制等因素，我们在本次实验中采用 ResNet-50 模型。由于原本的 ResNet-50 模型用于 ImageNet 分类，其末端的分类器有 1000 个神经元。为了使 ResNet-50 能够用于表情识别问题，需要将最后一个全连接层的输出个数改为 8，对应 8 种情绪。

考虑到 ResNet-50 模型层数较多，参数数目巨大，因此需要较多的训练数据才能保证训练过程收敛，负责容易陷入局部极小值，分类效果极差；或者甚至会出现损失函数发散的现象。但是目前能够采集到的训练数据远远不够，因此需要通过预训练的方法，使得网络参数初步逼近最优解，之后通过带标签的小数据集进行参数微调，让网络参数能够更加精细地趋近最优解。

在实际网络的预训练中，包含两个阶段。第一阶段使用了 FER 数据库中的带标签数据进行预训练。FER 数据库中包含 30000 余张带表情标签的图片，但是由于其中包含图片的人脸角度、光照环境等变化极大，因此不适合用来做微调数据。为了保证预训练数据的规模，需要使用切割（crop）、旋转（角度小于 5°）等数据扩展操作。最终用于预训练的数据集中包含 900000 张图片作为训练集、300000 张图片作为验证集（为避免出现验证集数据与训练集数据过于相似导致验证集准确率虚高的现象，在进行数据扩展操作时先将原始数据集分离为训练集和验证集，之后分别对两个集合进行相同扩展操作）。第二阶段使用了另外四个学术公开数据库的合集进行粗略的参数微调。在此阶段中，将原始数据分为验证集和测试集，并分别对两个集合的数据进行翻转操作以实现数据增强。最终用于第二阶段预训练的数据集中包含 6000 张图片作为训练集、2000 张图片作为验证集。

两个阶段的预训练结束之后，使用助教提供的数据集中的训练集部分进行最终的参数微调。不同于预训练使用的静态数据增强（对图片进行操作之后把所有得到的图片保存下来作为训练数据），在参数微调阶段使用的的是动态数据增强方法：每次读取数据时，添加随机旋转、随机切割、随机镜像、随机对比度改变、随机模糊等一系列数据增强的方法。由于所有操作都带有一定的随机性，因此每次读取图片操作之后得到的训练数据几乎可以确定是独一无二的。通过这种方式进行数据扩展不仅能够减少图片运算和保存的时间开销，而且可以通过设置足够多的迭代次数实现非常强的数据增广，对于提高分类效果很有帮助。

以上训练过程中，batch size 均设置为 10。预训练两个阶段中输入图片大小为 224×224，微调阶段中输入图片大小为 256×256，经过随机切割后变为 224×224。预训练第一阶段训练次数为 20 个 epoch，第二阶段次数为 100 个 epoch，参数微调阶段次数为 100 个 epoch。ResNet 分类流程图参见图 9。

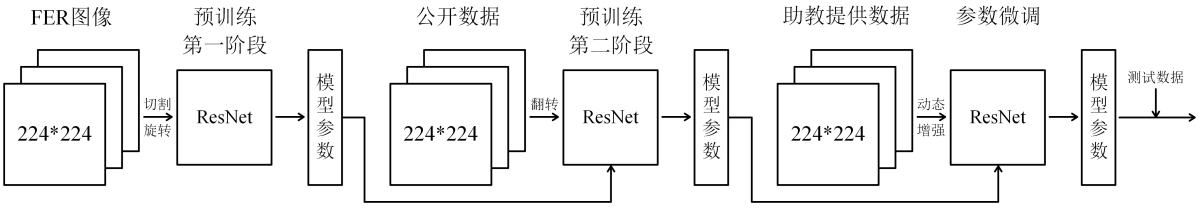

Figure 9: ResNet 分类流程图

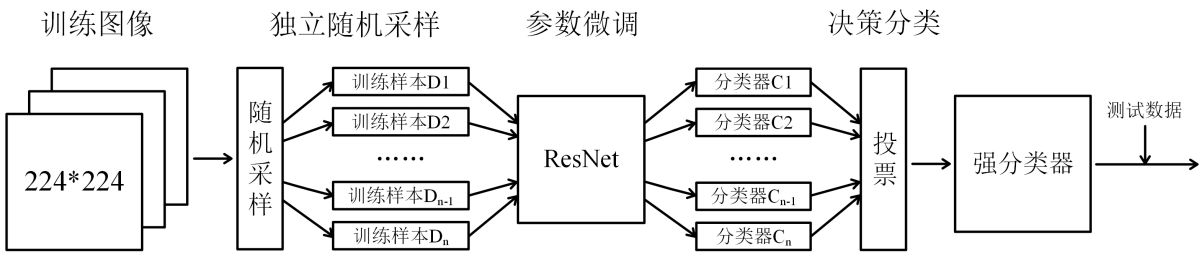

Figure 10: Bagging 分类决策流程图

4.5 分类决策

我们在分类时采用的决策方法是集成学习法（Ensemble learning），构建多个学习器来共同完成学习任务，应用训练数据构建一组基分类器，然后通过对每个基分类器的预测进行投票来进行分类。通常一个集成分类器的分类性能会好于单个分类器，以此来提高分类的准确率。

集成学习方法又可分为两类，一种方法个体学习器间存在强依赖关系，如 boosting 方法，必须串行生成的序列化方法；另一种方法是并行化方法，以 bagging 方法为代表。此次实验采用了 bagging 方法。Bagging 方法是一种以均匀分布从数据中有放回抽样的技术，在每次随机抽样生成的样本集上训练一个分类器，最终得到多个分类器，决策时对训练过的分类器的进行投票，最终置信度最高的类即为样本所属的类。这一方法降低了基分类器的方差并以此改善了泛化误差，其流程图参见图 10。

5 实验结果

5.1 LBPH 识别结果

线性分类器在测试集上的的准确率可以达到 79.9%，但对我们后续的程序来说，这样的准确率对更为“模棱两可”的图片识别和来源更为广阔的测试集准确率一定会有下降，所以这种方法被小组放弃。

5.2 SVM 识别结果

对于公开数据集，其识别准确率为 92.27%。对于给定的自采数据集，我们将训练集的前 80% 进行标定、训练，用已经得到的支持向量机模型对后 20% 进行预测，得到正确率为 33.7%。可见，这样的方法在训练集上也无法达到很好的拟合，故我们最终采用了精度更高的深度学习网络的方法。

5.3 ResNet 识别结果

对于公开数据集，在训练过程中验证集上的准确率最高为 96%，在测试集上的准确率高于 97%。对于自采数据集，由于在微调阶段使用的训练数据包含一部分自采数据，因此训练得到的模型对于自采数据的泛化能力也比较强。最终的测试结果为 66.04%。

6 讨论

此次实验中还存在许多不足之处。在数据预处理部分，MTCNN 无法处理图片中脸部过大的情况；在 LBPH 分类和 SVM 分类部分，没有深入研究分类方法和训练技巧；在深度学习部分，两次预训练采用静态增强的方法使得数据扩展能力受到很大限制，而且 ResNet 模型层数较多，占用显存空间较大，因此在进行 bagging 决策时对计算设备的显存要求很高。

参考文献

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *ECCV*, 2004.

[2] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, 1992.

[3] Li-Fen Chen and Yu-Shiuan Yen. Taiwanese facial expression image database. 2007.

[4] Lundqvist D., Flykt A., and Ohman A. The karolinska directed emotional faces - kdef. 1998.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893 vol. 1, 2005.

[6] Evelyn Fix and J L Hodges. Discriminatory analysis, nonparametric discrimination,consistency properties. *Randolph Field, Texas, Project 21-49-004, Report No.4*, 1951.

[7] Ian J. Goodfellow, Dumitru Erhan, and et.al. Challenges in representation learning: A report on three machine learning contests. *Neural networks : the official journal of the International Neural Network Society*, 64:59–63, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[10] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[11] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *FG*, 1998.

[12] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:971–987, 2002.

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016.