基于协方差池的面部表情识别及其教学监督应用

周子涵 李 丹
四川大学锦城学院 四川 成都 610065

【摘 要】二十一世纪以来，伴随着人们生活质量的提高，信息传递的方式也是不断的改革变化。其中，图像传递尤为重要。因此，人脸识别、面部表情识别等技术的运用领域也越来越广。将人脸表情进行分类需要抓取脸部的一些关键点。但是传统的 CNN 对面部关键点的捕获效果一般，那么如何更好地获得面部特征中的关键点呢？本文将研究协方差这种二阶统计量 [1] 在抓取面部关键点的优势。因此，本文主要介绍了基于 CK+ 和 SFEW2.0 数据集下的协方差池面部表情识别系统，并将其应用到了教学领域，进行教学监督。

【关键词】表情识别；协方差；教学监督

1 背景

1.1 研究背景及意义

目前，面部表情识使用的都是一些标准的网络结构，如 CNN、ResNet、VGG 网络等等。这其中很多识别系统都是在 fer2013 这个数据集上进行训练模型的。也有使用将多个卷积神经网络集成的训练方法：比如，在 2015 年的自然场景表情识别比赛中，冠亚军 [2] 都将多个卷积神经网络进行集成，然后对 fer2013 训练，并获得了较好成绩。在论文 [3] 中，作者在单层网络结构下得到了目前最高的识别精确率：54.82%。上面这些网络结构使用的都是传统的神经网络体系，它们只能抓取一阶的统计量。有人在此基础上进行了进一步的研究，发现类似于协方差这样的二阶统计量可以更好的反应特征之间的关系。文章 [4] 和文章 [5] 就是将协方差池作为描述符来提取数据特征。文章 [6] 提出了多种网络结构下的协方差池模型。文章 [7] 首次提出了可以与协方差池像结合并且用于黎曼流形的网络结构。因此，本文使用 SFEW2.0 和 CK+ 数据集，在协方差池与流形网络相结合的卷积神经网络上进行训练，并将训练后的模型加以封装，提供可视化界面，将其应用于课堂的教学监督上。

2 技术要领

2.1 CNN

本文主要选用的网络模型为 CNN。近些年，CNN 发展势头迅猛，应用范围也非常广泛。CNN 与传统的神经网络结构类似，只不过它将传统的三层结构细化成了卷积层、池化层、激活函数层和全连接层这几层网络结构。

2.2 协方差池

2.2.1 协方差

如果你想衡量两个随机变量的相关程度，那么就需要用到协方差，它通过公式（1）得到：

$$Cov(X,Y) = E(XY) - E(X)E(Y) \quad (1)$$

2.2.2 协方差矩阵

就像之前在研究现状中提到的一样，由卷积层、池化层和全连接层组成的传统卷积神经网络只能抓取普通的一阶统计量。虽然有能够引入非线性结构的激活函数“ReLU”，但是它也只能在单个像素级上引入。而协方差矩阵是从特征中计算得来，它能比一阶统计量更好地抓取区域特征。给定一个特征集，协方差矩阵能够简洁地提取这个特征集中的二阶信息。如果 f1,f2,……,fn ∈ R d 是这个特征集，可通过公式（2）得到协方差矩阵 C：

$$C = \frac{1}{n-1}\sum_{i=1}^{n}\left(f_i - \bar{f}\right)\left(f_i - \bar{f}\right)^T, \quad (2)$$

$$\text{where } \bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$$

只有在 f1,f2,……,fn 中的线性独立数字成分大于 d 时，才能得到对称正定的矩阵。即使这个矩阵只是正半定的，也能通过公式（3）来将其正则化：

$$C^+ = C + \lambda \mathrm{trace}(C) I \ , \quad (3)$$

其中，λ 是正则化参数，I 是恒等式矩阵。

其实所谓的协方差池就是用协方差矩阵算法来代替传统 CNN 常用的最大或平均池化中取 max 或平均值，进行池化，提取特征信息。

2.3 流形网络

本节是对本文所提及到的流形网络的解释和其中一些结构的介绍。

2.3.1 黎曼流形

黎曼流形是一种用于研究空间图像的重要工具 [8]。本文使用了黎曼流形中的 SPD 流形网络来与协方差池构成 CNN 的升级模型。

2.3.2 SPD 流形网络

如果我们直接将协方差矩阵展开或者应用到全连接层的话，可能会导致重要信息的丢失。解决此类问题的标准方法是：使用对数运算将黎曼流形展开，让欧几里得空间 [9][10] 的标准损失函数能在它上面运用起来。文章 [11] 引入了三层结构来对 SPD 矩阵进行降维，以便之后使用损失函数对矩阵进行操作。下面，我们简单介绍一下 [11] 中引入的三层结构。

（1）双线性地图层（BiMap）

协方差矩阵在进行降维后不能直接用到全连接层。此外，当降维时双线性地图层对保护集合结构也有重要作用。双线性地图层能在传统的 CNN 上完成这些任务。如果 SPD 的输入矩阵是 Xk−1，在满秩矩阵空间中 $W_k \in R^{d_k * d_{k-1}}$ 是权重矩阵并且 $X_k \in R^{d_k * d_k}$ 是输出矩阵，则第 k 个双线性地图层的 f_b^k 可以表示为：

$$X_k = f_b^k(X_{k-1};W_k) = W_k X_{k-1} W_k^T, \quad (4)$$

（2）特征值整流层（ReEig）

特征值整流层用来引入非线性，其作用类似于 ReLU 函数。如果 SPD 的输入矩阵是 X_{k-1}，输出矩阵是 X_k，并且 I 的系数是整流限度值，则第 k 个特征值整流层的可以定义为：

$$X_k = f_r^k(X_{k-1}) = U_{k-1}\max(\in I,\sigma_{k-1})U_{k-1}^T, \quad (5)$$

max 运算是矩阵运算，特征值分解 $X_{k-1}=U_{k-1}\sum_{k-1}U_{k-1}^T$ 得到 U_{k-1} 和 $\sum_{k-1}$。

（3）对数特征值层（LogEig）

如前文所述，黎曼流形与 SPD 矩阵密不可分，对数特征值层会给黎曼流形中的元素赋予一种 Lie's Transformation Groups 结构 [12]，该结构能使矩阵降维，然后就能够使用标准的欧氏运算进行操作。如果 X_{k-1} 是输入矩阵，X_k 是输出矩阵，则在第 k 层中的 LogEig 层的 f_l^k 可以定义为：

$$X_k = f_l^k(X_{k-1}) = \log(U_{k-1}) = U_{k-1}\log(\sum_{k-1})U_{k-1}^T, \quad (6)$$

最后将 BiMap 层和 ReEig 层合并使用，用 BiRe 来表示。构成具有 2-BiRe 层的 SPD 流形网络结构如图 1 所示。

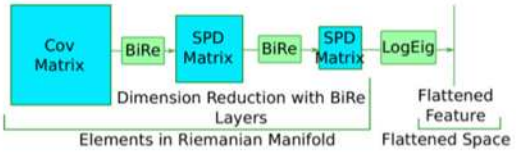

图 1 具有 2-BiRe 层的 SPD 流形网络示意图

3 实验

3.1 数据集

本文选取了 CK+ 和 SFEW2.0 两种数据集进行实验，通过数据集预处理、加入协方差与流形网络结构、训练模型、比较训练结果，最后得出结论。

3.1.1 CK+ 数据集

CK+ 数据集是多名志愿者在实验室中拍摄的 593 张图片，共有为中性、开心、害怕等七种表情。这些表情是每一名志愿者表情从中性（neutral）到激动的一个变化过程。它的表情分类标签是存储到 txt 文件中的。

3.1.2 SFEW2.0 数据集

SFEW 2.0 中包含 1394 张图片，其中 958 张用于训练，436 张用于验证，与 CK+ 一样包含 7 种表情。该数据集是从多部电影片段中选取有表情的某些帧截图构成的，并且采用基于零件混合模型的特征点检测方法 [13]，然后用这些获取的标记点进行校准。

3.2 训练过程

3.2.1 人脸检查

人脸检查及定位部分，本文使用的是 python 中的第三方视觉库 OpenCV[14] 提供的人脸识别模块以及写好的人脸检测分类器和人眼检测分类器。OpenCV 利用 Adaboost 对图片进行分类 [15]，得到算法制定规则的强分类器模型，之后就能通过此模型进行人脸的检查和定位。

3.2.2 数据集预处理

（1）灰度化。由于 SFEW2.0 大部分图像和 CK+ 少部分图像都为 RGB 彩色图像，不便于之后的其他操作。所以首先需要将这些 RGB 图进行灰度化，将其调整为灰度图。

（2）直方图均衡化。由于光线强度和人的肤色不同，两个数据集上不同图片的亮度也有所区别。运用到直方图均衡化，将图片的亮度调节到基本相似的程度。图 2 为是否运用直方图均衡化 [16] 来处理图像的对比图。

图 2 直方图均衡化对比图

（3）裁剪（归一化）

在 SFEW2.0 中，人脸只占整张图像的极小一部分。

为了确保人脸识别的速度与准确率，我们需要将图片进行裁剪。然后将裁剪完后的图片数据进行归一化，再将其转换为数据矩阵，便于之后的操作。

3.2.3 加入协方差池和流形网络进行训练

在确定面部位置、图像预处理后，接下来的工作就是将归一化的图像提供给 CNN。为了从卷积神经网络中池化出空间特征，本文使用协方差池与流形网络相结合的结构来深度学习二阶统计量。图 3 展示将协方差池和流形网络与 CNN 相结合的模型原理。

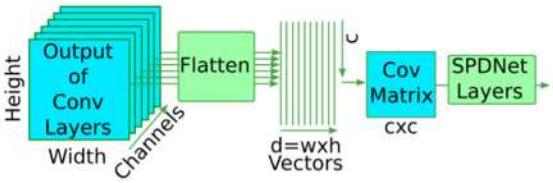

图 3. 扁平化卷积层的输出

为了证明协方差池能够提高模型准确率，本文在最后一层卷积层与全连接层之间使用协方差池与流形网络层，之后再对各种模型进行微调，然后进行实验。表 1 总结了微调后各种模型的细节（Model–0 为没有加入协方差池与流形网络的基础模型）。

表 1 考虑协方差池的各种模型（为了简洁，已忽略初始卷积层）

Model-0	Model-1	Model-2	Model-3	Model-4
	Cov	Cov	Cov	Cov
	BiRe	BiRe	BiRe	BiRe
	LogEig	LogEig	BiRe	LogEig
			LogEig	
FC2000	FC2000	FC2000	FC2000	FC2000
FC7	FC7	FC128	FC7	FC512
		FC7		FC7

3.3 结果分析

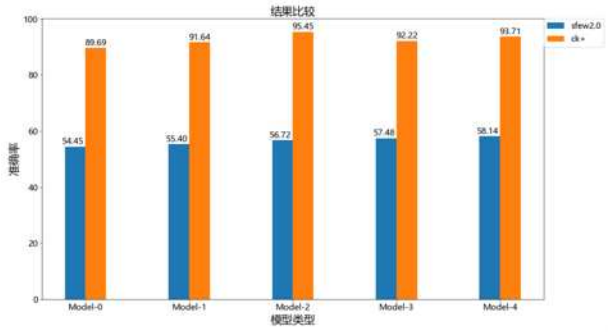

图 4 健康具有和不具有协方差池的两种数据集 5 种模型的识别准确率

表 1 中描述的各种模型及其准确率见图 4。通过对比观察，我们可以得到如下结论：

（1）由于 CK+ 数据集中的图像数据是从室内固定环境拍摄得到的，与 SFEW2.0 从电影中截取出的图像相比，外界背景干扰更小、表情更加丰富、图像处理更为方便，因此 CK+ 的准确率要明显高于 SFEW2.0。

（2）对于 SFEW2.0，它在 Model–4 上的准确率最高；对于 CK+，它在 Model–2 上得到最高准确率。

（3）无论是 SFEW2.0 还是 CK+，Model–0 的准确率都要低于其他的模型，因此可以判断协方差池加流形网络的结构确实能够提高模型的准确率。

4 应用——教学监督系统

学生在上课时，正常情况下的面部表情应该是 neutral 状态。如果有其他表情出现，则大概率没有认真听课。因此，本文将面部表情识别系统应用到课堂教学监督方向上来。

选择准确率为 95.45% 的 CK+ 数据集在 Modle–2 上训练得到的模型作为教学监督系统的模型。本文还为表情识别系统增加了 UI 界面，效果及预测结果如图 5 所示：

图 5 识别结果预览图

在平时上课时可开启实时摄像头识别，对学生进行实时监督，以此来提高学生听课效率和课堂教学质量。

5 结束语

本文将 SFEW2.0 和 CK+ 数据集作为研究对象，使用协方差池与流形网络结构与卷积神经网络相结合的模型实现了对人脸面部表情的识别，最终在 SFEW2.0 和 CK+ 数据集上分别得到了 58.14% 和 95.45% 的识别准确率。通过比较没有协方差池结构的 Model–0 与其他加入了协方差结构的模型的训练结果，证明了引入协方差池与流形网络结构来提高模型的识别准确率的方法是可行的。最后选择准确率为 95.45% 的 CK+ 数据集在 Modle–2 上训练得到的模型作为教学监督系统的模型，并为识别系统增加了 UI 界面，初步完成了课堂表情识别系统。

【参考文献】

[1] K. Yu and M. Salzmann. Second-order convolutional neural networks. CoRR, abs/1703.06817, 2017. 1, 2, 3, 7

[2] B.-K.Kim,H.Lee,J.Roh,and S.-Y.Lee.Hierarchicalcommitteeof-deepcnnswithexponentially-weighteddecisionfusion for static facial expression recognition. In Proceedings ofthe2015ACMon-InternationalConferenceonMultimodal Interaction,ICMI'15,pages427 - 434,NewYork,NY,USA, 2015. ACM. 2, 5

[3] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, pages 435 - 442, New York, NY, USA, 2015. ACM. 2, 5

[4] J. a. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In Proceedings of the 12th European Conference on Computer Vision - Volume Part VII, ECCV'12, pages 430 - 443, Berlin, Heidelberg, 2012. Springer-Verlag. 2, 3

[5] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classi-cation. In Proceedings of the 9th European Conference on Computer Vision -Volume PartII,ECCV'06,pages589 - 600,Berlin,Heidelberg,2006. Springer-Verlag. 1, 2, 3

[6] K. Yu and M. Salzmann. Second-order convolutional neural networks. CoRR, abs/1703.06817, 2017. 1, 2, 3, 7

[7] Z. Huang and L. V. Gool. A riemannian network for spd matrix learning. In AAAI, 2017. 1, 2, 3, 7

[8] 许春燕 . 基于黎曼流形的图像分类算法研究 [D]. 武汉 : 华中科技大学 ,2015.

[9] J. a. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In Proceedings of the 12th European Conference on Computer Vision - Volume Part VII, ECCV'12, pages 430 - 443, Berlin, Heidelberg, 2012. Springer-Verlag. 2, 3

[10] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classi-cation. In Proceedings of the 9th European Conference on Computer Vision - Volume PartII,ECCV'06,pages589 - 600,Berlin,Heidelberg,2006. Springer-Verlag. 1, 2, 3

[11] Z. Huang and L. V. Gool. A riemannian network for spd matrix learning. In AAAI, 2017. 1, 2, 3, 7

[12]Sophus Lie's Transformation Groups: A Series of Elementary Articles. 1897, 4(12):308-313.

[13] X. Zhu and D. Ramanan. Face detection, pose estimation and-landmarkestimationinthewild. InIEEEConferenceon Computer Vision and Pattern Recognition (CVPR), 2012. 4

[14] 薛同来 , 赵冬晖 , 张华方 , 郭玉 , 刘旭春 . 基于 Python 的深度学习人脸识别方法 [J]. 工业控制计算机 ,2019,32(02):118-119.

[15] 林志健 , 周设营 , 陈延清 . 基于 OpenCV 的人脸识别关键技术分析 [J]. 中国新技术新产品 ,2020(07):15-16.

[16] 李宽 . 基于浅层卷积网络的人脸表情识别方法研究 [D]. 合肥 : 中国科学技术大学 ,2019.