摘　要

从上世纪90年代开始人脸识别领域的科研工作者们就一直专注于寻找适合计算机理解的人脸的有效表达。在2005年之前，绝大多数的人脸表达都是基于底层特征的，这段时间的代表性的成果包括几何结构特征、子空间技术、小波编码和局部描述子等。虽然这些特征种类众多，并且对人脸识别的发展起到了不可磨灭的作用，但是底层特征这种方式并不完反映合人类视觉系统的机理。在人类的视觉系统中，可视信息处理是分级的进行。其中，眼睛负责接受视觉信号，然后传送给主视觉皮层V1区。主视觉皮层V1区针对视觉信号产生一系列的点和边缘的感受器，把视觉信号转化成点和边缘信号，再将它们汇聚到视觉皮层V2区。在视觉皮层V2区根据点和边缘信号产生了一些列的动作检测器、立体检测器以及颜色检测器，将信号转化成更强的关系并汇集到更高的视觉皮层。同时视觉皮层V2区也向主视觉皮层V1区反馈从高层视觉皮层而来的信号。信号根据这种方式最终进入最高层，形成帮助人类理解的有效的语义信息和行为信息。得益于丰富多彩的底层特征表达，从2005年开始，人脸识别领域的科研工作者们开始模仿大脑处理视觉信息的方式构造了大量的人工大脑。这些人工大脑从图像的底层特征中抽取有用信息，组合成越来越抽象并且越来越具有区分能力的高层特征。在这些人工大脑中，最卓越最为人所知的是Fisher向量编码和深度卷积神经网络这两种模型。Fisher向量编码是目前词袋模型中性能表现最好，概括力最强的模型，而深度卷积神经网络是深度模型中最适合处理图像与视频信息的模型。作为最前沿的模型，Fisher向量编码和卷积神经网络不断的在各个应用场景上刷新性能，甚至在人脸识别领域和物体识别领域形成了各占半壁江山的局面。因此，本文对Fisher向量编码和深度卷积神经网络进行深入的研究，在不同场景下尝试大量的评测，提出一些实用的经验，并对性能和速度进行改进。本文的主要工作包括：

在Fisher向量编码方面：本文研究了词袋模型和Fisher向量编码的基础理论，并从层次结构方面重新解释了Fisher向量编码；本文提出了Fisher向量编码的一种实现技巧，并与普通Fisher向量编码的实现做了对比；本文论证了Fisher向量编码在SIFT特征上归一化和Fisher编码上归一化的重要性；本文比较了不同提取方式的SIFT特征和位置信息等底层特征对Fisher向量编码性能的影响；另外，本文还从理论和实践上论证了在Fisher向量编码的早期阶段加入判别信息，能够增强混合高斯模型的判别性，进而进一步提升Fisher向量编码的性能；最后，本文研究了在资源受限的条件下，通过主成分分析法和最大边缘距离度量学习法从高维度Fisher向量编码中学习一种具有判别性的低维度表示，不仅降低了存储和计算成本，还提升了Fisher向量编码的性能。

在深度卷积神经网络方面：首先，本文详细的比较了最前沿的两种深度模型，深度信念网络和深度卷积神经网络，并从理论和实践上论证了在图像处理和视频处理

上深度卷积神经网络具有更强的建模能力和推广能力；在速度方面，本文通过寻找最优的数据存储方式、利用单指令多数据流做并行计算、利用矩阵乘法优化卷积操作、合并最大汇聚层和ReLU激励函数等方法，将深度卷积神经网络特征提取的速度提升了10倍左右；此外，本文还构造了从简单到复杂的深度卷积神经网络的级联，让简单模型学习大量普通数据，让复杂模型学习少量奇异数据，达到了不同能力的模型拟合不同难度数据的目的；另一方面，本文还提出了多准则与多返回流的深度卷积神经网络，它能够更好的稳定训练的收敛过程，并一定程度上提升了深度卷积神经网络的能力。

综上所述，本文针对层级特征表示最前沿的两个模型Fisher向量编码和深度卷积神经网络做了广泛的调研和深入的研究，有针对性的从速度和性能上提出了若干经验与改进方法，并进行了有效的验证，一定程度上提升了这些模型在人脸识别和人脸确认的性能。最后本文还提出了几个未完成的开放式问题与猜想，以供进一步研究与学习。

关键词：人脸确认， 词袋模型， Fisher向量编码， 深度学习， 深度卷积神经网络

Face Recognition using Hierachical Feature Representation

Fang Zhenpeng (Computer Application Technology)
Advised by Professor Shan Shiguang

Since 1990s, researchers in the field of face recognition have focused on finding effective representations of human faces, which are more understandable for computers rather than human beings. Before 2005, almost all the face representations are based on low-level features. Some representative works in this time are geometric features, subspace technology, wavelet coding and local descriptors. Though, there exist many types of low-level features, and some of them indeed play incredible roles in the field of face recognition. However, low-level representations are not fully consistent with the mechanism of human vision. In the human vision system, visual signal is hierarchically processed. Eyes are responsible for receiving visual stimulation. The primary visual cortex which also known as V1, receives information from visual field, and generates point and edge detectors to transmit visual signals into points and edges, and finally send them to higher level visual cortex. Visual area V2, receives forward point and edge information from V1, generates motion detectors, stereoscopic depth detectors and colour detectors, and send those strong connections to higher level cortexes. At the meanwhile, it also sends strong feedback connections to V1. Signals go forth in the same manner, and finally are aggregated into semantic representations and meaningful behaviors, which help humans know the world. Thanks to the booming development of low-level representations, high-level representations come into fashion since 2005. researchers mimic the mechanism of human vision, create lots of artificial brains. Those brains extract information from low-level features and combine them into more abstract and more discriminative high-level features like human beings. Two of the best known and standing out artificial brains are Fisher Vector, which is the representative work of bag-of-visual-words model, and Deep Convolution Neural Network, which is more suitable for images and videos compared to other deep learning model. As the main strength, they continuously update performance on various application scenarios, and even form the two-way race situation in the field of face recognition and object recognition. Therefore, in this paper, we in-depth study fisher vector and deep convolution neural network, and conduct a larger number of evaluations in various configurations, and try to point out some useful experience, and go a step further to make some improvements and integrations.

In the aspect of Fisher Vector: First, a brief methodology of Bag-of-Visual-Words and Fisher Vector is discussed, and Fisher Vector is re-explained from the hierarchical model perspective. Secondly, an implementation trick of Fisher Vector is proposed, and the tricky implementation is compared to the original one. And then, the importance of SIFT normalization and fisher vector encoding normalization are paid much attention. After that, different ways of local feature extraction including SIFT and location information are compared mutually. Moreover, discriminative information is added into the early stage of fisher vector to increase the discriminability of Gaussian Mixture Model and improve the performance of face verification. Finally, two kind of feature reduction approaches are proposed. Principal component analysis and large margin metric learning are used to learn a low-dimensional representation of the original high-dimensional fisher vector in resource-limited conditions, and further boost speed and performance.

In the aspect of Deep Convolution Neural Network: First, two kind of the-state-of-the-art deep models, Deep Belief Network and Deep Convolution Neural Network, are compared carefully. Deep convolution neural network is proved from theory and practice, that it has a stronger ability to adapt data and an easier capability of generalization in handling images and videos. On the speed hand, this paper seeks a better data storage manner to ensure the continuity of memory address, uses single instruction multiple data technique to do parallel computation, substitutes convolution by matrix multiplication and merge maxpooling layer with ReLU layer. With all the technique above, deep convolution neural network is accelerated by ten times of the original version. What's more, a schema of cascading deep convolution neural network from simple to complex is proposed to learn simple deep networks for abundant ordinary data and complex deep networks for small amounts of singular data, and make different networks adapt to different data with the corresponding capacities. This paper also proposes multiple guidelines and multi-backward steams deep convolution neural network, which speeds up the training process and boost the performance of deep convolution neural network.

In conclusion, this work studied two of the most advanced works in face recognition, those are fisher vector and deep convolution neural network. And this work proposed some improvement and pointed out a lot of meaningful experience, and did extensive experiments, and enhanced the performance of face recognition and face verification. In the end of our work, several interesting and open problems for further research are given.

Keywords: face recognition, bag of visual word, fisher vector, deep learning, deep convolution neural network