本科毕业论文

题　目　基于深度神经网络的手语识别技术

英　文　Handshape Recognition Based on

题　目　Deep Neural Network

院　系　少年班学院

姓　名　尹沛劼　学　号 PB09000615

导　师　李厚强　教授

日　期　2013 年 5 月 24 日

致谢

在本文的完成过程中，真挚地感谢李厚强老师的指导，李老师帮我确定了手语识别的大方向，支持我对深度学习的各个方面进行了系统的学习与研究，并鼓励我积极的进行改进和创新。感谢唐傲师兄领导的科大手语识别研究组的全体同学，特别是与我共同工作学习的黄杰同学，你们对我的帮助和支持对我这篇文章的完成意义重大。

在本人的大学四年生活中，感谢各位任课老师对我的教诲；感谢班主任倪晓玉老师对我的关心；感谢室友们对我的包容与照顾；感谢我的女朋友季梦晨大人在生活上给了我体贴入微的关怀，在事业上帮我找到奋斗的目标和动力，并不厌其烦的督促我尽快动笔，让我按时完成了本文。

最后，感谢在我二十二年生命中始终在我身边陪伴我的父母，没有你们就没有我今天的一切。

此文献给我的大学四年，和我最爱的人们。

目录

摘要

自 2006 年 Hinton 等人发表了关于深度学习（Deep Learning）的新型算法[30]以来，带有多隐层的神经网络研究再次得到了学术界的重视。本文介绍了深度神经网络的基本概念和基本思想；介绍了目前流行的两种深度神经网络结构：卷积神经网（CNN）和深度置信网（DBN）；介绍了卷积神经网的卷积训练算法与权值共享的思想；介绍了构成 DBN 的基本网络限制波尔兹曼机（RBM）和训练 DBN 的无监督贪婪逐层学习算法；并提出了将两种神经网络结合并应用在手语分类中的具体方法。

CNN 作为长期以来神经网络的佼佼者，以基于完善的生物学背景著称，在手写数字识别[28]、OCR[29]得到了广泛的应用。DBN 作为一个强大的概率生成模型，已在多个领域取得了达到或击败 state-of-art 的结果，如降维[20]，3D 物体分类[21]、语音分类[22]、回归[24]，高维时间序列建模[26]、图像转换[27]等等。但是 CNN 和 DBN 在手语识别方面的运用还非常罕见。本文介绍了 CNN、DBN 在手形分类上的应用。

本文介绍的手语子单元（Sub-Units）识别系统以模块化架构整合手语识别中的各个部分。手形分类作为整个系统的前端模块直接处理手形的原始数据，提取手形特征并分类，可以为语义识别模型提供手形信息。

与过去依靠人工提取特征的方法不同，深度神经网络模拟人的视觉系统，能够直接处理图像像素信息，网络自动提取层次特征，从低层到高层的特征表示（representation）抽象度逐步提高，不同类型的数据在特征空间的距离也越明显，有利于进一步降噪和分类。实验证明 DBN 和 CNN 在手形分类上均取得了较高的分类效果。

本文的最后，作者给出了结论并展望了神经网络在手语识别中的进一步应用。

关键字：卷积神经网络，深度置信网， 限制波尔兹曼机，手形分类，手语识别

Abstract

The new learning algorithm of Deep learning proposed by Geoffrey E. Hinton initiated a wave of research on neural network. This paper introduces the basic concepts and basic ideas of deep neural network. Furthermore, it introduces the two most popular structure of Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Restricted Boltzmann Machine (RBM), which is building block of DBN and unsupervised greedy layer-wised learning algorithm.

While the deep learning has reached or beat the state-of-art result in areas such as Regression, Nonlinear embedding, Reducing Dimension, 3D Object Recognition, Speech Recognition and so on, little attention has been paid to Sign Language Recognition (SLR) yet. This paper show a promising way of applying CNN or DBN to hand shape recognition, an important module in SLR. Up to now, most of the researches in this area was did by manually chosen features, like SIFT and HOG. Although they accomplished some accepted results, they are somehow not good enough, and it costs much time to moderate model for such results.

Deep neural network can simulate human vision system and process pixel information directly. Network extracts abstract feature automatically. As the structure becomes deeper, the feature extracted by network becomes more abstract and expresses semantics more clearly. So it is easier to make recognition. Experiment proves the surprisingly ability of CNN and DBN on hand shape classifications.

Hand shape classification as a sub-module of the whole sign language recognition system can process the raw data of hand shape, extract features and make classification. It also can provide information about hand shape to semantics recognition model like Hidden Markov Model (HMM), then information about hand shape is observation sequence.

At the end of this paper, I draw a conclusion and propose some future work about how to build semantics model based on feature of hand shape.

Key words: Convolutional Neural Network, Deep Belief Network, Restricted Boltzmann Machine, Hand shape classification, Sign Language Recognition

第 1 章 绪 论

1.1 课题背景及意义

全世界共有 5 亿左右的聋哑人，其中，中国就有 2700 万人。聋哑人主要通过手语同他人进行交流、沟通。然而，由于手语的普及程度较差，造成聋哑人与正常人沟通困难,即使是聋哑人之间，也会因为地域差异而难以交流，例如大陆与台湾手语存在较大差异，我国手语与其他国家手语例如美国手语的差异就不言而喻了。但是如果能够通过计算机来自动的分类聋哑人的手语，再将手语翻译以文本或者语音的形式表示，将具有重要的现实意义。

人机交互技术的研究是计算机技术研究领域的重要组成部分。当人与人进行面对面的通讯时, 包括口语及书面语等自然语言与包括手语、表情、体势及口型等人体语言同时或联合传递信息。因而研究人体语言的感知模型及其与自然语言的信息融合, 对于提高计算机自然语言理解水平和加强人机接口的可实用性是极有意义的。手语分类作为人体语言理解的一部分, 有着非常重要的作用。一方面, 它是虚拟现实人机交互的主要手段; 另一方面它又是聋哑人利用计算机与正常人交流的辅助工具。每个手语是由一个手势序列组成, 而每个手势是由手形(posture) 变化序列组成。根据手语输入介质的不同, 手语分类系统可分为两种: 基于视觉的手语分类系统和基于设备输入(如数据手套、位置跟踪器等) 的手语分类系统。本文所处理的数据是基于视觉获取的

1.2 研究现状

目前，一般手势分类多基于小词汇集，有时为简化问题需要加入人造手势。手语是固定的大词汇集、语法约束完备的手势集，以手语分类为开端，可以为将来研究普遍的手势分析工作积累经验。手语是一个多模态并发的信号，既包括手部信息，如手形、朝向、位置等，还包括身体其他部分的运动，例如：人脸表情、头势、躯干运动等，手语分类需要同时处理这些并行信息，对它们进行同步、融合。因此，深层次理解手语的本质有助于人体运动和行为的分析和理解，例如：人脸和面部表情分类、人体跟踪与人体运动分析、体势分类等。手语的高度结构性使得它可以作为一个很好的研究平台，国内外都有不少学者正在进行深入的研究。

1.2.1 国内研究现状

台湾大学的 Liang 等[1]专注台湾手语分类的研究，其实现实时连续手语分类的研究基于数据手套所采集的数据，利用姿势、位置、方向和动作等特征参数，

51个自定义的基础姿势，实现基于自定义的250个词汇的分类，分类率达到80.4%。Liang 与 Ming 建立了美国[1]手语字母连续流的分类系统，该系统利用 VPL 型号数据手套作为手势输入设备，采用窗口模板匹配策略实现实时连续分类，可分类美国手语中 26 个字母的连续字母流。

哈尔滨工业大学计算机系利用边缘检测、神经网络等方法，实现了基于视觉的、可分类 13 个静止手势的手势分类系统和基于视觉的简单的连续变化手势分类面向大词汇量的连续中国手语分类系统的研究与实现系统[2]。该系统不要求用户戴任何特殊手套。

中科院计算所马继勇[3]对基于数据手套的中国手语分类算法进行了研究，提出了与打手语人位置无关的特征提取方法、流捆绑技术、过渡帧模型等技术，实现对自定义词典中孤立词和一些连续语句的分类，但是由于是建立过渡模型，所以对大量词汇的分类速度较慢。

1.2.2 国外研究现状

美国 CMU 的 Christopher Lee 与 Yang [4]开发了一个基于 Cyber glove 数据手套的手势分类系统。该系统利用 HMM 技术对美国手语字母中的 14 个手势进行了分类，这个系统可以作为机器人远程操作及示例规划交互接口的一部分。

日本 ATR 研究室的 Takahashi 与 KI Shino[5,6,7]建立了一个基于数据手套的分类系统，该系统综合使用了主成分分析及聚类分析技术，对手的关节角度及方向进行编码，可分类 46 个日本手指字母中的 34 个。这种方法对动态变化描述能力比较差，只能用于静止手势的分类中，如果分类动态手势，还需要与其他方法结合使用。

Vogler 与 Metaxas[8,9]的美国手语分类系统，采用一个位置跟踪器及三个互相垂直的摄像机作为手势输入设备，利用计算机视觉方法提取打手势者手臂运动的三维参数，利用 HMM 分类技术，完成了 53 个孤立词及 486 个连续手语语句的分类，分类率为 89.91%。此外，他们还将手势词进一步细分为更小的基元“音子”，包括静态和动态两种，然后为这些基元建立模型，对由 22 个词构成的句子的测试结果表明，这种方法的分类率和传统方法的分类率相近。

麻省理工学院（Massachusetts Institute of Technology）的 Starner 等[10]在 1995 年实现了一个基于单目视觉的美国手语分类系统，该系统能够分类由 40 个美国手组成的连续语句。他们采用 395 个句子作为训练数据，99 个句子作为测试数据。在有较强语言模型的情况下取得了 97%的分类率，没有语言模型的情况下取

得了 91%的分类率。

Drew 等[10]用动态规划的方法进行手部跟踪。该方法利用动态规划思想确定目标物体的路径，同时能够对多目标物体进行跟踪。利用整个视频序列的信息能够避免陷入局部搜索目标物体带来的错误跟踪。

Bui 等[11]利用微电子机械系统来分类越南手语中的手势。他们发明了 AcceleGlove 数据手套，该手套包含 5 个手指传感器和 1 个手背传感器。他们用该系统来分类 23 个越南手语中的手势以及两个标点符号—句号和逗号。

1.3 研究趋势与前景

目前语音分类技术已经逐渐趋于成熟，很多方法都得到广泛成功的应用，并且效果良好。但是对于手语分类而言，由于研究起步就比语音晚得多，所以还需很长的路需要探索。近十来年，手语分类才开始逐渐被人们赋予较高的期望。HMM 的植入使得手语分类成为可能，SVM、LDCRF[13]等方法的引入慢慢的给手语分类领域带来希望。人们开始慢慢的进入深一层次的研究，以后得更高的分类率。手语分类的发展趋势以及前景主要可以归结为以下几点：

（1）孤立词汇分类准确率的提升。

孤立词的分类从最开始的不到半成的分类率，到目前为止已经可以实现少量词汇高精度的分类。对于大量词汇的分类还有待加强。

（2）提取好的特征。

语义分类效果的好坏很大程度上依赖于特征提取的好坏，如果能找到具有不变性和可区分性的特征来描述手语，那么分类效果将会有质提升。

（3）过渡手势的分类以及连续手语的准确分割。

分割绝对是现在困扰手语分类研究者的主要问题，也是手语分类的主要瓶颈，正因为此问题不能得到很好的解决，使得手语分类未能成功实现于应用。但此问题将会慢慢得到解决，道路是曲折的。国内，由参考文献[1]可知，台湾科技大学梁容辉等人提出使用 TVPs 模型，国外，由参考文献[11]可知 Pitsikalis 提出使用 sub-units 模型。

（4）最终适用于聋哑人的应用应运而生。

手语分类的最终目的就是为了能够帮助语言障碍的人能和正常人进行无障碍的交流，如若同时结合语音技术，还能实现盲人与聋哑人之间的交流。盲人通过说话，利用应用将语音转化为手语，而聋哑人通过手语利用应用转化为语音，所以应用前景还是值得期待的。

1.4 手语分类存在的问题

手语分类工作不同于语音分类工作，手语分类目前还处于初级阶段，各工作无法直接进行对比，因为没有公认的数据集提供同等比较。但却有一些相同的问题困扰着手语分类的研究者。

之前大部分研究是基于特定人的手语研究，基于特定人的手语分类系统，就是用来训练模型的数据和模型用来分类的数据采集自同一个人的手语分类系统，但这明显不符合实际的应用情形，手语分类系统应该能够对不同人的基本相同的手势进行分类，这种情况下的手语分类研究近几年才开始，而且用不同人的训练数据进行分类大大降低了之前特定人分类研究的效果。一种解决该问题的方法是采集大量的用户的数据，训练一个普适的非特定人模型。然而，非特定人模型的训练会遇到一些问题：

（1）采集非特定人的数据耗时、耗力、且成本又高。

中国手语词典中定义的手语词有 5000 多个，可以想象如果采集多人的 5000 多个词汇的数据需要耗费大量的时间和成本。然而目前基于视觉的，即利用摄像头之类的进行直接提取视频特征的方法还未成熟，对小数据集的效果还行，但是不适用与大数据集的特征提取，所以当需要采集不同人的大数据集时还是要采用传统的数据手套和位置跟踪等比较昂贵又容易损坏的设备，可见数据采集的成本是比较大的。

（2）来自不同人的数据在模型训练时不易收敛。

虽然相同的词义遵循一定的规则，然而受到各种因素的影响，如不同人的手型，采集数据的环境，不同人的身材情况等等，总之不同人做同一动作时的差异都可能会很大。差异较大的模型在训练时是很难达到收敛。

（3）非特定人模型推广能力不强。

即便能够训练得到非特定人模型，由于其模型是由众多非特定人的数据训练得到，其参数的分布必定是平滑的，因此，该模型能够在大多数用户的数据集上取得可以接受的分类结果。但是，对于某一特定用户，其性能与特定人模型相比，差距明显。

（4）其他信号对手语分类的影响

手语是一个多源信号的组合，除去传统意义的双手的手形、位置、朝向和运动轨迹之外还包含一些其他非手部信号，例如：脸部表情、头势、体势、唇动等等，这些信号能够对手部信号做有益的补充，如：某些通过手部信号难以区分的

手语词很容易通过非手部信号加以区分。

（5）运动插入对手语分类的影响

手语在不同情况下会出现变化，主要包括两手语词之间的运动插入和一些数据流的变化所导致的词义改变。在连续手语分类中，两个手语词之间会出现一些不属于任何一个手语词的一些冗余数据，即：运动插入，这些数据给连续手语分类带来很大的困难。通常，可以利用一些特征子集作为线索来显式分割词的边界。

1.5 本文的主要工作

研究学习两种深度神经网络结构—CNN（卷积神经网）和 DBN(深度置信网)，通过文献学习 CNN/DBN 在手语分类中的应用，知道如何搭建一个深度神经网络结构，通过贪婪逐层训练算法训练模型，避免模型收敛局部最优解，也避免模型过拟合,能够达到好的分类效果。本论文框架如下：

第 1 章、绪论。主要介绍了手语分类的背景和意义、国内外现状、存在的问题、发展趋势和本文主要工作。

第 2 章、手语识别系统简介。主要介绍整个手语识别系统的实现基本方法和流程。

第 3 章、深度卷积网和深度置信网，主要介绍了 CNN、RBM（限制玻尔兹曼机）、DBN 的原理，学习算法，训练方法。

第 4 章、基于 CNN/DBN 的手形分类，分析了运用 DBN 实现手形的分类的原理。

第 5 章、实验结果。主要 CNN、DBN 实现手形分类的实验结果以及结果分析。

第 6 章、总结和今后工作展望。

第 2 章 手语识别和手形分类系统简介

本章主要介绍整个手语识别系统的框架和基本流程，系统基本实现了手语识别的功能，但是各模块的功能和效果包括特征的提取，连续句子的分词，孤立词的识别率等都有待提升。本论文研究的手形分类系统旨在为手语识别系统提供一种提取手形特征的方法。

2.1 手语识别系统简介

手语识别系统可分为特征提取，模型训练和识别两个部分。首先对手语视频

提取特征，然后把特征作为训练和识别的输入。识别依赖于预先训练的模型。模型的训练在线下完成，模型训练好之后即可进行实时的手语识别。

2.1.1 特征提取

首先利用 Kinect，跟踪获得手语测试者打手语时的骨骼点位置信息，包括左手、右手、左肘、右肘四点的 3 维坐标一共 12 维；由于不同的人身体形态存在差异，即便是相同的手语动作，手的位置坐标会不同，为了消除这种差异，需要对坐标进行归一化处理，得到左手、右手、左肘、右肘四点的相对坐标，建立一个 12 维向量作为 Feature Vector；由于每一帧的图像都会获得一个 12 维的 Feature Vector，数据量太大不利于之后要建立的 HMM，所以对 12 维特征进行量化。目前使用无监督方法 K-Means 进行聚类，将训练集所有的 Feature Vector 聚成 K 类（目前 K=100，尚未测试其他取值），给每个类一个标号从 0 到 99。建立 KDTree，对每一个 12 维特征向量用 KDTree 寻找分类，用 1 维的标号代替原来的 12 维向量，使数据量化并且降维，称 1 维特征为状态。

2.1.2 模型训练

手语视频经特征提取并量化得到状态序列（每一帧提取一个状态），以词为单位，用词的状态序列样本来训练 HMM，即每个词训练得到一个 HMM

2.1.3 孤立词识别

对孤立词的视频数据每帧都提取 12 维特征，然后量化得到状态序列，带入每个词的 HMM 中进行匹配，计算得到概率最大的 HMM 所对应的词即为识别结果。

2.1.4 连续句子识别

对句子的识别关键在于对句子按词进行分割，然后对词采用上面提到的孤立词识别方法来识别，最后用 N-Gram 语义模型对句子重组。

系统中我们对句子的分割采用一种简单的方法，认为词与词之间的过渡状态是静止的，根据前后帧坐标的变化可以计算手的移动速度，当速度小于阈值时判定为静止状态，即认为该时间点是分割点。分割得到词的状态序列用 HMM 来识别，每个词保留前 5 个候选词。建立 N-Gram 语义模型在所有候选词中找出一条最大的概率路径（Viterbi 算法），由此校正连续词识别结果。

2.2 手形分类系统简介

上面提到的系统的特征提取模块只提取了手的位置信息，考虑到手形变换是打手语时不可或缺的重要环节，其中包含了大量能反映语义的信息，所以我们打

算在获得手部位置特征的基础上加入手形特征。如果能对手形正确地分类，得到更丰富的特征，那么必定能提高 HMM 识别的效果。手形分类系统包括两部分，一是手形图像的获取，另一部分是建立手形分类模型。

2.2.1 手形图像的获取

首相根据 Kinect 获取手的彩色图像，深度图像和骨骼点信息；由骨骼点信息获取手心点的位置。再将手心点投影到彩色图像上，在手心点附近寻找皮肤颜色并闭合边缘，以此获取手形。然后将手形区域投影到深度图，根据深度图信息对手形区域进行校正。最后计算手形区域范围，令背景为黑色，手部取 RGB 值。

2.2.2 手形分类模型

深度神经网络有自主学习数据特征的能力，并且随着网络层数加深得到的特征越抽象，对原始数据的猜测越少，越有利于分类。本文主要研究基于 DBN 的手形分类，通过分析手语手形特点，将所有手形归结为 61 种基本手形。分类模型的目的就是把所有的手形正确地识别出对应的基本手形。

第 3 章 深度卷积网与深度置信网

3.1 卷积神经网络（Convolutional Neural Network）

在 2006 年深度置信网的训练算法提出以前，基于生物学研究成果，由 Fukushima、Lecun[28]等人提出并发展的卷积神经网络是为数不多的，可以在避免过拟合情况下，进行多隐层训练的神经网络，具有结构简单、训练参数少和适应性强等特点。特别是在图像识别领域，可以直接对二维图像（而无需一维化）进行建模，对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性。

3.1.1 提出与发展

1962 年 Hubel 和 Wiesel 通过对猫视觉皮层细胞的研究，提出了感受野(receptive field)的概念[31]，其核心思想为“人在视网膜上的神经元只感知其附近的视觉信号（所谓“感受野”），从而获得图像的局部特征，且同一位置会有多个神经元感受不同的局部特征。1984 年日本学者 Fukushima 基于感受野概念提出的神经认知机(neocognitron)可以看作是卷积神经网络的第一个实现网络[32]，也是感受野概念在人工神经网络领域的首次应用。

Lecun[29]在 1998 年提出了首个进行了实际应用的神经网络模型：Lenet，他在普通的卷积网络每个卷积层之间添加了一个子采样(Subsampling)层，并利用 MaxPooling 技术，在保证准确性的情况下，成功的减少了参数的数量，降低了模

型的自由度，同时大大提高了训练的速率。以这种网络为核心算法的 FPGA 芯片已在美国邮政编码的识别工作中取得了广泛的应用。本文采用的卷积神经网络，就是基于 Lenet 的结构建立的。

卷积神经网络由于其权值共享、参数自由度较小的特点，深层训练时，不易发生过拟合。特别是其卷积的特性使得其在处理二维图像上有着很强的优势。

3.1.2 网络结构

一个典型的处理二维图像数据的卷积神经网络，是一个多层的神经网络，每层由多个二维平面组成，而每个平面由多个独立神经元组成。网络中包含两种神经元，分别记为 C-元（Convolutional）和 S-元（Subsampling）。C-元聚合在一起组成 C-面，C-面聚合在一起组成 C-层，用 LayerC 表示。S-元、S-面和 S-层之间存在类似的关系。网络的中间层由 C-层与 S -层串接而成，而输入层为单层网络，直接接受二维视觉信号（raw data）。图 1 即是一个典型卷积网的示意图[37]。

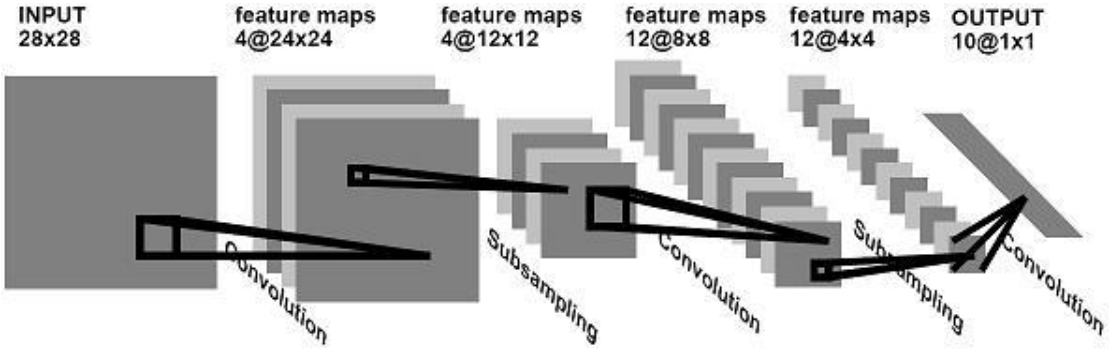

图 1 卷积神经网的典型结构

卷积神经网络的巧妙之处在于，其训练过程不仅是一个训练样本的过程，也是样本特征的学习过程。其特征提取的步骤已嵌入到卷积神经网络模型的互联结构中。

LayerC 为特征提取层，每个神经元的输入与前一层的局部感受野相连，并提取该局部的特征，一旦该局部特征被提取后，它与其他特征间的位置关系也随之确定下来，每个卷积核将输出一个二维矩阵，相当于是对图片进行了特征映射，从而得到相应的特征图，而多个卷积核则可以得到一系列不同的特征图（feature maps）；容易发现，在同一张 feature map 上的神经元共享相同的卷积核，也即是输入权值，形成了权值共享（weight share），这样就极大地减少了需要训练的参数复杂度。

LayerS 是子采样层，利用图像局部相关性的原理，对图像进行子抽样，可以

减少数据处理量同时保留有用信息。本层把卷积层输出的 feature maps 作为输入，分别进行子抽样运算后输出与输入图像数量相同的低维图像，而该图像保留了原图像的绝大部分有用信息。网络中神经元的输出连接值符合“最大值检出假说”，即在某一小区域内存在的一个神经元集合中，只有输出最大的神经元才强化输出连接值。所以若神经元近旁存在有输出比其更强的神经元时，其输出连接值将不被强化。根据上述假说，就限定了只有一个神经元会发生强化。因此，子抽样点的值是原图像对应位置相邻几个点的最大值。

最后一层通常为全连接层，即是普通的神经网络（非卷积），一般通过 logistics（二分类）或 softmax（多分类）激活函数来将其分为多层。

3.1.3 学习算法

卷积神经网络的训练算法分为有监督（Supervised）学习与无监督（Unsupervised）学习，其中无监督的学习主要用于图像的聚类，或者无标签数据的特征提取；本文主要使用的算法是利用向前算法得出结果，再用倒推（BP）算法进行相应的权值调整。

具体步骤描述如下：

卷积神经网络的训练样本集由（输入向量，理想输出）二元向量对构成的。所有这些对，都应来源于网络系统的真实值，它们可以是从实际运行系统中采集来的，也可以根据特定参数模拟，需要有着内在的结构化联系。

在开始训练前，所有的权都应该用选择小随机数进行初始化。“小随机数”保证了网络既不会因为权值过大而导致训练饱和，也不会因为权值相同而学习能力过差。实际上，如果用相同的数去初始化权矩阵，则网络无能力学习。我们也注意到了文献[36]提到的神经网络初值选取的重要意义，但是在实际的手语应用中，用适当增加迭代次数的方法可以有效的减小初值选择的影响。

训练算法主要包括 4 步，这 4 步被分为两个阶段：

第一阶段，向前传播阶段：

①从样本集中取一个样本(X ,Y)，将 X 输入网络；

②用式 3.1 计算相应的实际输出 O。

$$o = F_n(\cdots(F_2(F_1(X\ W^{(1)})W^{(2)})\cdots)W^{(n)}) \tag{3.1}$$

其中 Fk 表示第 k 层的激活函数（一般采用 sigmod 或 tanh），而 W(k)表示第 k 层的权值矩阵。

第二阶段，向后传播阶段

①计算实际输出 O 与相应的理想输出 Y 的差；

②按 BP 的方法调整权矩阵。

其中，BP 算法在文献[33][34][35]等中均有详细的介绍，在此不再赘述。

3.2 限制玻尔兹曼机（Restricted Boltzmann Machine）

限制玻尔兹曼机（RBM）是建立 DBN 的基本结构。DBN 可以看做是多个 RBM 堆叠而成的深层神经网络。因此在介绍 DBN 之前，有必要先介绍 RBM 的模型结构，学习算法。

3.2.1 模型

RBM 是一个两层的神经网络，一层是可视层，即数据输入层（V），一层是隐藏层（h），隐藏层是输入数据的一种特征表示，所以 RBM 可视为一个特征提取器。每一层的结点之间没有连接，层间是全连接的。如图 2

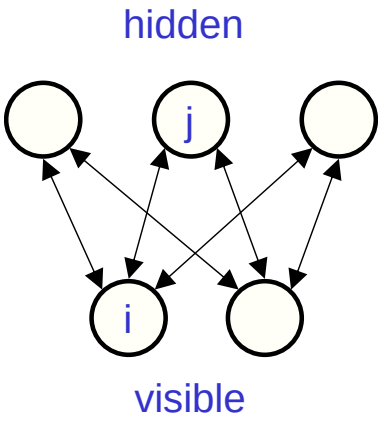

图 2 RBM 结构

为了简洁起见，我们假设所有的结点都是随机二值变量结点，即只能取 0 或者 1 值，同时假设全概率分布 $P(v,h)$ 满足 Boltzmann 分布。RBM 联合组态的能量可以表示为：

$$E(v,h)=-\sum_{i,j}v_i h_j w_{ij}-\sum_i b_i v_i-\sum_j a_i h_j \qquad (3.2)$$

其中 vi，hj 分别表示第 i 个可见单元与第 j 个隐单元的状态，wij 则为第 i 个可见单元与第 j 个隐单元的连接权重，ai，bj 分别为可视层和隐藏层的偏置单元。而某个组态的联合概率分布可以通过 Boltzmann 分布（和这个组态的能量）来确定：

$$P(v,h)=\frac{\exp(-E(v,h))}{Z} \qquad (3.3)$$

$$Z = \sum_{(v,h)} \exp(-E(\tilde{v}, \tilde{h})) \quad (3.3)$$

其中 Z 为归一化因子。要计算归一化因子 Z 需要考虑所有可能的组态，所以难以有效地计算由参数 w 确定的这个分布。由图可见，当给定可见单元的状态时，隐单元的激活概率是条件独立的。则第 j 个隐单元的激活概率为，

$$P(h_j = 1 \mid v) = sigmoid(v^T w_j) = \frac{1}{1 + \exp(-v^T w_j)} \quad (3.4)$$

由于 RBM 的对称结构，相应地当给定隐单元的状态时，可见单元的激活概率同样是条件独立的，

$$P(h_j = 1 \mid v) = sigmoid(w_j h) = \frac{1}{1 + \exp(-w_j h)} \quad (3.5)$$

3.2.2 RBM 学习算法

对于一个现实的问题，我们最关心的问题是由 RBM 所定义的关于观测数据 v 的分布 P(v)，即联合概率分布 P(v,h)的边缘分布：

$$P(v) = \frac{\prod_j (1 + \exp(v^T w_j))}{Z} \quad (3.6)$$

RBM 的参数便是通过最大化 RBM 在训练集上的对数似然度学习得到的。即：

$$W^* = \arg\max_w \sum_n \log P(v^n; w) \quad (3.7)$$

为了获得最优参数 W,我们可以使用梯度上升发。对于训练数据 v(n)，其对数似然度的梯度为，

$$\frac{\partial \log P(v^{(n)})}{\partial W} = -\left\langle \frac{\partial E(v^{(n)}, h)}{\partial W} \right\rangle_{P(h|v^{(n)})} + \left\langle \frac{\partial E(v, h)}{\partial W} \right\rangle_{P(v,h)} \quad (3.8)$$

其中<.>P 表示求关于分布 P 的期望。3.8 式的被减数表示输入样本数据的自由能量函数期望值，而减数表示模型产生样本数据的自由能量函数期望值。可惜的是，如前文所述，由于归一化因子的存在，我们无法有效地计算训练数据的似然度。因此我们需要使用采样方法来获得对数似然度梯度的近似。下面介绍 RBM 中的应用吉布斯采样（Gibbs sampling）方法。

3.2.3 RBM 中的吉布斯采样

吉布斯采样（Gibbs sampling）[15]是一种基于马尔可夫链蒙特卡罗（Markov chain Monte Carlo，MCMC）[16]的采样方法。考虑一个 K 维的随机向量 X，其分量为 $X_1, X2, \ldots, X_K$。我们无法求得关于 X 的联合分布 P(X)，但是我们知道给定 X

其他分量时，X_k 的条件分布，即 $P(X_K \mid X_{K-})$，$X_{K-}=(X_1, X2,\ldots, X_{K-1}, X_{K+1},\ldots, X_K)$。那么我们可以从 X 的一个任务分布状态开始（如$[x_1(0), x_2(0),\ldots.., x_k(0)]$），利用上述的条件分布，迭代地对其他分量依次进行采样，随着采样次数的增加，随机变量$[X_1(n), X_2(n),\ldots.., X_k(n)]$的概率分布将以 n 的几何级数的速度收敛于 X 的联合概率分布 P(X)。也就是通过上述的步骤，我们可以在未知联合概率分布 P(X)的条件下对其进行采样。

RBM 的对称结构，以及其中神经元状态的条件独立性，使得对于 RBM 使用吉布斯采样方法我们可以有效地得到服从 RBM 定义的分布的随机样本。在 RBM 中进行 k 步吉布斯采样的具体算法为，随机初始化可见单元 v0 的状态，交替进行如下采样，

$$\begin{aligned} h_0 &\sim P(h \mid v_o) \\ v_1 &\sim P(v \mid h_0) \\ h_1 &\sim P(h \mid v_1) \\ v_2 &\sim P(v \mid h_1) \end{aligned} \qquad (3.9)$$

采样步数 k 足够大的情况下，我们便可以得到服从 RBM 定义的概率分布的样本，这可以方便地考察 RBM 作为一个生成模型到底学到怎样的分布。此外，使用吉布斯采样我们便可以得到等式 3.8 中第三项的一个近似。

3.2.4 对比散度

尽管利用吉布斯采样我们可以得到训练数据对数似然度梯度的近似，但通常需要使用较大的采样步数，这使得 RBM 的训练效率仍旧不高，尤其当观测数据在高维空间时。

2002 年 Hinton 提出对比散度（Contrastive Divergence，CD）[17]来近似式 3.8 的第三项。与吉布斯采样不同，Hinton 指出当使用训练数据初始化 v0 时，我们仅需要使用 k 步的吉布斯采样我们便可以得到足够好的近似。由此，当使用随机梯度上升优化式 2.14 时，给定训练数据 v(n)，隐单元 j 的特征 wj 的迭代更新式为，

$$\triangle w_j = P(h_j = 1 \mid v^{(n)}) \cdot v^{(n)} - P(h_j = 1 \mid v^{(n)-}) \cdot v^{(n)-} \qquad (3.10)$$

$$v^{(n)-} \sim P(v \mid \hat{h}) \qquad (3.11)$$

$$\hat{h} \sim P(h \mid v^{(n)}) \qquad (3.12)$$

过程如图 3。

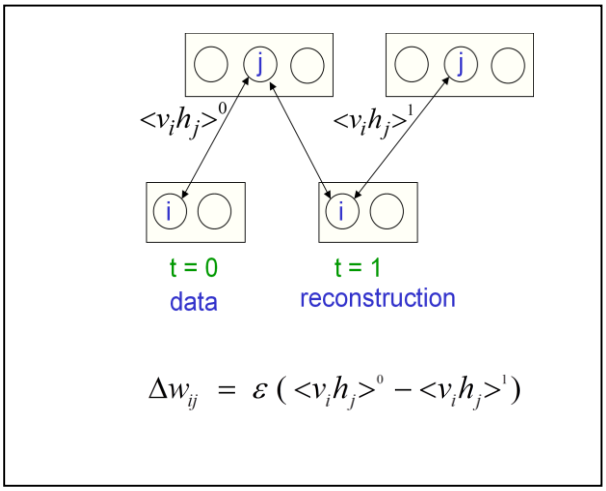

图 3 对比散度

CD 算法实际上是针对训练那些归一化因子不可求的模型的一种一般性的算法，文献[19]通过大量实验指出 CD 算法并不会收敛到最大似然的解，确切地说，CD 算法会的到最大似然解的一个有偏估计。但是，当目标为求最大似然解时，文献[19]指出可以使用 CD 算法进行初始地训练，在后面的训练中逐步地增加 CD 算法中吉布斯采样的步长，通常可以得到最大似然解的一个很好地近似。

3.3 深度置信网（Deep Belief Network）

3.3.1 结构

DBN 的结构如图 4 所示，最上面的两层是一个 RBM，剩下的各层构成一个有向图。

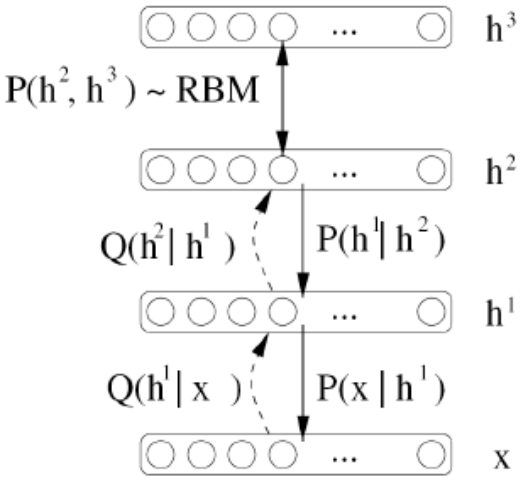

图 4 DBN 结构

假设 DBN 是一个 l 层的神经网络，它定义了一个关于观测数据 x 和 l 层隐单元 $(h^1,.....,h^l)$ 的联合分布，

$$P(x,h^1,.....,h^l)=(\prod_{k=0}^{l-2}P(h^k\mid h^{k+1}))P(h^{l-1},h^l) \qquad (3.13)$$

其中 $x=h^0$， $P(h^k\mid h^{k+1})$ 是一个潜在的 RBM 条件概率分布（可以把 hk+1 看作这个 RBM 的隐单元，将 hk 看作相应的可见单元），$P(h^{l-1},h^l)$ 则是一个 RBM 定义的联合概率分布。训练图 4 的模型的困难之处在于无法有效地计算模型后验概率 $P(h^{k+1}\mid h^k)$。文献[15]指出，当使用 l 个 RBM 初始化一个 l 层的 DBN 后，可以使用 RBM 定义的条件分布（如图 4 中所示的 $Q(h^2\mid h^1)$）来近似这个 DBN 的后验概率。

3.3.2 DBN 训练算法

无监督贪婪逐层算法[15]是由 Hinton 提出的,算法的基本思想是把一个 DBN 看作是多个 RBM 的堆叠，通过训练 RBM 得到的 DBN 的权值，该过程也称为预训练，如图 5。预训练使网络得到一个较好的初始权值，然后我们可以用有标记的数据对网络进行微调，这样网络对数据的拟合效果更好。

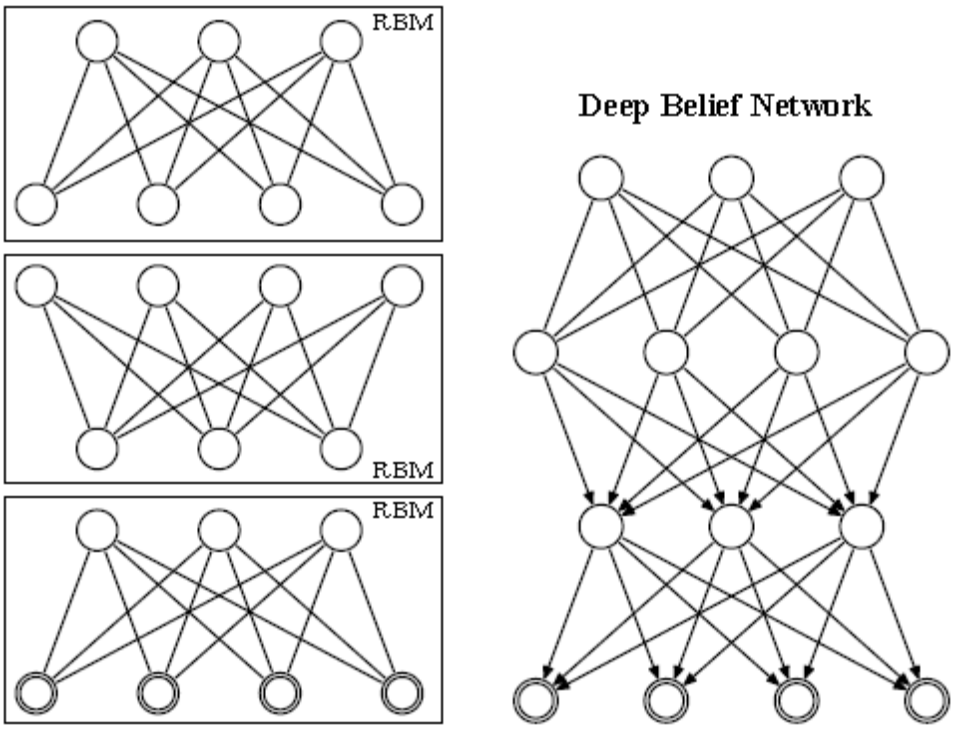

图 5 预训练过程

预训练: 首先，使用训练数据集训练第一个 RBM，并使用这个 RBM 的连接权重来初始化 DBN 的最底层参数。然后计算输入数据在第二层的输出 h1，形成新的训练集，训练第二个 RBM。如此迭代，直到最后一层。

微调：在 DBN 的顶层添加一个分类层，数据的标签将被附加到分类层的顶层，一般采用 BP 算法反向传播去调整整个网络的权值。如图 6

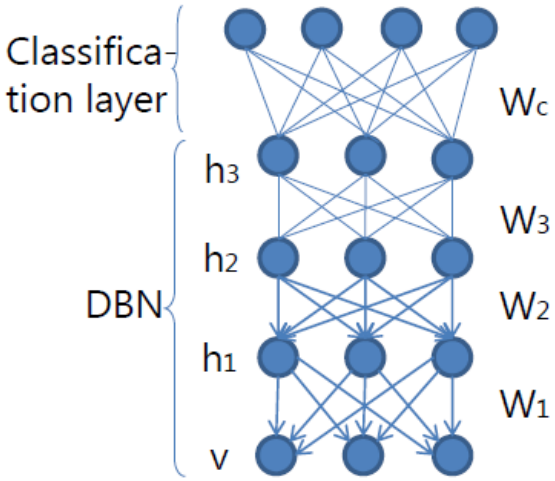

图 6 DBN 结构-2

训练过程可以直观地解释，预训练类似神经网络的随机初始化初值过程，但是不同于随机初始化，这是通过学习输入数据的结构得到的，因而这个初值更接近全局最优，从而能够取得更好的效果。之后只需要用 BP 算法对权值参数空间进行一个局部搜索。

第四章 基于 CNN 和 DBN 的手形分类

句子可以由词组合而成，大部分手语词中涉及的手形都可以归结为 61 种基本手形的组合。如果能对基本手形作出准确的分类，那么基于手形信息的手语分类的效果将大大提升。

本文通过建立 DBN 来进行分类，以手形轮廓图像作为模型输入，输出的 61 个单元代表 61 种基本手形。深度神经网络的样本训练是关键，因为目标函数是非凸的，所以很容易收敛到局部最优，使得分类效果不好。Hinton 提出的非监督贪婪逐层训练算法[11]是一种快速有效的深度神经网络训练方法，算法的基本思想是把深层网络看做是多个 RBM（限制波尔兹曼机）的堆叠。

4.1 基本手形

在手语中，手语词是最小的，有意义的单位。手形的位置，动作和方向共同构成了手语词。因此一个手语词可以包含一个或者多个的手形。手形作为手语表达过程中最重要的信息，如果能够很好地提取手形特征，对手形进行分类，那么将大大提高手语语义分类的准确率。根据[14]的研究，手语的手形并非杂乱无章，根据手指数量，手指形态和手指的组合，最终可以归结出 61 个基本的手形。表 1 列出了 61 个基本手形的名称。图 7 是部分基本手形的展示。因此本论文的主要目的是对手语表达过程中的手形进行分类，分类出对应 61 个基本手形中的哪一个。

序号	手形	序号	手形	序号	手形	序号	手形
1	1-伸	17	1+5-捏	33	2+3+4-伸	49	五-S
2	1-弯	18	1+5-弯	34	2+3+5-伸	50	五-捏
3	2-伸	19	2+3-伸	35	3+4+5-伸	51	五-O
4	2-弯	20	2+3-并	36	3+4+5-弯	52	五-D
5	3-伸	21	2+3-弯	37	1+2+3+4-M	53	五-CH
6	5-伸	22	2+3-交	38	2+3+4+5-伸	54	五-OK
7	5-弯	23	2+5-伸	39	2+3+4+5-并	55	五-P
8	1+2-伸	24	1+2+3-伸	40	2+3+4+5 弯	56	五-WC
9	1+2-并	25	1+2+3-捏	41	五-伸	57	五-兰花指
10	1+2-弯	26	1+2+3-弯	42	五-并	58	五-床
11	1+2-平	27	1+2+3-警察	43	五-聚	59	五-姜
12	1+2-环	28	1+2+3-N	44	五-开	60	五-仿 d
13	1+2-捏	29	1+2+3-K	45	五-侧开	61	五-毛笔
14	1+2-十字	30	1+2+3-SH	46	五-弯		
15	1+2-半	31	1+2+3-除号	47	五-平		
16	1+5-伸	32	1+2+5-伸	48	五-C		

表 1 61 种基本手形

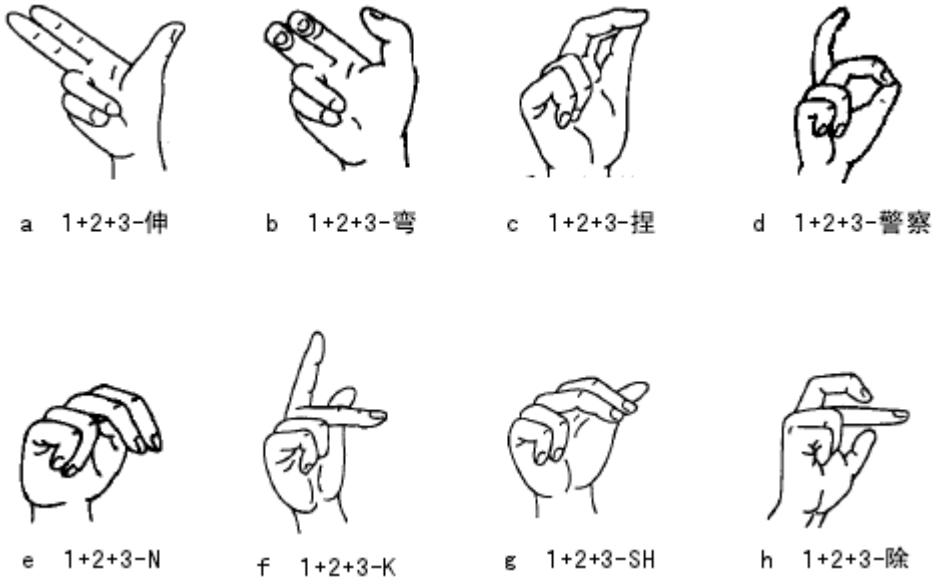

图 7 部分基本手形

4.2 应用深度神经网络实现手形分类

分类效果依赖于特征，好的特征应具有不变（形大小、尺度和旋转等）和可区分性在图像领域出现了不少好的特征，例如 Sift 和 hog。但它们也不是万能的。然而，手工地选取特征是一件非常费力、启发式（需要专业知识）的方法，而且它的调试需要大量的时间。相比之下，深度神经网络可以自动地学习图像的特征，很好地解决了这一问题。

正如 3.1 节中提到的那样，文献[31]发现了视觉系统的信息处理”：可视皮层是分级的。这个发现激发了人们对于神经系统的进一步思考。神经-中枢-大脑的工作过程，或许是一个不断迭代、不断抽象的过程。从原始信号，做低级抽象，逐渐向高级抽象迭代。人类的逻辑思维，经常使用高度抽象的概念。例如，从原始信号摄入开始（瞳孔摄入像素 Pixels），接着做初步处理（大脑皮层某些细胞发现边缘和方向），然后抽象（大脑判定，眼前的物体的形状，是圆形的），然后进一步抽象（大脑进一步判定该物体是只气球）。

总的来说，人的视觉系统的信息处理是分级的。从低级的 V1 区提取边缘特征，再到 V2 区的形状或者目标的部分等，再到更高层，整个目标、目标的行为等。也就是说高层的特征是低层特征的组合，从低层到高层的特征表示越来越抽象，越来越能表现语义或者意图。而抽象层面越高，存在的可能猜测就越少，就越利于分类。例如，单词集合和句子的对应是多对一的，句子和语义的对应又是多对一的，语义和意图的对应还是多对一的，这是个层级体系。

深度神经网络就是借鉴这个过程。直接把手形图像作为网络的输入，让每一

个隐层都去提取手形的抽象特征。低层的单元提取图像的初级信息，高层的单元对低层的特征进行组合得到更加抽象的特征，随着层数的不断加深，网络对于手形的“理解”趋于正确，并且对于不同的手形图像，网络有不同的“理解”，根据不同的“理解”来实现分类。

第五章 实验与结果

5.1 数据获取与处理

实验数据采集自 9 个人录制的共 549 段视频转成的图片，包含 61 种基本手形。视频为经过 Kinect 基于深度信息抠取的双手手势，尺度为 128*128，每秒 25 帧，每只手共有约 10 万张图片。原始图片如图 8。

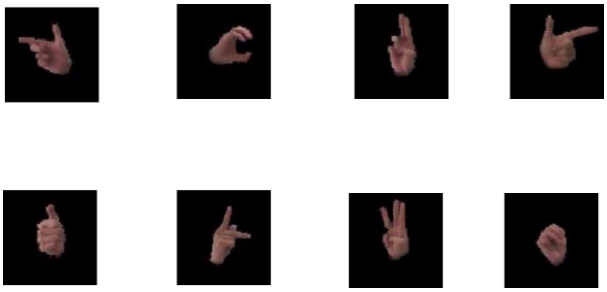

图 8 原始图片

图像的格式选择 32*32 的灰度图，训练集从中随机选取 30 万个样本，剩余 20 万个样本作为测试集。考虑到即使是同一个人不同时间做同样的手形也会存在角度的差异，为增加分类效果的鲁棒性，研究者对数据进行了扩充，将录制的数据进行了 5 次正负 60 度之间随机旋转，最终得到约 50 万张图片，如图-9。其中我们随机抽取了 60%作为训练集，剩余的 40%作为测试集。

图 9 旋转过的基本手形

本文使用了 Matlab 2012b 和 C# 进行编程实践，其中引用了 DeeplearningToolbox 和 Accord 框架下 DeepLearning 模块中的部分函数。

5.2 CNN 手形分类

建立包含四个隐层的 CNN 网络模型，分别为：输入层–卷积层–子采样层–卷积层–子采样层–输出层，神经元个数分别为 1024-4704-2352-768-384-61。

将 32*32 图片直接作为输入放入神经元，理想输出向量为一个 61 维{0,1}向量，在对应的手形编号位置上取 1，其余部分取 0。神经网络的输出选择其预测向量 O 中，最大值所在的位置（argmax（O_i））。

对未经旋转的图片进行测试，迭代 200 次，识别率高达 99.3%，

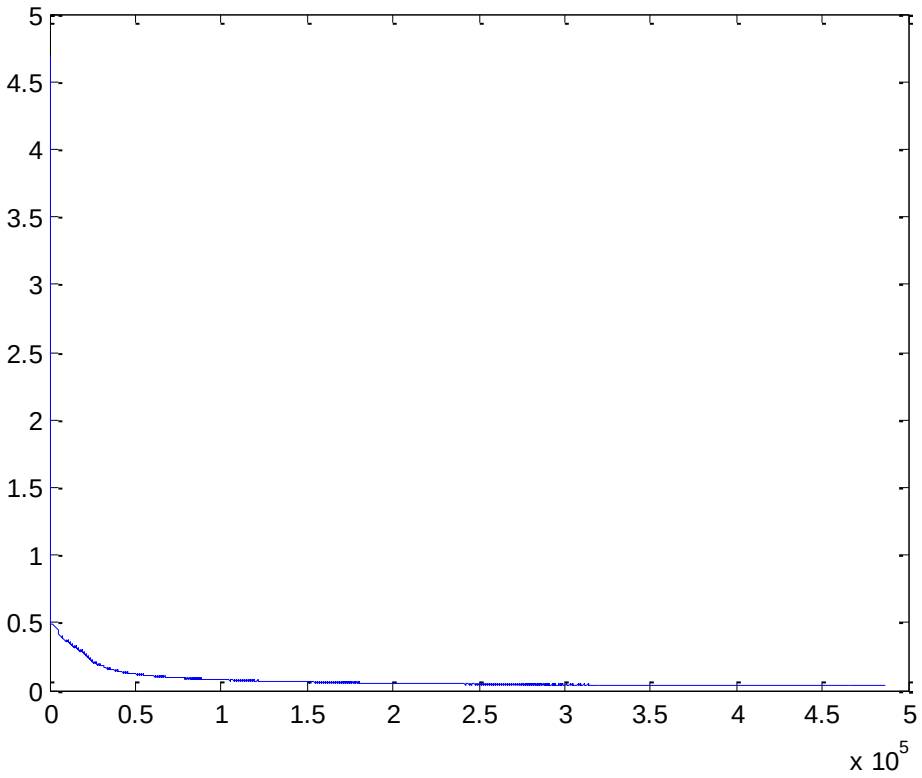

图 10 未旋转图片的收敛情况

选择迭代次数为 400 次，训练过程的收敛情况如下：

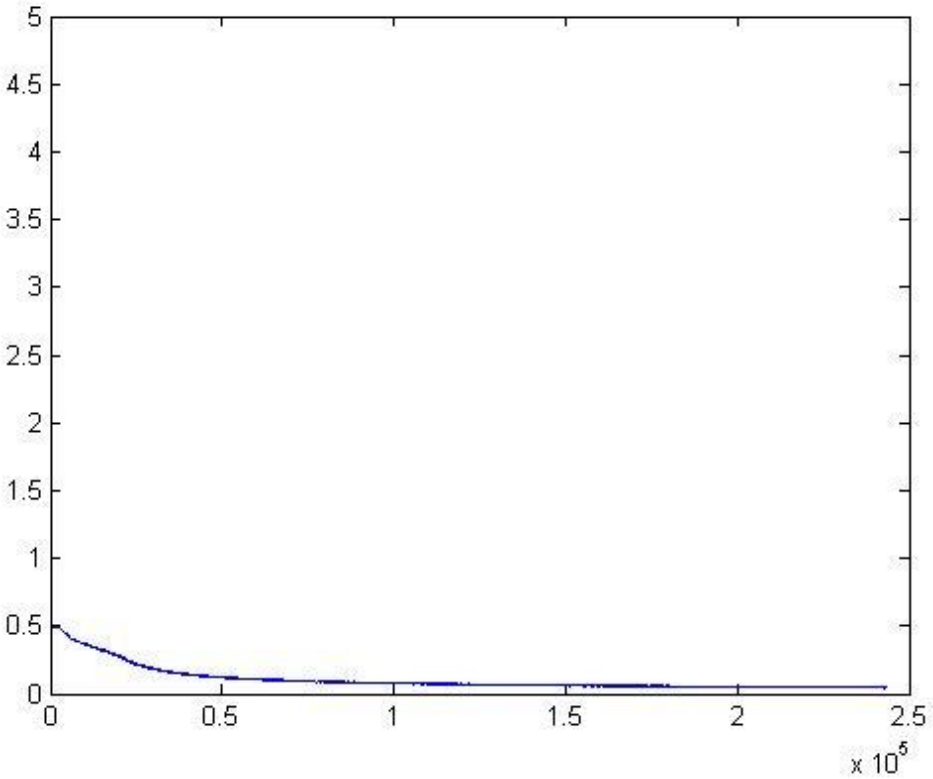

图 11 旋转图片的收敛情况

最终的识别率高达 98.9%，远远超过了文献[38][39]，也超过了 state-of-art[40]的结果。

5.3 DBN 手形分类

搭建一个包含三个隐层的 DBN 网络结构，从输入到输出的神经元个数依次为：1024-100-100-2000-61。因为系统每次处理一帧图像，所以 32*32 的灰度值矩阵被拉成一个 1024 维的数组作为 DBN 的输入，输出的 61 个单元对应 61 个基本手形，如果输出表示属于第 i 个类，则第 i 个单元的值为 1，其他单元的值为 0。

预训练时，每个 RBM 迭代 50 次，模型收敛情况如图 12、13、14。

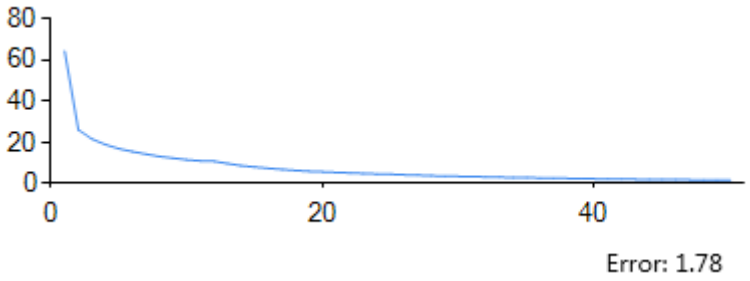

图 12 第一层 RBM 收敛情况

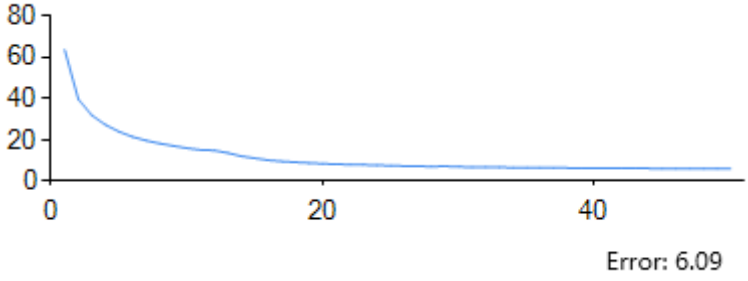

图 13 第二层 RBM 收敛情况

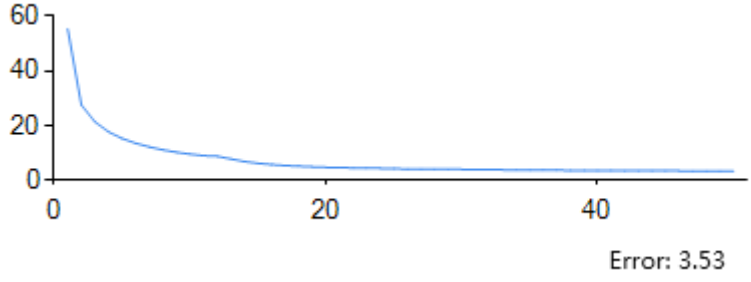

图 14 第三层 RBM 收敛情况

微调时采用 BP 算法，迭代 100 次。错误率曲线如图 15

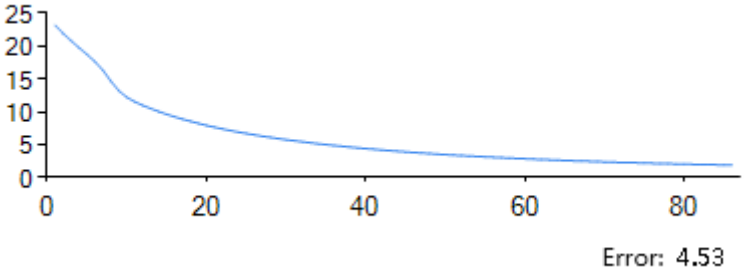

图 15 微调（fine-tune）误差收敛曲线

经过微调之后的模型，识别率达到 95.5%，虽然与 CNN 的结果有所差异，但同样实现了很好的分类效果。

5.5 结果分析

通过对 CNN 训练出的权值进行分析（图 16-19），我们可以看出其提取出的特征与我们通常提取的特征有相似之处，但又进行了一些针对手语的优化。 同时也可以发现，经过卷积处理的图片确实是逐层抽象的

图 16 第 1 卷积层手部特征卷积核

图 17 第 2 卷积层手部特征卷积核

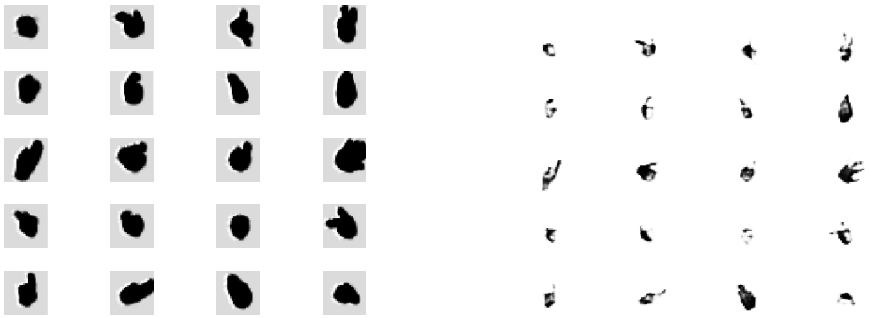

图 18 第一卷积层提取的部分手部特征图

图 19 第 2 卷积层提取的部分手部特征图

第 6 章 结论和展望

人的视觉系统的信息处理是分级的，深度神经网络模拟人的视觉系统，对输入数据提取层次特征，从低层到高层的特征表示越来越抽象，越来越能表现语义或者意图。而抽象层面越高，存在的可能猜测就越少，就越利于分类。实验证明 CNN/DBN 对手形图像具有很好的分类效果。

手形分类作为整个手语系统的一个子模块，目的是为手语语义的分类提供手形信息。可以从两个方面考虑建立基于手形分类的手语分类。一是建立一个字典，字典记录每个手语词和基本手形组合序列之间的映射关系，因为大部分的手语词都可以由基本手形组合而成，所以只要识别出基本手形的组合序列，就可以在字典中查找对应的语义。另一方面可以考虑建立更有理论依据的概率生成模型，比如 HMM 和 HCRF，此时基本手形构成的序列可以作为模型的观测序列。

实验中还存在一些可改进的地方，本文介绍的手形图片是 32*32 的灰度图，如果使用 128*128 的 RGB 彩色图，那么可以对更多，更加复杂的手形进行分类。但是因为输入的数据信息更多，训练所需的时间也会增加。也可以考虑增加层数以及单元数来提升特征提取的效果。还可以使用 DBN 和 CNN 的组合，增加模型的复杂度使得模型能更好地模拟人的视觉系统。

参考文献

[1] Liang R H, Ouhyoung M. A real-time continuous gesture recognition system for sign language[C]//Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998: 558-567.

[2] Fang G, Gao W, Chen X, et al. Signer-independent continuous sign language recognition based on SRN/HMM[M]//*Gesture and sign language in human-computer interaction. Springer Berlin Heidelberg*, 2002: 76-85.

[3] 马继勇. 手语理解的统计模型研究[D]. 中国科学院研究生院（计算技术研究所） 2001

[4] Lee C, Xu Y. Online, interactive learning of gestures for human/robot interfaces[C]//*Robotics and Automation*, 1996. Proceedings., 1996 IEEE International Conference on. IEEE, 1996, 4: 2982-2987.

[5] Takahashi T, Kishino F. A hand gesture recognition method and its application[J]. *Systems and Computers in Japan*, 1992, 23(3): 38-48.

[6] Takahashi T, Kishino F. Hand gesture coding based on experiments using a hand gesture interface device[J]. *ACM SIGCHI Bulletin*, 1991, 23(2): 67-74.

[7] Takahashi T, Shima N, Kishino F. An image retrieval method using inquiries on spatial relationships[J]. *Journal of information processing*, 1992, 15(3): 441-449.

[8] Vogler C, Metaxas D. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods[C]//*Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation*., 1997 IEEE International Conference on. IEEE, 1997, 1: 156-161.

[9] Vogler C, Metaxas D. ASL recognition based on a coupling between HMMs and 3D motion analysis[C]//Computer Vision, 1998. Sixth International Conference on. IEEE, 1998: 363-369.

[10] Starner T E. Visual Recognition of American Sign Language Using Hidden Markov Models[R]. MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF BRAIN AND COGNITIVE SCIENCES, 1995.

[11] Pitsikalis V, Theodorakis S, Vogler C, et al. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2011

IEEE Computer Society Conference on. IEEE, 2011: 1-6.

[12] Dreuw P, Deselaers T, Rybach D, et al. Tracking using dynamic programming for appearance-based sign language recognition[C]//Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on. IEEE, 2006: 293-298.

[13] Bui T D, Nguyen L T. Recognizing postures in Vietnamese sign language with MEMS accelerometers[J]. Sensors Journal, IEEE, 2007, 7(5): 707-712.

[14] 骆维维. 《中国手语》手形研究[D]. 北京师范大学 2008

[15] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. *Neural computation*, 2006, 18(7): 1527-1554.

[16] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images[J]. *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, 1984 (6): 721-741.

[17] Liu J S. Monte Carlo strategies in scientific computing[M]. Springer Verlag, 2008.

[18] Hinton G E. Training products of experts by minimizing contrastive divergence[J]. *Neural computation*, 2002, 14(8): 1771-1800.

[19] Carreira-Perpinan M A, Hinton G E. On contrastive divergence learning[C]//*Artificial Intelligence and Statistics*. 2005, 2005: 17.

[20] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, 313(5786): 504-507.

[21] Nair V, Hinton G. 3-d object recognition with deep belief nets[J]. *Advances in Neural Information Processing Systems*, 2009, 22: 1339-1347.

[22] Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]//*NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*. 2009.

[23] Garofolo J S. TIMIT: Acoustic-phonetic Continuous Speech Corpus[M]. *Linguistic Data Consortium*, 1993.

[24] Salakhutdinov R, Hinton G. Using deep belief nets to learn covariance kernels for gaussian processes[J]. *Advances in neural information processing systems*, 2008, 20: 1249-1256.

[25] Salakhutdinov R, Hinton G. Learning a nonlinear embedding by preserving

class neighbourhood structure[C]//*AI and Statistics*. 2007, 3: 5.

[26] Höffken M, Oberhoff D, Kolesnik M. Switching hidden Markov models for learning of motion patterns in videos[M]//*Artificial Neural Networks–ICANN 2009. Springer Berlin Heidelberg*, 2009: 757-766.

[27] Minsky M L, Papert S A. Perceptrons - Expanded Edition: An Introduction to Computational Geometry[M]. Boston, MA:: MIT press, 1987.

[28] P. Simard, Y. LeCun and J. Denker: Memory Based Character Recognition Using a Transformation Invariant Metric, in IAPR (Eds), *Proc. of the International Conference on Pattern Recognition*, II:262-267, IEEE, Jerusalem, October 1994,

[29] R.E. Howard, B. Boser, J.S. Denker, H.P. Graf, D. Henderson, W. Hubbard, L.D. Jackel, Y. Le Cun and H. S. Baird: Optical character recognition: a technology driver for neural networks, *IEEE International Symposium on Circuits and Systems*, 1990, 3:2433-2436, 1990,

[30] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. *Neural computation*, 2006, 18(7): 1527-1554.

[31] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. *The Journal of physiology*, 1962, 160(1): 106.

[32] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. *Biological cybernetics*, 1980, 36(4): 193-202.

[33] Fahlman S E. An empirical study of learning speed in back-propagation networks[J]. 1988.

[34] Van Ooyen A, Nienhuis B. Improving the convergence of the back-propagation algorithm[J]. *Neural Networks*, 1992, 5(3): 465-471.

[35] Hirose Y, Yamashita K, Hijiya S. Back-propagation algorithm which varies the number of hidden units[J]. *Neural Networks*, 1991, 4(1): 61-66.

[36] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//*Proceedings of the International Conference on Artificial Intelligence and Statistics* (AISTATS'10). Society for Artificial Intelligence and Statistics. 2010.

[37] LeCun Y, Jackel L D, Bottou L, et al. Comparison of learning algorithms for

handwritten digit recognition[C]//*International conference on artificial neural networks*. 1995, 60.

[38] Grobel K, Hienz H. Video-based handshape recognition using a handshape structure model in real time[C]//*Pattern Recognition, 1996., Proceedings of the 13th International Conference on. IEEE*, 1996, 3: 446-450.

[39] Vogler C, Metaxas D. Handshapes and movements: Multiple-channel american sign language recognition[M]//*Gesture-Based Communication in Human-Computer Interaction*. Springer Berlin Heidelberg, 2004: 247-258.

[40] Potamias M, Athitsos V. Nearest neighbor search methods for handshape recognition[C]//*Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments. ACM*, 2008: 30.