清 华 大 学

综 合 论 文 训 练

题目：深度学习与多模态信息融合策略探究

系　　别：计算机科学与技术系

专　　业：计算机科学与技术

姓　　名：高博

指导教师：李洪波助理研究员

2016 年 6 月 15 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签　名：_____ 导师签名：_____ 日　期：_____

中文摘要

物体的识别、标记，以及机器人的各类感知、操作的实现都基于对信息的处理、理解和利用。一方面，在实际操作中，单一模态的信息往往不能提供我们需要的精确度和全面性。这就为我们使用多模态信息，探究多模态信息的融合策略提供了动机。而另一方面，传感器技术的飞速发展，又为收集更高质量、更大规模的不同模态的信息提供了技术支持与可能性。所以本文就在此背景下进行多模态信息融合策略的研究，以便为机器人精细感知和精细操纵的实现提供技术基础。

本文针对多方法多模态信息融合策略以及深度学习在多模态信息融合中的应用做出了一系列研究，具体的工作有：利用传统的方法来进行颜色和深度信息的提取和融合；提出一种决策层的带权投票信息融合策略；引入了深度学习方法；提出了一种将深度信息转换为合法图片的正规化方法；使用预训练模型提取多模态信息；提出了结合了分层匹配追踪、随机权值递归神经网络以及深度学习三种方法的综合多模态信息识别模型，提升了识别的准确率和稳定性。

关键词：信息融合；深度学习；识别；神经网络

ABSTRACT

The processing, understanding and utilization of information serve as not only the basis of recognition and segmentation tasks, but also the key part in robotic fine perception and manipulation. On one hand, one single modal of information can not always provide the precision required by the practical case, therefore makes multi-modal information fusion a meaningful topic. On the other hand, the rapid development of sensor technology has made multi-modal information fusion viable by providing multi-modal datasets of larger scale and higher quality. In this article we present a multi-modal information fusion strategy combined with deep-learning methods, in order to provide technical backing for robotic perception and manipulation.

We systematically propose a series of approches that can effectively fuse multi-modal information and multi-method features on different levels. Specifically, our research includes the following parts: using hierarchical matching pursuit and randomized CRNN to extract features from both RGB and depth cues; proposing a decision level fusion method based on SVM and weighed voting; taking adventage of deep-learning approach; proposing a depth normalization method that can apply different normalization scales to different parts of depth cues; using pretrained model to extract features; proposing a generalized fusion model based on hierarchical matching pursuit, randomized CRNN and deep-learning, which achieves high classification accuracy and stability.

Keywords: Information fusion; Deep learning; Recognition; Neural network

目 录

主要符号对照表

HMP	分层匹配追踪 (Hierarchical Matching Pursuit)
NN	神经网络 (Neural Network)
CNN	卷积神经网络 (Convolutional Neural Network)
RNN	递归神经网络 (Recursive Neural Network)
BP	反向传播算法 (Back propagation)
FP	正向传播算法 (Forward propagation)
RGB	颜色
RGB-D	颜色和深度
Finetune	微调
SVM	支持向量机 (Support Vector Machine)
Linear SVM	核函数支持向量机
Kernel SVM	核函数支持向量机
Normalization	正规化
STD	标准差 (Standard Deviation)
OMP	正交匹配追踪 (Orthogonal Matching Pursuit)
SVD	奇异值分解 (Single Value Decomposition)
K-SVD	K-奇异值分解
AlexNet	Alex 神经网络
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IDL	实例距离学习 (Instance Distance Learning)
SP	空间金字塔 (Spatial Pyramid)
BoW	词袋模型 (Bag of Words)
DAN	深度信息自适应正规化方法 (Depth Adaptive Normalization)

第 1 章　引言

本章内容安排如下：1.1 小节介绍本文研究背景，要解决的问题及研究的目的； 1.2 小节整理并回顾相关的国内外研究成果，并讨论其优缺点； 1.3 小节介绍本文研究讨论的主要内容； 1.4 小节介绍本文的主要贡献点；最后， 1.5 小节介绍本文结构组织。

1.1　研究背景

多模态信息的融合有助于更好地利用不同模态信息之间的互补性，挖掘信息的潜力，提高机器人识别、感知和操作的精确度和鲁棒性，提高机器人系统的稳定性。近年来机器人的识别感知技术在诸如空间探索 [1–3]、机械制造 [4]、灾难应对 [5,6]、军事安保 [7] 以及医疗卫生 [8,9] 等领域获得了广泛应用。而随着各式各样的机器人应用日趋普及，以及对于机器人感知精度要求的不断提高，信息融合策略 [10] 也变得日益重要。同时，由于传感器技术不断进步，可以收集的信息模态越来越多 [11]，也为信息融合策略的研究提供了背景和技术支持 [12,13]。现在我们已经可以较好地提取包括颜色信息、深度信息、触觉信息、滑觉信息、声音信息、电学信息等在内的多种模态的信息 [14]。因为不同模态的信息（例如 RGB 信息和深度信息）含有的目标物体的特征是不同方面的 [15]，例如，通过对颜色信息的处理我们可以得到目标物体的颜色特征，图案纹理特征和光照特征等；而通过对深度信息的处理我们可以获得目标物体在空间中的形状特征和更加清晰的边缘特征等信息。所以不同模态的信息之间天然地具有较强的互补性。不过我们在研究过程中注意到，由于每种信息提取方式在提取信息时都会选择性地保留部分原有信息 [16]，所以不同的信息提取方式之间应该也具有一定的互补性 [17]。

此外，深度学习 [18] 的飞速发展使得大量识别任务的准确率迅速提升。例如，[19] 使用深度学习来进行音频分类；[20,21] 使用深度学习来进行人脸识别；[22,23] 使用深度学习来提取 RGB-D 的信息。所以我们也打算使用深度学习的方法来进行多模态信息融合。不过由于深度学习对于数据量的要求是巨大的 [24]，而目前已有的多模态数据集大小往往达不到深度学习的要求，所以我们使用已

经预先训练的深度学习模型来进行信息提取 [25]。同时，由于我们选用的模型是使用 RGB 图片信息训练的，所以在使用它提取深度信息之前我们还使用了深度信息对其进行微调 [26]，以便使原模型可以更好地适应深度信息。

综上所述，本研究要解决的问题是：

1. 寻找适用于不同模态的信息和不同信息提取方法的不同层次的信息融合策略，有效利用模态间和方法间数据互补性。
2. 深度学习在多模态信息融合中的应用。
3. 深度学习中多模态信息数据量不足的问题。

而本研究的主要目的是进行深度学习与多模态信息融合策略的研究，以便为实现机器人的精细感知和精细操纵提供技术支持。

1.2 文献综述

最早的有关信息融合的应用于第二次世界大战中出现在军事领域， [27] 将高射炮火控系统中的雷达传感器和光学传感器融合起来应用，达到了既提升了雷达系统的测距精度，又增强了系统的鲁棒性的效果，使其抗恶劣天气干扰能力大幅增强。而数据融合（Data Fusion）概念，也是由美国国防部下属研究机构于 1973 年正式提出 [28]，并应用于水下声纳系统，并随后在各类军用系统中得到广泛使用。进入 20 世纪 80 年代，军事上对于信息精度的要求迅速提升 [29]，使得 Data Fusion 得到了各国研究人员的广泛关注。

目前有关信息融合的研究，按信息的来源可以分为：

1. 多传感器信息融合。
2. 多模态信息融合。

按融合层次可以分为 [30]：

1. 数据层融合。
2. 特征层融合。
3. 决策层融合。

多传感器信息融合的意义在于，每种传感器都有自己的精度、误差范围以及作用范围 [31]。而在实际操作中通过将多个同一种类的传感器组合使用，往往可以提高测量精度，提升系统稳定性以及增大测量范围 [32]。 [33] 通过使用多个传感器，在多个定位结果中求解最小带权均方误差的方法给出最终定位结果，

提升了定位系统的稳定性，并将此方法应用到了机器人操作中。综合使用多个传感器取均值或投票来产生最终结果的方法在机器人感知识别和精细操纵，以及各种智能系统的构建中比较常见，例如 [34,35]，而 [36] 则对多传感器融合集成做了综述。 [37] 是对颜色信息中的多个通道进行信息融合，求其非局部平均值，以降低噪声对感知结果的影响，提高数据信噪比和感知质量。根据 [38] 的整理和综述，多传感器融合的主要方法有：加权平均法，例如 [39]；多数投票法，例如 [40]； [41] 和 [42] 使用卡尔曼滤波的方法来融合信息； [43] 使用了贴近度算法； [44] 介绍了基于粗糙集和模糊理论的信息融合策略； [45] 中还介绍了可用于信息融合的随机集方法；当然还有经典的贝叶斯融合方法 [46–48]；以及神经网络方法 [49,50] 等。具体地说，[40] 使用了 SVM（支持向量机）对多个传感器的分类结果进行了多数投票产生最终结果，有效地改进了分类的准确率。不过由于通常情况下的 SVM 只能给出分类结果，而不能给出分类的自信度（即分类结果属于每一类的概率），所以使用 SVM 意味着只能使用多数投票法。不过这就对参与投票的传感器的数量有了要求，或者需要其他的规定来避免平局的出现。而在本文中我们使用了带权的 SVM，可以进行带权投票。 [47] 使用了贝叶斯网络来实现空中监控的威胁评估（threat evaluation），可以融合多种因素和信息，并且具有一定的容错性。 [49] 使用经过预先训练的神经网络在经过卡尔曼滤波提取出的多传感器特征中学习，用来实现目标追踪，并进而实现机器人运动控制。

多模态信息融合的意义在于，不同模态信息之间往往具有天然的互补性，例如 RGB 图片可以提供颜色和纹理特征，深度图片可以提供三维形状信息和边界特征，触觉信息可以提供物体的力学特征，滑觉信息可以提供物体的摩擦力特征等等。而利用这些模态之间的互补性有助于挖掘信息的潜力，提高机器人识别、感知和操作的精确度，提高机器人系统及智能系统的稳定性。多模态融合的方法与多传感器融合的方法有类似之处。例如 [51] 使用了朴素贝叶斯方法将颜色和材质信息融合起来进行图片分割，获得了优于只使用单一模态进行图片分割的效果。而 [52] 则使用了相机产生的除了图片信息之外的元数据，包括曝光时间、闪光信息等，将其与图片信息在概率层进行贝叶斯融合，并在区分室内室外和识别日落场景等任务中都取得了更好效果。 [53] 将视觉和触觉信息融合起来，实现了对于机械手的高速操纵，和更加熟练的控制。另外， [54] 将一个低光摄像头，一个短波红外摄像头和一个长波红外摄像头获得的实时图片信息融

合起来输入神经网络，构建出夜视图片，获得了较好的效果。不过多模态融合的方法与多传感器融合的方法也有不同之处。因为不同模态的信息结构和维度都可能有很大不同，所以有时数据层和特征层的融合是不可行的，只能进行决策层的融合 [55] 。如果想要将不同模态的信息在数据层或特征层融合起来，可以像 [56] 一样采用为每一模态学习不同的度量的方法来解决，这样可以通过不同的度量来正规化每个模态的信息，使它们在融合时得到平衡。我国在信息融合领域一直紧跟研究前沿，近年来的研究成果包括 [57]、 [58]、 [59] 等。

机器学习的快速发展带给了我们诸多便利 [60] 。现在我们不仅可以使用 Bag of words 来提取信息 [61] ，使用 SVM [62] 或贝叶斯分类器 [63] 来分类信息，还可以使用各种各样的神经网络来完成模型学习，提取特征，以及分类、识别、感知、分割等任务 [64,65] 。近年来深度学习日趋成熟，被广泛应用于图片、语音、视频的识别 [66]，并且也逐步被应用于多模态融合。[67] 提出了一种使用深度神经网络学习多模态信息的方法，并在图像——声音数据集中得到了验证。而使用 RGB-D 信息融合的研究也越来越多。[68] 在 2011 年收集了一个大型的包含属于 51 种日常用品的 300 个物体实例的 RGB-D 数据库，并给出了一些识别、分割任务的 Benchmark，比如使用 Linear SVM 和 Kernel SVM 以及随机森林（Random Forest）的识别准确率。接下来被应用于 RGB-D 信息融合的方法有：实例距离学习（Instance Distance Learning）[69]，核函数描述符（Kernel descriptor）[70,71]，分层匹配追踪（HMP）[72]，空间金字塔与分层匹配追踪（SP-HMP）[73]，卷积递归神经网络 [74,75]，决策树（decision trees）[76]，以及深度学习 [77] [78]。具体地说，[69] 将每一个新的场景信息跟已有的信息相比，并计算出距离。而词袋模型（Bag of Words）方法是将原始信息投影到另一个稀疏空间，转换为一种稀疏编码 [79]，使其更加可分。 [73] 使用了两层 HMP，第一层用来进行信息提取，第二层用来进行空间压缩，并且可以和第一层联合起来构成空间金字塔，有效地提升了准确率。[74] 使用一组随机的递归神经网络在 CNN 特征的基础上进行信息提取，证明了随机权值的神经网络也可有效地提取信息。不过 [74] 虽然使用了神经网络，却没有使用神经网络进行学习，而是采用随机的权值。[76] 使用了一组随机的决策树投票产生最终结果。 [77] 则使用了在 ImageNet [80] 中训练过的 Caffe [81] 深层卷积神经网络模型分别在 RGB 和深度图片中提取特征，即将神经网络中的最后一个卷积层的特征作为输出。 [78] 也使用同样的模型，其不同之处是：第一，并没有直接使用该模型，而是在原有基础上使用 RGB-D 的

信息进一步微调；第二，也不是使用此模型提取信息，而是直接改变了此神经网络的结构使其可以接受 RGB-D 信息，输出分类结果。

1.3 研究内容

本文的研究内容整体上可以分为三个部分，分别对应着1.1 小节中所提出的三个问题：

1. 寻找适用于多模态信息和多信息提取方法的不同层次的信息融合策略，有效利用模态间和方法间数据互补性的方法。
2. 深度学习在多模态信息融合中的应用。
3. 深度学习中多模态信息数据量不足的问题。

研究内容:

1. 多模态信息方法间的融合策略

   首先，由于 RGB-D 识别的应用需求巨大，应用前景广阔，我们选用了上文中提到的由华盛顿大学收集并处理的 RGB-D 数据集 [68] 作为我们的训练集。其次，我们选用了理论发展较为成熟且在分类、识别等领域的效果已经得到证实的信息提取方式：分层匹配追踪和以及随机权值卷积递归神经网络，对颜色和深度图片分别进行信息提取。其中在使用随机权值神经网络时我们使用了预先在更大的数据集上学习的卷积过滤器（filter）来进行图片卷积特征提取。然后使用一组随机产生的递归神经网络对卷积模式进行降维处理，相当于构建了空间金字塔（Spatial Pyramid），然后选取空间金字塔的最上层，作为分类特征。实验表明，虽然神经网络的权值是随机产生的，但是将一组（含 64 或 128 棵递归神经网络）结合起来之后往往可以取得较好的识别效果。如前所述，因为不同模态之间具有“天然的”互补性，所以我们将颜色和深度的高层特征提取出来之后直接进行连接，使用 SVM 分类器分类即可取得不错的准确率的提升；不过由于不同方式提取的信息之间结构、维度差异较大，直接将它们在特征层融合起来的效果并不好，所以我们使用了决策层融合的方式。为了高效地产生决策层投票权值，我们使用并改进了 SVM 分类器，使其可以输出分类概率，然后在此基础上采用最大似然的方法进行方法间的决策层信息融合，并探究其方法互补性。

多模态信息方法间的融合策略的意义在于，利用优秀的方法间融合策略可以更高效地执行（节省计算资源），更好地利用不同方法之间的互补性，弥补不同方法的不足。在分类、识别等任务中获得更好的效果（提高准确率，降低不确定性）。并且具备更好的可拓展性（即可以更加方便地加入新的分类方法）。而且根据之前的文献综述可以看出，大量已有工作都是局限于利用多传感器或不同模态之间的融合，而我们引入的不同方法之间的融合，可以成为一个新的研究方向。

2. 引入深度学习方法

   其次，由于深度学习发展迅速，在许多任务中已经超过传统的机器学习方法，所以本文为了突破传统信息提取方法的局限，还引入了深度学习方法。在实际操作中我们使用了八层的深度卷积神经网络在颜色信息和深度信息中分别进行深度学习，然后使用学习出来的模型将相应的特征层分类信息提取出来，并且对不同模态在特征层进行融合，再使用 SVM 分类器给出识别结果。此外，本文还提出了一种将深度信息转换为图片信息的正规化方法。此种方法对于深度信息的不同部分采用不同的正规化尺度，这种方法的优势是：一方面保留了目标物体与背景（和噪声）之间的间隔，另一方面还保留了详细的目标物体之内立体形状变化的信息。

   引入深度学习方法的意义在于，它在许多方面可以弥补传统信息提取方法的不足，突破传统方法的局限。尤其在数据量足够大的时候，可以获得远高于传统方法的分类、识别准确率。

3. 引入 Pretrained 模型及综合模型的提出

   同时，为了解决深度学习中数据量不足的问题，我们使用了预先在海量图片数据集 ImageNet [80] 中训练过的模型来进行信息提取。由于已有模型是在颜色信息数据集中训练的，对于颜色信息我们直接使用了 Pretrained 模型，在对深度信息进行提取的之前，我们使用已有深度数据对原有模型进行了微调，以便使它可以更好地适应深度信息的提取。最后，我们基于以上研究提出了综合分层匹配追踪、随机权值卷积递归神经网络以及深度学习三种方法的信息融合识别模型。将多种方法及不同模态在不同层次采用不同的融合方法进行综合使用，得到最终的分类结果。

   引入 Pretrained 模型的意义在于，在训练数据不足的时候也可以使用深度学习获得更好的效果，同时还提高了训练效率。综合分类模型的意义在于，

提升识别的准确率和稳定性，并且对于大规模的数据集和小规模的数据集均可取得较好的识别结果。

1.4 主要贡献

本文的贡献主要有以下五点：

1. 本文提出了利用方法间的互补性来提高多模态信息的利用效率的思路。使用并改进了 SVM 分类器，使其可以输出分类自信度，然后在此基础上采用带权投票的方法将包括分层匹配追踪、神经网络在内的多种信息提取方法融合起来，并探究其方法互补性。实验证明，将不同模态的信息或不同信息提取方法进行融合都可以使得我们对目标物理的特征的理解变得更加全面，可以获得更好的识别效果。
2. 本文在传统信息提取方法之外还引入了深度学习方法。即使用多层 CNN 在颜色信息和深度信息中分别进行学习，然后将相应的特征层信息提取出来，并进行特征层的信息融合，使用 SVM 分类器给出识别结果。
3. 本文还提出了一种将深度信息转换为图片信息的正规化方法。此种方法对于深度信息的不同部分采用不同的正规化尺度，从而实现了既保留了目标物体与背景之间的间隔，同时又保留了目标物体之内形状变化的目的。实验证明，所提的深度信息正规化方法可以提升深度学习的识别准确率。
4. 本文使用了 Pretrained 模型来辅助深度学习提取信息。而且由于已有模型是在颜色信息数据集中训练的，我们使用原模型直接进行颜色信息的提取，使用在深度信息数据集上微调过的模型进行深度信息的提取。实验证明，使用 RGB 图片预先训练的深度学习模型可以较好地识别颜色和深度信息。
5. 最后，本文基于以上研究提出了综合分层匹配追踪、随机权值卷积递归神经网络以及深度学习三种方法的信息融合识别模型。实验证明，我们的融合模型可以较好地利用模态间以及方法间的信息互补性，提升识别的准确率和稳定性，并且对于大量的数据和小量的数据均可取得较好的识别结果。

1.5 本文结构

本文接下来的章节有：第 2 章介绍了一些与本文研究相关的背景知识；第 3 章主要介绍了方法内以及方法间的多模态融合策略探究，包括其融合思想、操作步骤和实验结果；第 4 章介绍了使用深度学习进行多模态信息融合的思路、具体操作和实验结果；第 5 章综合之前的研究提出了一个使用了多模态融合的综合识别模型；第 6 章给出了对本文的总结和对未来研究的展望。

第 2 章　背景知识

本章主要介绍一些与本文研究相关的背景知识。

2.1　多模态信息融合

2.1.1　多模态信息融合的概念

信息，根据香农（Shannon）的定义，是“用来消除事物不确定性的数据”，并且使用熵来量化描述信息的质量 [82] 。控制论奠基人维纳（Norbert Wiener）认为信息是“人在与外部世界交互、适应并反作用于外部世界时，同外部世界交换的内容” [83] 。以上可以看作是对于信息的经典的定义。不过现代科学技术日新月异，现在将使用信息的主语限定于人已经不是十分恰当了，因为计算机和机器人也可以使用信息做出判断，维纳研究的控制论就是计算机利用信息的例子 [84] 。狭义的多模态信息指的是使用不同种类的传感器收集到的信息，这其中的重点是“多种传感器”，它相对于多传感器信息，强调因为传感器和信息种类的不同而产生了不同来源、不同结构的多模态信息 [85] 。而广义的多模态信息则主要强调信息结构上的异构性，因为异构性而产生的互补性 [86] 。

对于多模态信息融合的定义，我们选用 Joint Directors of Laboratories（JDL）的定义：对多种信息在多个层次上、使用多种方法的联合处理过程，以便实现多种信息的互相组合、互相关联和联合利用 [87] 。使用多模态信息融合的意义是更好地利用不同模态的信息之间的互补性，以便提供相辅相成的综合认知和理解，提高整体决策的可靠性和鲁棒性 [88] 。以及对多个或多种传感器收集的信息进行联合利用，获得比单个或单一种类的传感器更加准确的结果 [32] 。

2.1.2　多模态信息融合的层次

实际操作中我们对于信息的处理往往都是分层进行的 [89] ，而对于多模态信息的处理，往往可以分为三个层次 [90] ，分别是：

1. 数据层
2. 特征层

3. 决策层

下面逐一进行说明。

数据层。对于不同传感器收集到的原始信息直接进行处理，称为数据层的操作。在这一层次进行处理和融合的优点有：包含第一手原始数据，因为没有经过信息提取，也就没有信息损耗，可以说含有的信息量是最大的。缺点包括：数据量庞大，数据维数高，表征效率低，因而给信息提取造成了方法上和效率上的困难；同时受噪声影响比较大。

特征层。对底层数据进行初步处理，并且根据需求提取出来的特征，称为特征层。特征层的特征基于对底层数据的提取和抽象，可能已经包含一定的经过提炼的模式信息，比如使用单层 CNN 提取的信息会包含物体的边、角、形状等模式 [91] 。在这一层次进行处理和融合的优点有：经过提取的信息往往已经对于原始数据中某些规律有所把握，具有更适合于某种特定任务（比如分类、分割）的结构，很大程度上消除了噪声的影响，数据维度适中，因而在此层面上的处理效率较高；同时和决策层相比，特征层信息的灵活性更大。缺点包括：在提取过程中可能造成一定的信息损失，且没有数据层的操作灵活。

决策层。对数据层或特征层的信息加以自动化或半自动化的分析提炼，得出具有一定语义的结果，即为决策层信息 [92] 。决策层基于特政策和数据层，经过进一步的提炼，已经得出了可以被人理解和使用的语义信息。在这一层次进行处理和融合的优点有：数据维度已被高度压缩，在这一层的处理和判断时间及空间复杂度很低，抗噪声干扰的能力最强；并且由于具有语义信息，因而可以被人直接理解和利用。而缺点则包括：由于信息高度精炼，可能有较多的信息损失，且在这一层实施融合策略的灵活性很低，往往只可以使用投票 [93] 、专家系统 [94] 等高层方法。

第 3 章　方法间的多模态融合策略探究

本章主要介绍方法内以及方法间的多模态融合策略探究，包括其融合思想、操作步骤和实验结果。融合策略在多模态信息利用理解中起着举足轻重的作用，优秀的融合策略可以提高处理的效率（降低时间复杂度），降低成本（降低空间复杂度），更好地利用相辅相成的数据以及方法之间的互补性，弥补不同模态以及方法的不足。在分类、决策、识别等任务中获得更好的效果（提高准确率和鲁棒性）。好的融合策略往往还具备更高的可拓展性。在本章，我们首先利用传统的分层匹配追踪，以及随机权值卷积递归神经网络来进行颜色和深度信息的提取和融合，并探究不同模态的信息之间的互补性。然后我们在这两种方法的基础上而引入了不同方法之间融合策略探究，提出了一种决策层的信息融合策略，并以此探究不同方法之间的互补性。实验表明，不仅仅不同模态的信息之间具有互补性，不同信息提取方法之间也具有互补性。将不同模态的信息或不同信息提取方法进行融合都可以使得我们对目标物体特征的理解和把握变得更加全面，可以获得更好的识别效果。大量已有工作大都关注多传感器或不同模态之间的融合，而我们引入的不同方法之间融合，可以进一步利用数据的互补性，弥补各个方法的不足。

3.1　分层匹配追踪

图 3.1　分层匹配追踪的工作流

分层匹配追踪（HMP）[72] 的中心概念是自动地学习低层和中层的特征，而不是使用人工选择的特征，例如 SIFT [95] 等。HMP 从原始数据中学习到一组稀疏字典，并使用这些字典给数据分层地编码，获得一个空间金字塔。如图 3.1 所示。具体地说，给定一组 h 维的观测值 $Y = [y_1, ..., y_n] \in R^{h \times n}$，K-SVD 的学习目标是找到一组字典 $D = [d_1, ..., d_m] \in R^{m \times n}$（这里的 d_i 可以称为过滤器），和一组稀疏编码 $X = [x_1, ..., x_n] \in R^{m \times n}$，使得如下重建误差最小化：

$$\min_{D,X} ||Y - DX||_F^2 \qquad s.t.\ \forall i,\ ||x_i||_0 \leq K \tag{3-1}$$

其中 $||X||_F$ 表示 X 的弗罗贝尼乌斯范数 (Frobenius norm)；x_i 是 X 的一列；$||*||_0$ 是零范数，数值等于 x_i 中为零项的个数；K 是稀疏级别，表示每一组数据中为零项个数的上限。因为 3-1 中的问题是非凸的，很难给出精确解，所以在实际操作中我们采用近似解法 OMP [96] 来贪心地求解。具体地说，OMP 会迭代地交替执行以下两个过程，逐步逼近最终结果。

$$\min_{X} ||Y - D^* X||_F^2$$

$$\min_{X} ||Y - DX^*||_F^2$$

通过 K-SVD 的提取出来的信息，可以较好地重构出原来的图片，如图3.2所示。重构出的图片与原图质量相差无几，说明尽管在使用分层匹配追踪时信息的维度和结构发生了较大变化，但大部分有用信息都被保留了下来。

原始图片

使用K-SVD信息重建图片

图 3.2　K-SVD 保留信息可视化

如果将 K-SVD 学习得到的字典可视化，可以得到如图3.3所示的效果，可以看出：在深度、颜色信息之中学习到的字典（过滤器）均包含一定的有规律可循的模式，比如说边、角、点。其中 RGB 图片的字典含有颜色信息，而深度图片学习到的字典虽然没有颜色信息，但是边角分界处往往更加锋利清晰，这也体现了两种模态信息的互补性。

空间金字塔最大池化（Spatial pyramid max pooling）是一种高度非线性的信息处理过程，可以从局部稀疏编码中生成更高层次、更低维度的表征方式。它的结构如图 3.4所示 [79]。这种方法在多个层次上递归地进行最大池化。具体地说，假设最大池化的接受域为 2×2，那么第零层每个单元就代表原始数据中的一个单元, 第一层每个单元代表原始数据中的四个单元，以此类推。设 U 为池化前的数据，U 的每一列为原始数据中的邻域，z 代表着池化之后的数据，那么最大池化可以表示为：

$$\begin{aligned} z &= F(U) \\ z_j &= max\left\{|u_{1j}|, \ldots, |u_{Mj}|\right\} \end{aligned} \tag{3-2}$$

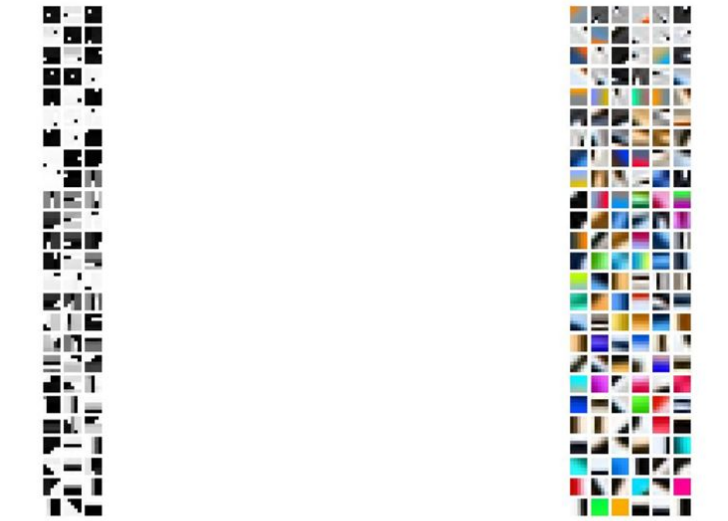


图 3.3　K-SVD 字典可视化

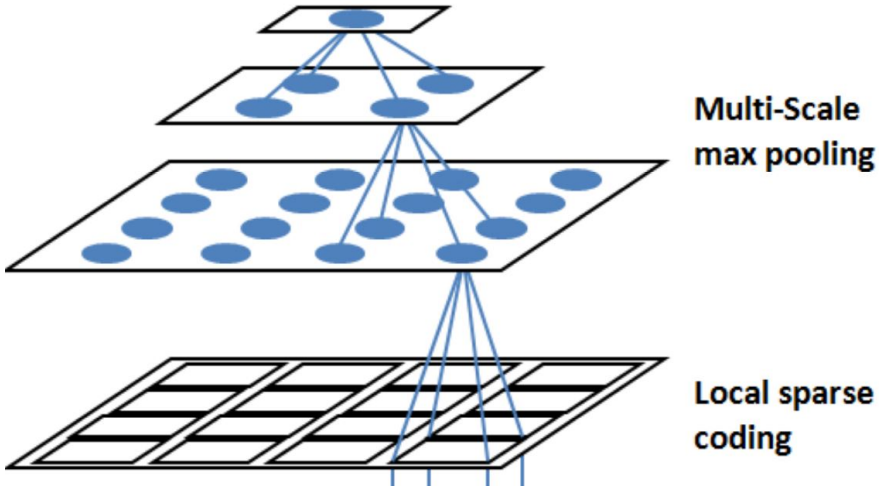


图 3.4　空间金字塔最大池化示意图

具体的参数及实验效果见 3.4 小节。

3.2　随机权值神经网络

如果说分层匹配追踪是一种目的明确的监督学习（supervised learning），那么随机权值神经网络（CRNN）的信息提取过程就显得比较“随意”。在这里，我

们引入 CRNN 作为另一种参与融合的信息提取策略，一方面是因为它可以引入一定的随机性，另一方面是因为通过神经网络的信息提取过程与传统的词袋模型（Bag of Words）具有很大不同，提取出的信息在结构上也与之有较大差别，所以可以预见 CRNN 与传统信息提取方法有着较大的互补性，将它们融合可以获得较大的准确率上的提升。由于一定随机性在许多人工智能策略中都可以避免陷入局部最优，提升策略表现，比如遗传算法 [97] 和蒙特卡洛方法 [98] 等。所以我们也试图使用 CRNN 来给我们的模型引入一定的随机性。CRNN 的工作流程图如图3.5 所示。

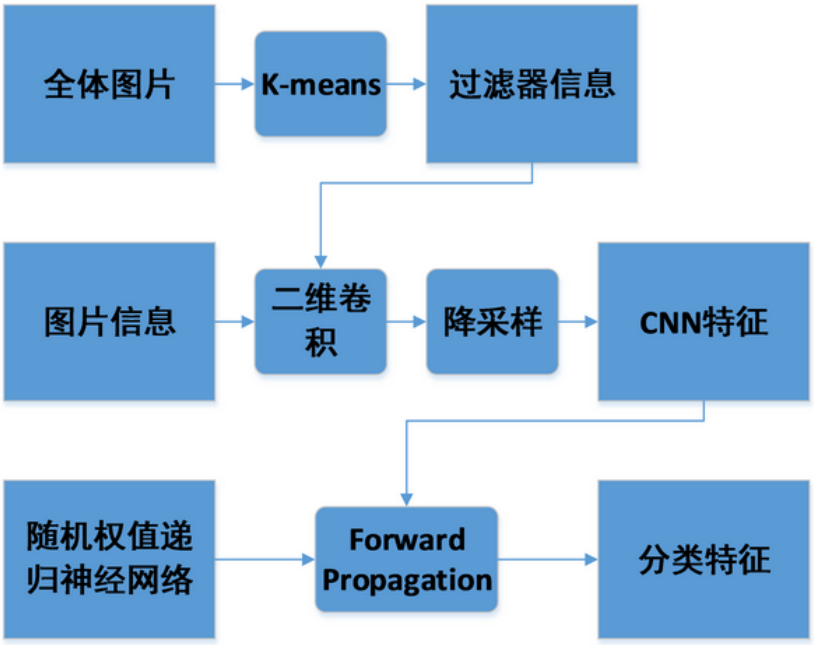


图 3.5　CRNN 的的工作流

如图3.5所示，我们需要先得到一组过滤器，这个过滤器不一定需要在我们的训练集上学习出来，也可以在更大、更全面的数据集上预先通过 K-means 学

习出来。K-means 的目标函数是：

$$\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{3-3}$$

其中 k 是 K-means 产生的中心的数量，S 是分类信息，μ_i 是 K-means 产生的中心，$x \in S_i$ 说明信息 x 以 μ_i 为中心。学习到的过滤器如图 3.6所示。

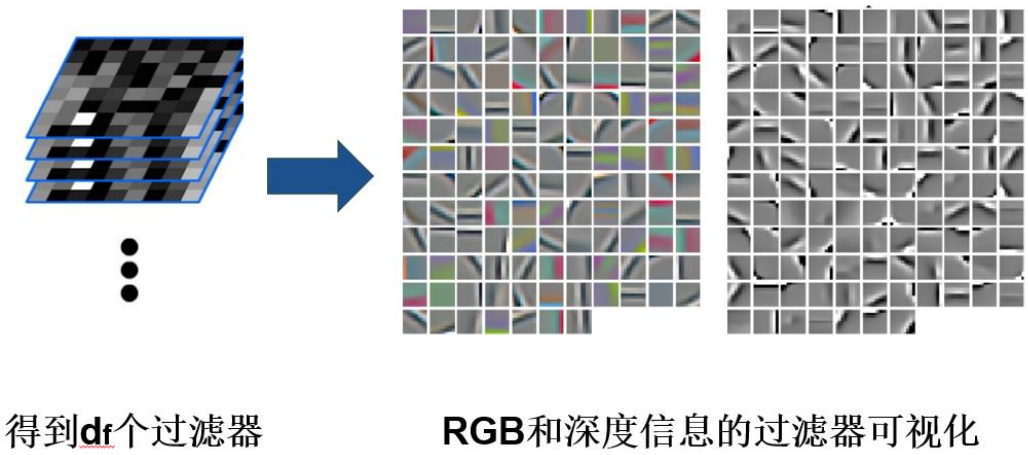


图 3.6　CNN 的过滤器可视化

然后我们使用 Pretrained 的过滤器信息与图片信息进行二维卷积和降采样，如图 3.7 所示，以便得到图片的 CNN 特征。

由于我们使用一组过滤器分别与原图片卷积，所以对于每一张图片输入会得到一组 CNN 的模式（pattern）。通过 CNN 提取出的某个目标物体（一个苹果）的 pattern 如图 3.8所示。可以看出，左边的 RGB pattern 含有一定的光照、纹理信息；而右边的 Depth pattern 则含有更加锋利明显的边界信息，这也体现了两种模态信息的互补性。

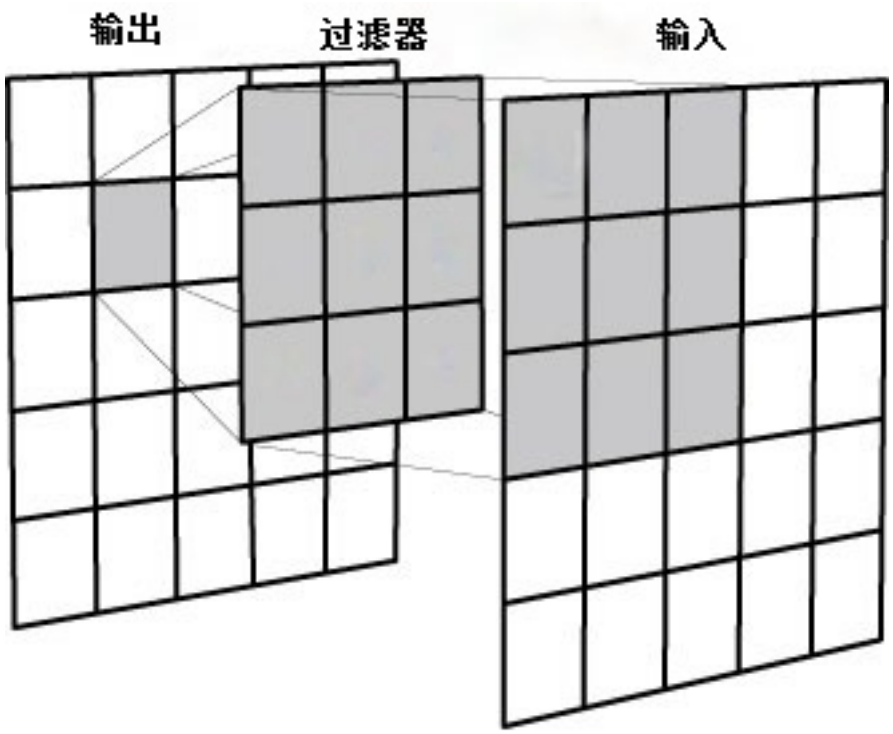


图 3.7　CNN 的结构

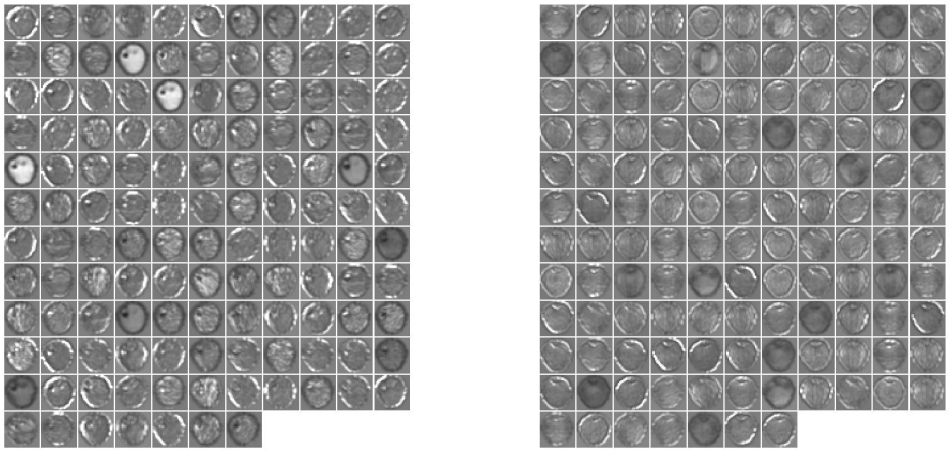


图 3.8　CNN 所提取出的模式（左为 RGB patter，右为 Depth patter）

然后我们将得到的 CNN 模式输入一组递归神经网络（RNN）。注意，这里

我们虽然使用了神经网络，但是由于其权值是随机产生的，不会在信息提取的过程中改变，并且我们只使用前向传递（forward propagation）而不使用后向传递（back propagation），所以这并不是一个监督学习的过程，而是一个非监督（unsupervised）的信息提取过程。在这一步中我们对于每一层的每个接受域（相近的指定大小的区域）中的数值施加一个变换，得到一个数值作为更高层相应单元中的数值。具体地说：

$$x_i^h = f(X_i^{h-1}) \tag{3-4}$$

其中，x_i^h 代表 h 层的单元 i 的数值，X_i^{h-1} 代表 h 层的单元 i 在 h-1 层对应的接受域内的数值矩阵。可以看出，RNN 其实是空间金字塔最大池化（Spatial pyramid max pooling）的推广——将 3-4 式中的 $f()$ 函数设置为取 max 就是空间金字塔最大池化。不过我们在分类时往往只使用上层的信息。RNN 的工作示意图如图 3.9 所示。

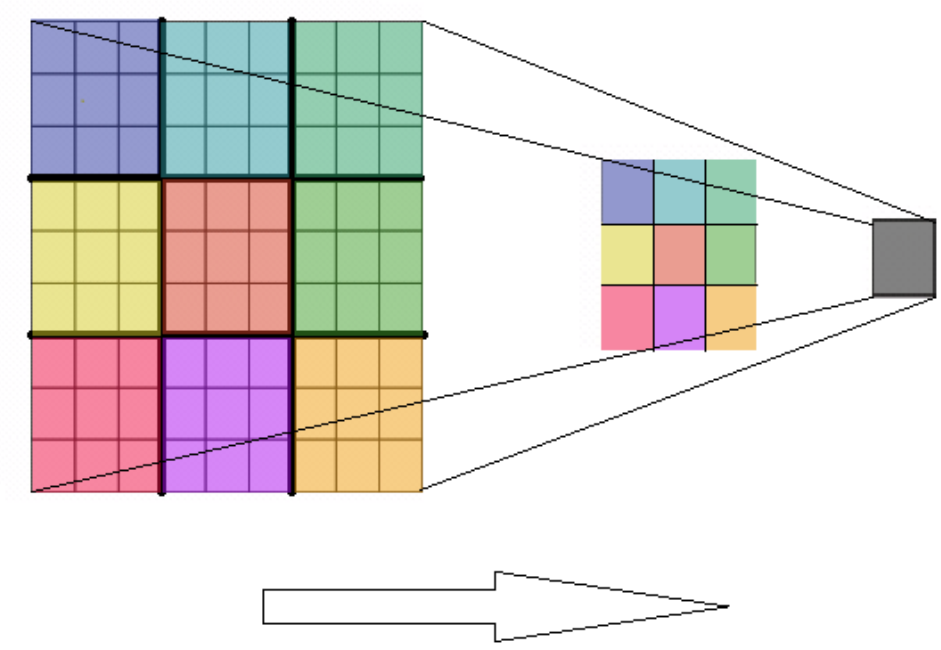


图 3.9　RNN 工作示意图

具体的参数及实验效果见 3.4 小节。

3.3 多模态融合策略

3.3.1 特征层融合

对于同一方法内，不同模态间的信息融合，我们倾向于使用特征层的融合策略。因为正如前文所述，RGB 信息和深度信息有着天然的互补性，而这些互补性在特征层体现得最明显。如果我们先分别利用颜色信息和深度信息得到一个分类的估计值，这样虽然具有了一定的语义信息，也可以从贝叶斯的角度来增大最终分类的准确率，但是这样就不能完全挖掘颜色和深度信息之间的联系与互补性。在实验中我们尝试了使用决策层的融合策略和特征层的融合策略，发现对于同一方法内的融合，在特征层的操作确实在大多数情况下会取得更好的结果。

具体地说，我们使用连接作为特征层融合的方法：

$$X_{feature} = [X_{RGB}, X_D] \tag{3-5}$$

其中 $X_{feature}$ 是特征层融合得到的信息，X_{RGB} 是颜色图片提取出的信息，X_D 是深度图片提取出的信息。

3.3.2 决策层融合

而对于不同方法之间的信息融合，我们更倾向于使用决策层的融合策略。原因如下：

1. 由于方法的不同，各组特征之间维度差异较大，在特征层难以权衡各个方法在总体决策中占的权重，维数越高的信息在特征层融合之后占总体决策权重就会越大。
2. 各组特征结构差异较大，如果使用特征层融合，难以捕捉其中的联系。相反，有可能造成过拟合。
3. 由于在特征层融合之后特征的维度往往比较高，对于接下来的决策步骤中的计算资源要求也较高。而采用决策层融合的策略可以分而治之，将特征压缩为带有语义的信息再做处理。

在实验中，由于需要两个分类器带权投票，我们需要能够产生分类自信度的分类器，所以优先试验了贝叶斯分类器。但是发现贝叶斯分类器在信息矩阵

维数较高时计算过于缓慢，且效果也不如 SVM。所以我们决定使用 SVM 分类器来进行分类决策。不过因为原生的 SVM 只能够产生分类结果，而不能产生分类自信度，为了实现决策级的信息融合，我们使用 Sigmoid-Softmax 方法将 SVM 的决策值转化为自信度估计，如 3-6 式所示。

$$P(C_j|I_i) = \frac{Sigmoid(dec\_value_{ij})}{\sum_k Sigmoid(dec\_value_{ik})} \tag{3-6}$$

最终，我们在使用 3-6 式得出两种方法各自的分类自信度之后投票决定最终输出结果。实验证明，我们提出的融合策略可以有效地捕捉方法间的信息互补性，使得最终分类结果有所提升。整体融合策略的框架如图 3.10 所示。

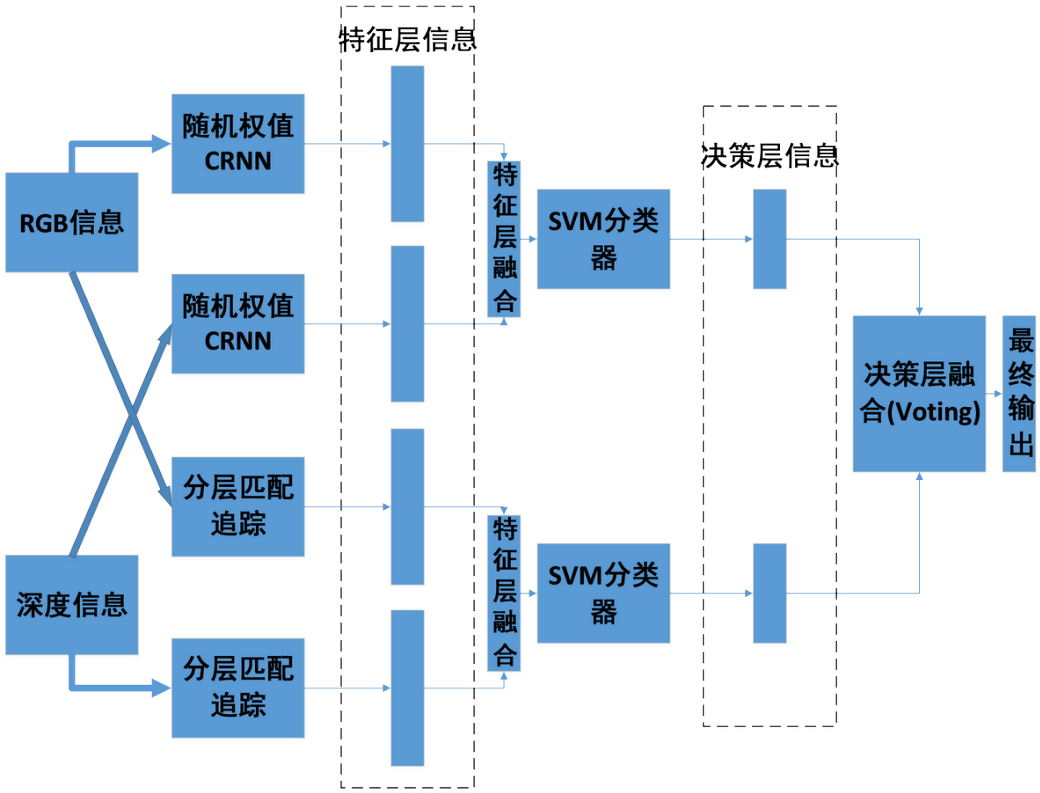


图 3.10　整体融合策略的框架

具体的参数及实验效果见 3.4 小节。

3.4 试验验证

3.4.1 试验综述

我们选用了上文中提到的华盛顿大学 RGB-D 数据集 [68] 作为我们的训练集。此数据集包含分别属于 51 个类别的日常用品的 300 个物体实例，如图 3.11所示。总共包含超过 20 万组匹配的 RGB-D 图片，在实验中我们为了与 [68] 的 baseline 对比，所以选用了其中的 41877 组 RGB-D 信息。

图 3.11　华盛顿 RGB-D 数据集

在本节中我们对颜色信息和深度信息分别用 CRNN 和 HMP 两种方法进行分类，然后将两种模态在不同层次融合起来进行分类，以观察模态间融合的效果提升。然后我们为了探究不同模态之间的互补性，对比了两种模态的混淆矩阵，并对其中典型的例子进行了探讨。最后我们统一使用 RGB-D 信息，用两种方法进行分类并尝试在不同层次进行融合，以探究方法融合的效果。其中 SVM 分类器的实现我们选择使用 liblinear [99] 。

3.4.2 试验参数

对于分层匹配追踪，我们需要在两个层次上设置相应的参数。

第一层。从 1000000 个 5×5 的深度图片样本中学习出 75 组稀疏度为 5 的灰度字典和深度字典；从 1000000 个 $5\times5\times3$ 的 RGB 图片样本中学习出 150 组稀疏度为 5 的颜色字典和曲面法线字典。接下来我们使用上述字典，使用 OMP 算出原图对应的稀疏编码，并使用空间金字塔最大池化的方法产生 16×16，4×4，2×2，1×1 的信息，将它们拼接在一起压缩成一维，作为第一层的输出特征。

第二层。在第一层输出的基础上从 1000000 个 5×5 的深度图片和 RGB 图片产生的输出特征中学习出 1000 组稀疏度为 10 的灰度字典、深度字典、颜色字典和曲面法线字典。接下来我们使用上述字典，使用 OMP 算出第一层输出对应的稀

疏编码，并使用空间金字塔最大池化的方法产生 3×3，2×2，1×1 的分割信息，将它们拼接在一起压缩成一维，得到总共 4(*channel*)×1000(*dict*)×(3×3+2×2+1×1) = 54000 维向量，作为第一层的输出特征（其中 RGB 和深度信息分别为 28000 维）。因为第一层信息维度过高，所以在实验中我们只使用第二层的信息。

对于 CRNN，我们同样需要在两个层次上设置参数。

CNN 层。学习出 128 个过滤器，分别和 RGB 和深度图片卷积之后降采样得到 27×27 的图片块，所以对于每一张 RGB 或深度图片分别可以得到 128×27×27 的图片块（见图片 3.8）。

RNN 层。利用 CNN 层得到的维度为 $128 \times 27 \times 27$ 的图片块，将每一个过滤器对应的 27×27 的图片块分别通过 128 个随机产生的 RNN，每个 RNN 得到 1×1 维特征。所以最终我们得到每张 RGB 和深度图片对应的信息分别为 $128 \times 128 = 16384$ 维，总共有 32768 维输出信息。

3.4.3 试验结果

在实验中，我们采用以上的参数设置，得到如下实验结果：

表 3.1 模态间融合实验结果

模态 / 方法	RGB	Depth	RGB & Depth	
			特征融合 *	决策融合 **
CRNN	81.7 ± 1.4	78.9 ± 1.6	**87.6 ± 1.2**	87.3 ± 1.1
HMP	75.8 ± 3.4	78.4 ± 1.9	85.4 ± 2.6	**86.3 ± 2.1**
决策融合	83.1 ± 2.6	82.5 ± 1.9	**89.6 ± 2.1**	89.1 ± 1.8

*：使用 3-5 式

**：使用 3-6 式

通过表 3.1 我们可以看出，多模态信息融合可以明显提升分类准确率。下面我们以匹配追踪为例，来探究两种模态的信息互补性。两种模态信息分类的混淆矩阵（Confusion Matrix）如图 3.12 所示。混淆矩阵 [100] 是在监督学习中对于分类结果的可视化，其每一列表示了一个预测类别；其每一行表示了一个真实类别，一个数据实例属于第 i 行第 j 列当且仅当此实例数据第 i 类，被分类模型分到第 j 类。颜色越深的区域中含分类实例越多，所以理想化的混淆矩阵应该是对角矩阵。通过对两者混淆矩阵的比较可以看出，它们是有一定的互补性的。例如图片 3.13 所示，如果只使用深度信息，容易将 peach 分类为 sponge，如果

只使用颜色信息，又容易将 sponge 分类为 rubber eraser。而将两种信息结合起来之后，上述三种物品均可以得到较好分类。

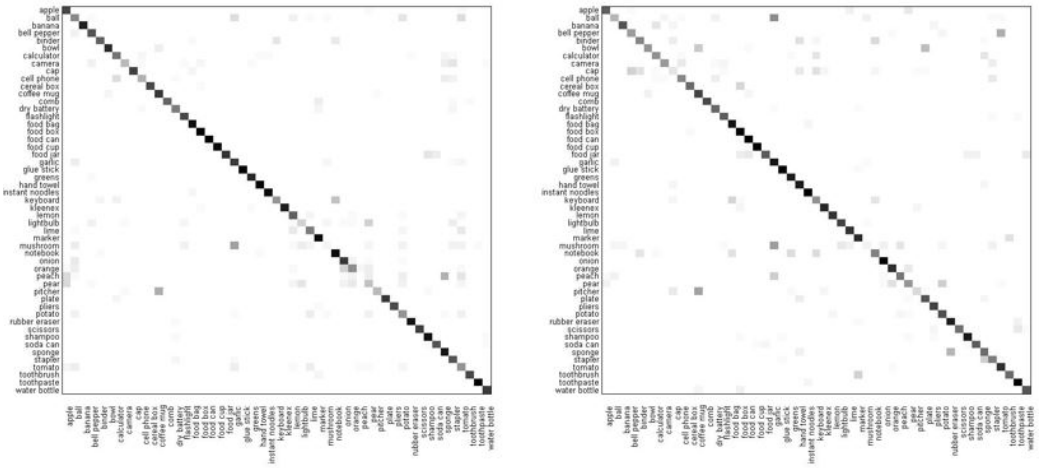


图 3.12　两种模态信息分类的混淆矩阵

图 3.13　分类效果举例

表 3.2　方法间融合实验结果

CRNN	HMP	CRNN & HMP 特征融合 *	CRNN & HMP 决策融合 **
87.6 ± 1.2	85.4 ± 2.6	88.7 ± 2.0	**89.6 ± 2.1**

本表展示的均是使用 RGB-D 信息得到的结果。

*：使用 3-5 式

**：使用 3-6 式

接下来我们使用 RGB-D 信息，探究两种信息提取方法之间的融合策略。得

到结果如表 3.2 所示。可以看出，不论是模态间还是方法间的信息融合都可以明显提高分类的准确率。而且正如我们预测的一样，在方法间采用决策层的融合可以更好的提升效果。我们在图 3.10 中提出的融合模型，可以综合使用图 3.1 和图 3.5 中的方法，并得到有效的效果提升。

第 4 章　深度学习在多模态融合中的应用

本章主要介绍使用深度学习进行多模态信息融合的思路、具体操作和实验结果。随着深度学习的飞速发展，包括分类、识别、检测乃至决策判断等诸多任务的记录都已经被刷新。在图片的识别中深度学习更是大显身手，展示出了惊人的威力，早已超过了传统的机器学习做法。所以本文为了突破传统的 Bag of Words 信息提取与处理的局限，引入了深度学习方法。并且我们不是使用深度学习方法直接给出分类结果，而是使用深度学习的模型来提取特征层信息，然后使用 SVM 分类器进行分类。这样做的原因是为了便于和其他方法在多个层次上进行融合，更加具有拓展性。

此外，本文还提出了一种将深度信息转换为图片信息的深度信息自适应正规化（Depth adaptive normalization）。深度信息自适应正规化的核心思想是对于深度图片的不同组成部分（例如目标物体、背景、噪声等部分）采用不同的正规化尺度。这种方法的优势有：可以保留目标物体与次要部分（比如背景和噪声）之间的间隔，使得目标物体形状轮廓清晰可辨；同时可以保留甚至放大目标物体内部三维形状变化的信息，使得目标物体含有的信息更加具体丰满；还可以一定程度上降低噪声和背景等对于识别结果的干扰。这种方法相比于其他深度信息到图片的编码方法相比，计算量小，不需要任何额外的信息。且可以与其他编码方法结合使用，即在这种正规化的方法之上再加入其他的编码信息。

最后，因为深度学习对于数据量需求巨大，而我们使用的多模态匹配的数据集往往具有数据量不足，过拟合的现象。为了解决前述问题，我们使用了预先在当前最大的 RGB 图片数据集 ImageNet [80] 中训练的模型来辅助信息提取。同时，由于已有模型的训练集只包含 RGB 图片，对于颜色信息我们可以直接使用原 Pretrained 模型，而对于深度信息的提取，我们尝试预先使用现有的深度图片信息对原有模型进行微调（finetune），使它更好地适应于深度信息的提取。

4.1　深度学习模型描述

在实验中我们使用了八层的深度卷积神经网络 AlexNet [101] 在颜色信息和深度信息中分别进行深度学习训练出一个适用于 RGB 的模型和一个适用于深度

信息的模型，然后使用相应的模型将高层特征提取出来，使用 SVM 分类器和 3-6 式做出决策并给出分类自信度。最后与其它分类模态在决策层进行融合。引入深度学习的意义在于，Deep Learning 在许多方面可以与传统词袋信息提取方法互相补充，有助于突破传统方法的局限。同时，深度学习最终提取出的特征由于高度压缩，维度较小。最后，在数据规模足够大的时候，深度学习能够获得远高于词袋方法的分类及识别准确率。使用深度学习与传统方法互相补充，就好比使用一个新型的、更高精度的传感器与传统的传感器联合做出决策。

AlexNet 的结构如图 4.1 所示（图片引自 [101]）。在实验中，我们提取 AlexNet 决策层之前的两层（分别称为 fc6 和 fc7）作为分类信息，将其提取出来并与其它特征进行融合。之所以选择这两层作为分类信息是因为以下原因：

1. 原模型第八层（最后一层）为决策层，会给出图片在 1000 个类别中的分类结果，且包含维度过少，有较大的信息损失。
2. 原模型第六层、第七层之前的信息层维数过高（例如：第五层含有 $13 \times 13 \times 128 \times 2 = 43264$ 维，数据冗余严重。而第六层第七层加起来一共只有 $2048 \times 4 = 8192$ 维）
3. 由于决策层实际上是一个 SoftMax 分类器，所以决策层之前的特征层的信息往往具有很好的可分性。

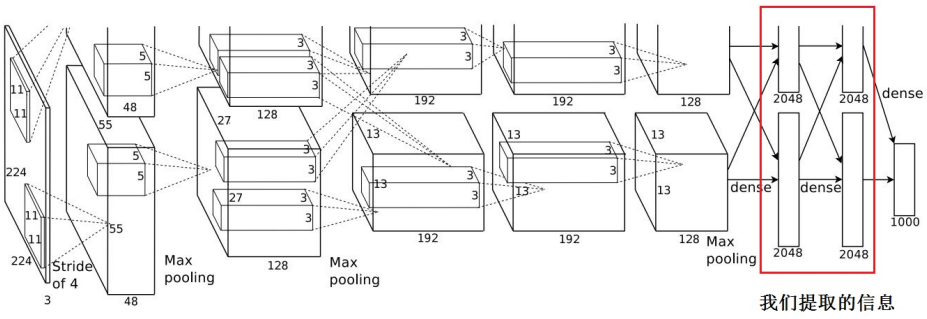


图 4.1　AlexNet 结构示意图

使用上述步骤我已经可以成功地提取信息并得出分类结果。具体的参数及实验效果见 4.4 小节。

4.2 Pretrained 模型引入

在上一小节我们已经可以提取信息，不过分类效果并不是很好。这主要是因为如前所述，我们的深度学习模型因为数据量不足而发生了过拟合。所以，为了解决深度学习中数据量规模不足的问题，我们引入 Pretrained 模型作为辅助。我们使用的 AlexNet 模型，是 ILSVRC12 比赛中在 ImageNet 中训练出来的 [102]。

由于已有模型是在 RGB 图片数据集中训练的，我们在使用它处理深度信息之前一方面需要把深度信息转为合法的深度图片，另一方面会先使用已有 Depth 信息对其进行微调（Finetune），使它更好地适应 Depth 信息的提取工作。试验证明，使用 RGB 图片预先训练的深度学习模型可以较好地提取颜色和深度信息。使用已有模型进行微调比从头开始训练有如下好处：

1. 微调是在之前训练的结果的基础上继续训练，相当于数据集被扩大了，训练出的模型的表现也可以超过从头训练的模型。
2. 训练过程缩短，节省时间和资源。

在微调过程中，我们保留 AlexNet 前七层的结构（如图 4.1 所示）与参数，将第八层替换为一个输出维度为 51 的决策层（因为我们的数据共有 51 类）。以深度信息为例，从头训练与微调的训练损失（Training loss）和测试准确率（Test Accuracy）的变化对比如图 4.2 和图 4.3 所示，可见在微调中，原模型的训练损失迅速下降，测试准确率迅速上升，这说明原模型可以很快地适应新的数据集。

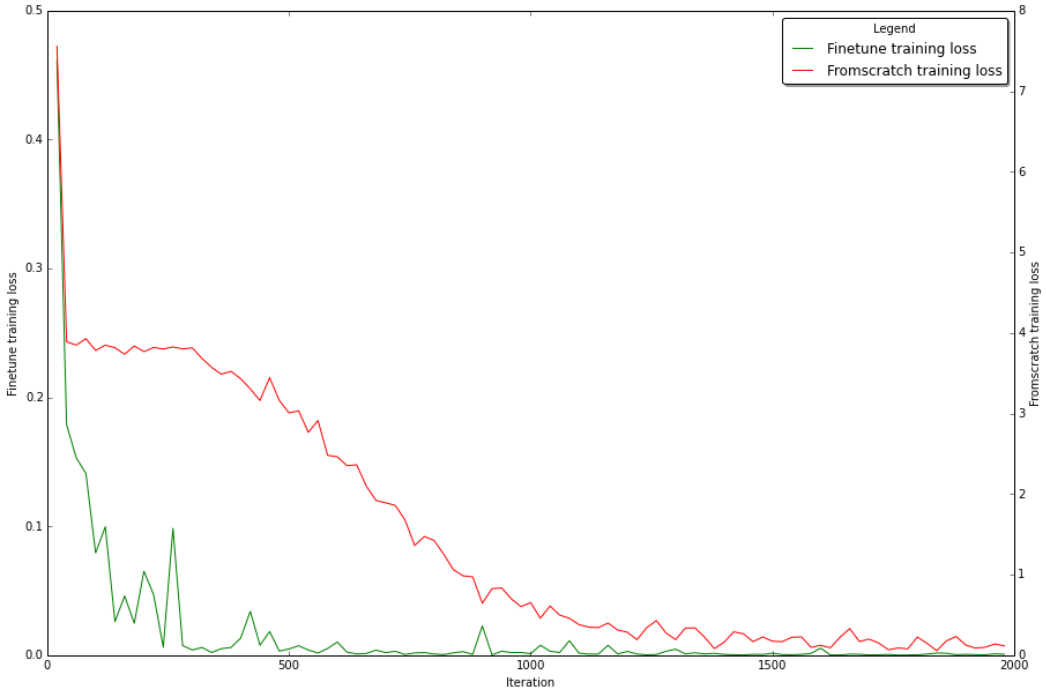


图 4.2　从头训练与微调的训练损失（Training loss）对比

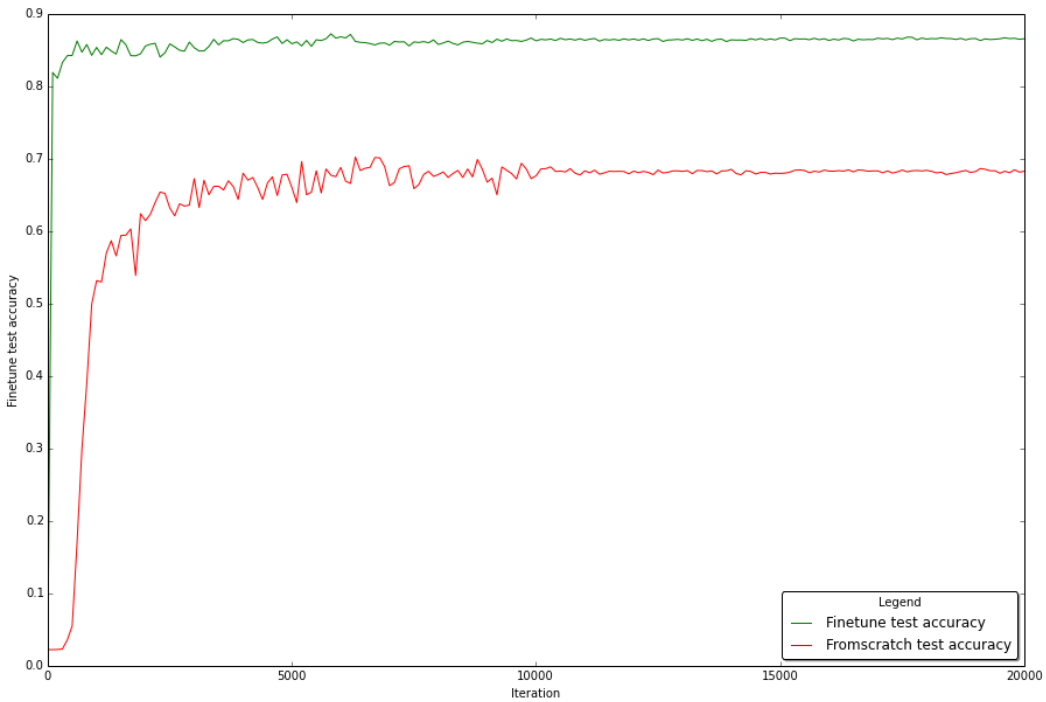


图 4.3　从头训练与微调的测试准确率（Test Accuracy）对比

通过引入 Pretrained 模型，我们可以比较成功地提取 RGB 信息并获得 85% 以上的准确率，已经接近 HMP 或 CRNN 方法中综合使用两种模态信息的表现。具体的参数及实验效果见 4.4 小节。

4.3 深度信息自适应正规化

通过引入 Pretrained 模型，我们已经成功地提取出 RGB 信息并只使用 RGB 信息获得 85% 以上的识别准确率。然而使用深度信息效果却不甚理想，经过研究发现，这是因为我直接将深度图片传入了 Caffe 而没有将它们正规化为合法的图片。由于深度信息是由单通道的深度数值构成的，而且每一个像素点上的数值范围可能不合法。而合法图片信息是由三通道构成的，且每一个像素点的数值在 [0 – 255] 之间。所以我就把深度图片正规化到了 [0 – 255] 之间，并将单通道扩展为三通道。不过这样做之后效果仍然不好。所以我又尝试了一些的预处理方法，比如说试图使用一些 mask 把背景遮住，或者使用 interpolation 把深度信息中的噪声点都修补上，但是效果都善乏可陈。

最后我想到，可能是正常的正规化的过程使得深度的“纹理”信息有损失。例如，原始深度图片上，物体的深度分布在 [720 – 780] 之间，而背景的深度分布在 [2000 – 3000] 之间，噪声的深度接近 0。而如果把这张深度图 normalize 到 [0 – 255] 之间的话，原来分布在 [720 – 780] 之间的物体信息就变成了 $[\frac{720}{3000} \times 255 - \frac{780}{3000} \times 255] = [61.2, 66.3]$ 之间，也就是说物体自身上的深度差异很大程度上被忽略了；当然，这样做也有好处，就是物体和背景之间的 gap 特别大，但是我觉得适当缩小 gap 并不会影响物体形状轮廓的可辨识度。

所以我提出了一种**深度信息自适应正规化方法**（Depth Adaptive Normalization, DAN），会有选择地保留、加重、淡化、忽略某些信息。规范化的描述见算法 1。

具体地说，就是先根据明显的 Gap 将深度信息划分为几个部分，然后根据需要保留相应的部分（比如目标物体信息），忽略不需要的部分（比如 gap，背景信息），如图 4.4 所示。

Algorithm 1 Depth Adaptive Normalization

Require: DepthMap $\in R^{n \times m}$, MaxGapNear, MaxGapFar, ThresholdGap

$DepthArray \leftarrow Squeeze(DepthMap)$

$DepthArray \leftarrow Sort(Unique(DepthArray))$

targetNear $\leftarrow$ min(DepthArray), targetFar $\leftarrow$ max(DepthArray)

Find the two gaps

for e_i in DepthArray **do**

  if e_i - e_{i-1} > ThresholdGap **then**

    if targetNear == -1 **then**

      targetNear $\leftarrow e_i$

    else

      targetFar $\leftarrow e_{i-1}$

    end if

  end if

end for

Normalize the two gaps

for e_i in DepthMap **do**

  if e_i < targetNear **then**

    $e_i \leftarrow 0$

  else if e_i > targetFar **then**

    $e_i \leftarrow$ (targetFar - targetNear) + MaxGapNear + MaxGapFar

  else

    $e_i \leftarrow e_i$ - (targetNear + MaxGapNear)

  end if

end for

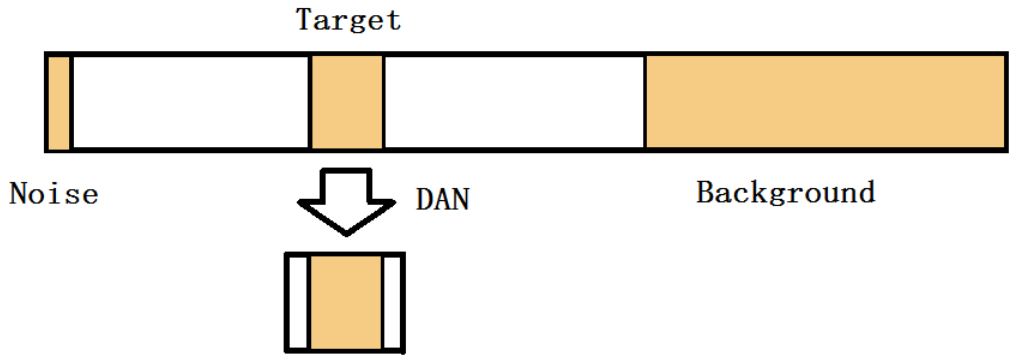


图 4.4　深度信息自适应正规化方法示意图

比如说上例中，设置 $MaxGap = 20$。0 到 720 之间 gap 为 720，把它缩小到 20，780 到 2000 之间 gap 为 1280，把它缩小到 20，2000 到 3000 之间很容易判断为背景，所以这部分数值都用 2000 代替（不需要考虑背景的纹理）。所以最终深度信息变为 [0 – 20][20 – 80][80 – 100]（其中 [20-80] 是原来的 [720 – 780] 之间的物体信息），把这些信息再缩放到 [0 – 255] 之间即完成了正规化。这样做的好处有：

1. 保留了目标物体与其他部分（比如背景和噪声）之间的间隔，使得目标物体形状轮廓清晰可辨。
2. 保留了（有时甚至可能放大）目标物体内部三维形状变化的信息，使得目标物体含有的信息更加具体丰满。
3. 一定程度上忽略了噪声、背景对于识别效果的干扰。

具体的，我们将算法 1 中的 MaxGapNear 和 MaxGapFar 设为 5，ThresholdGap 设为 100，在对水壶的深度信息处理之后可以可到的的深度图像如图 4.6 所示，而不使用 DAN 得到的深度图像如图 4.5 所示。图 4.6 包含了更加细致的水壶壶身弧度的变化，所以显然包含更加详细的深度信息。

通过本小节介绍的深度信息自适应正规化方法，我们可以成功提取深度信息并获得 80% 以上的分类准确率。具体的参数及实验效果见 4.4 小节。

图 4.5　使用普通正规化得到的深度图像

图 4.6　使用 DAN 得到的深度图像

图 4.7　深度信息自适应正规化效果

4.4　试验验证

4.4.1　试验综述

华盛顿大学 RGB-D 数据集的物体由 41877 个含家用物品的 RGB-D 图像组织成，它们分别属于 51 个不同的种类和一共 300 个实例，这些图像是在三个不同视点捕捉的。我们使用间隔 5 帧的采样来评估我们的算法。我们使用和 Lai 等人 [68] 相同的十折交叉验证分割，来评估我们的方法在类别识别任务中的表现。每个分割包括大约 35000 组训练 RGB-D 数据和 7000 组测试 RGB-D 数据。在每个分割中，每一个对象类中都有一个实例是被留下做测试的。所以我们在剩下的 300-51 = 249 个实例中进行训练。在测试时任务是给每个之前没见过的实例分类。

我们使用了开源的深度学习框架——Caffe [81] 来完成我们深度学习的试验。同时，我们使用 Krizhevsky 等人 [101] 训练的 AlexNet 作为预训练模型。为了使得结果更加清晰，本章的试验采用控制变量法的方式展现。分别根据：深度学习方法（是否使用预训练的模型、是否使用深度信息自适应正规化），使用的 AlexNet 中的信息层（第六层、第七层、两层结合），以及微调（是否微调）为变量展示了相关实验结果。特征层融合使用的是 3.3.1 小节中介绍的方法，决策层融合使用的是 3.3.2 小节中介绍的方法。

4.4.2　试验参数

如先前所描述，我们使用 AlexNet 作为我们融合网络的基础。它由五个卷积层（在第一、第二和第五之后需要最大池化），其次是两个完全连接层和 SOFTMAX 分类层。除了最后一层之外所有的层都用到了线性修正单元。在从头训

练的情境中，我们将所有信息层的参数都设置成按照一种固定学习率（learning rate）的模式在改变（最开始学习率取为 0.01，经过 40K 次迭代后变为 0.001，并在 60K 次迭代后停止训练）。

在微调过程中我们使用预先训练的模型的前七层（如图 4.1 所示）的权重和偏差来初始化我们网络的前七层，丢掉了最后的 SOFTMAX 分类层，将第八层替换为一个输出维度为 51 的决策层（因为我们的数据共有 51 类）。然后我们继续使用分阶段训练。在微调的情境中，我们将模型前七层的参数设置为按照一种固定学习率（learning rate）的模式在改变（最开始学习率取为 0.001，经过 10K 次迭代后变为 0.0001，并在 20K 次迭代后停止训练），将最后一层的学习率设置为始终是前七层的十倍（即最开始学习率取为 0.01，经过 10K 次迭代后变为 0.001，并在 20K 次迭代后停止训练）。这样设置的原因是前七层由于继承了原模型的参数，已经趋于稳定，需要改变的幅度较小；而第八层由于是从头训练，需要改变的幅度较大。我们尝试了微调 RGB 和深度信息，但是最后发现微调 RGB 信息并不能带来识别效果的改进。

未经特殊说明，我们将算法 1 中的 MaxGapNear 和 MaxGapFar 设为 5，ThresholdGap 设为 100。

我们从此任务的十种数据分割方法中随机选出一种分割方法作为确认分割，而训练的迭代次数以及其他参数都是基于在确认分割上的测试结果所选定的。如果没有特殊说明，我们使用固定为 0.9 的 momentum 和固定为 128 的 mini-batch。我们直接将图片缩放至 224×224 的维度，没有使用随机水平翻转。

4.4.3 试验结果

模型及方法对深度学习结果的影响如表 4.1 所示，可以看出 Pretrained Model 可以有效提高颜色信息的分类准确率，而 Pretrained Model 加上 DAN（即将深度信息正规化为合法图片）可以有效提高深度信息的分类准确率。注意，我们并没有对颜色信息使用微调。

信息层对于深度学习结果的影响如表 4.2 所示，可以看出使用原神经网络的第六层和第七层结合可以获得比较高的准确率和稳定性，不过这些结果与 fc6 的分类结果相比相差无几。同时我们需要考虑，fc6 对于单模态信息只有 4096 维特征，如果加上 fc7 就将特征维数变为原来的两倍，而识别准确率几乎没有变化，说明 fc7 与 fc6 的信息互补性非常弱，所以在最终的综合模型中我们选择只使用

表 4.1　模型及方法对深度学习结果的影响

模态 / 方法	RGB	Depth	RGB & Depth
From scratch	80.3 ± 2.6	70.7 ± 2.3	82.8 ± 2.4
Pretrained*	**85.5 ± 1.5**	76.2 ± 2.1	87.1 ± 1.8
Pretrained & DAN**	**85.5 ± 1.5**	**81.6 ± 2.8**	**90.3 ± 1.8**

本表展示的深度信息数据均是经过微调得到的结果，颜色和深度信息数据均是采用 fc6 & fc7 两层信息得到的结果。

*：如 4.2 节所述

**：如 4.3 节所述，只对深度信息使用 DAN

fc6 层信息。注意，使用其它信息层（比如 fc8）并不会提高分类准确率。

表 4.2　信息层对于深度学习结果的影响

模态 / 信息层	RGB	Depth	RGB & Depth
fc6*	85.3 ± 1.8	81.5 ± 2.8	90.2 ± 1.7
fc7**	83.0 ± 1.9	78.8 ± 2.6	88.3 ± 2.1
fc6 & fc7	**85.5 ± 1.5**	**81.6 ± 2.8**	**90.3 ± 1.8**

本表展示的深度信息数据均是经过 DAN 得到的结果，颜色和深度信息数据均是采用 Pretrained Model 得到的结果。

*：fc6 是图 4.1 中的第六层

**：fc7 是图 4.1 中的第七层

微调对深度学习结果的影响如表 4.3 所示，可以看出使用深度信息微调原模型有利于深度信息的提取和利用。注意在我们的试验中微调并不会提高颜色信息的分类准确率。

表 4.3　微调对深度学习结果的影响

模态	Depth	RGB & Depth
不微调	78.1 ± 2.1	89.6 ± 1.7
微调	**81.6 ± 2.8**	**90.3 ± 1.8**

本表展示的深度信息数据均是经过 DAN 得到的结果。

第 5 章　综合分类模型

5.1　综合模型分类结果

综合第 3 章的多模态融合策略和第 4 章的深度学习方法，我们提出了一个用于分类的多模态融合的综合模型，整体架构如图 5.1 所示。我们将颜色和深度信息分别作为 HMP、CRNN、以及深度学习的输入，提取出特征然后在特征层将不同模态进行融合，最后在方法间使用 3.3.2 小节描述的决策层融合方法进行融合。

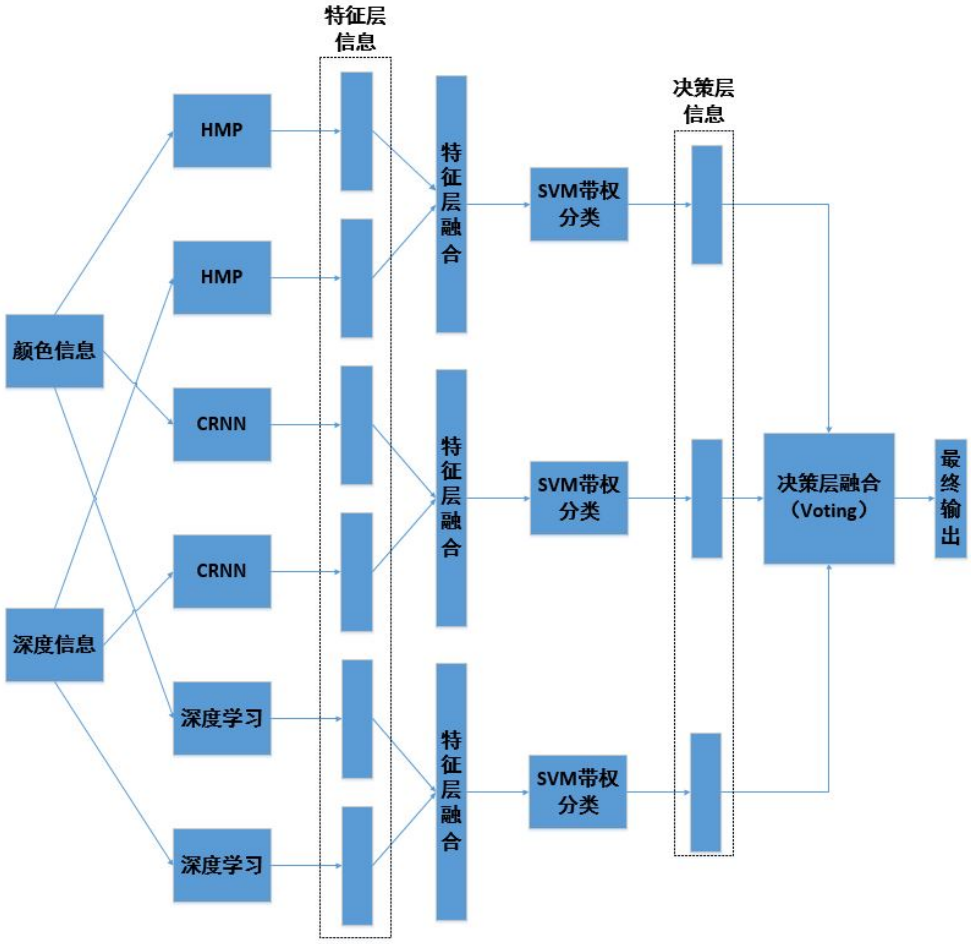


图 5.1　综合分类模型工作流

在现有的方法中，传统词袋方法的识别精度不能尽如人意。而深度学习学习虽然准确率较高，但需要大规模数据库的支持。而我们提出的综合分类模型则结合了多种方法的优点：在没有大规模数据库支持时（即表 5.1 中的 FS[1]），深度学习表现一般，但有词袋模型作为准确率的支撑；在有大规模数据库或相关大规模数据库中预训练的模型时（即表 5.1 中的 PM[2]），则深度学习识别的准确率超过传统方法，对最终分类结果可以发挥极大的贡献。最终模型的识别结果如表 5.1 所示。注意由于综合模型中是三方融合，我们还尝试使用了多数投票（Majority Voting）作为融合策略，最终识别结果为 91.3%，略低于我们的带权投票。

表 5.1　综合模型实验结果

方法 \ 模态		RGB	Depth	RGB & Depth
CRNN		81.7 ± 1.4	78.9 ± 1.6	87.6 ± 1.2
HMP		75.8 ± 3.4	78.4 ± 1.9	85.4 ± 2.6
CRNN & HMP		83.1 ± 2.6	82.5 ± 1.9	89.6 ± 2.1
深度学习	FS[1]	80.3 ± 2.6	70.7 ± 2.3	82.8 ± 2.4
	PM[2]	85.5 ± 1.5	81.6 ± 2.8	90.3 ± 1.8
综合模型	FS[1]	84.1 ± 2.3	81.9 ± 2.5	90.1 ± 1.6
	PM[2]	86.9 ± 1.3	83.8 ± 2.9	91.6 ± 1.2

本表展示的深度信息数据均是经过微调和 DAN 处理得到的结果，深度学习数据均是采用 fc6 单层信息得到的结果。且方法内均使用 3.3.1 小节描述的特征层融合方法，方法间均使用 3.3.2 小节描述的决策层融合方法。

1：不借助其他数据集从头训练（from scratch）

2：使用在其他数据集上预训练的模型（pretrained model）

各个类的分类准确率如图 5.2 所示，可见最终综合模型的百分之八十的类别分类准确率都在百分之九十以上，并且对于大部分类别，综合模型的分类准确率都高于三种方法单独的分类准确率。

不过如果我们重点看最分类准确率最低的十个类（如图 5.3 所示），可以发现，这些类别分类效果差一方面是因为数据本身问题——三种分类方法准确率都不高；另一方面也是由于分类融合策略在某些类别中被表现较差的方法“拖了后腿”，而没有发挥出表现好的类别的效果。这应该是我们下一步重点探讨的问题。

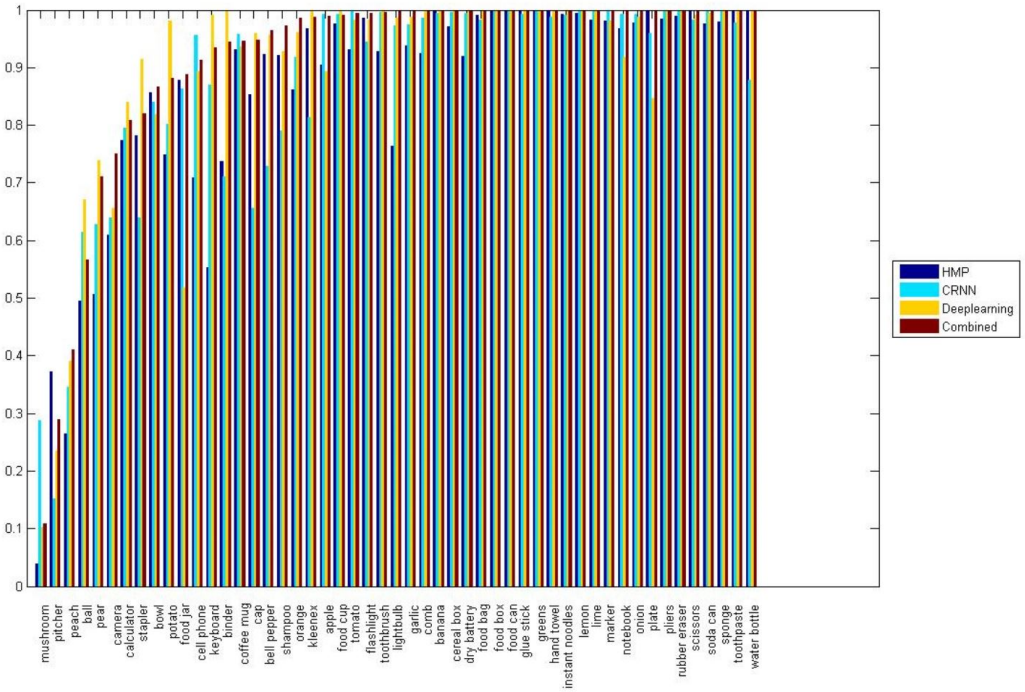


图 5.2　分类准确率一览

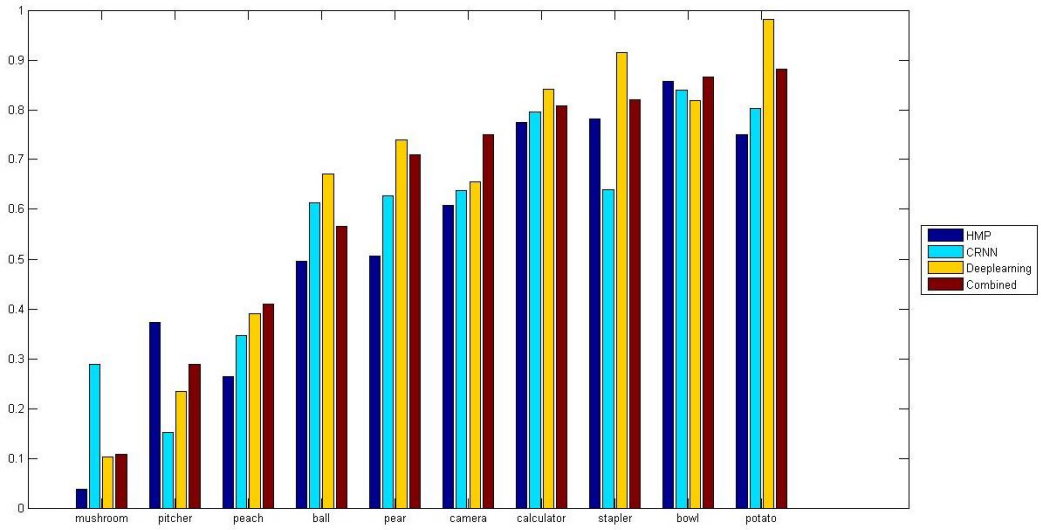


图 5.3　分类准确率最低的十个类

5.2 多模态识别程序

我们按照 5.1 节描述的综合模型实现了具有图形界面的多模态识别实例程序。

程序的流程图如 5.4 所示，程序的界面如图 5.5 所示。在程序开始时会自动载入训练好的 Model，在识别时可以输入 RGB 图片和深度图片信息，并选择使用单模态或多模态分类器进行分类。具体分类效果由模型决定，在 Qt 程序中只做简单的预处理。

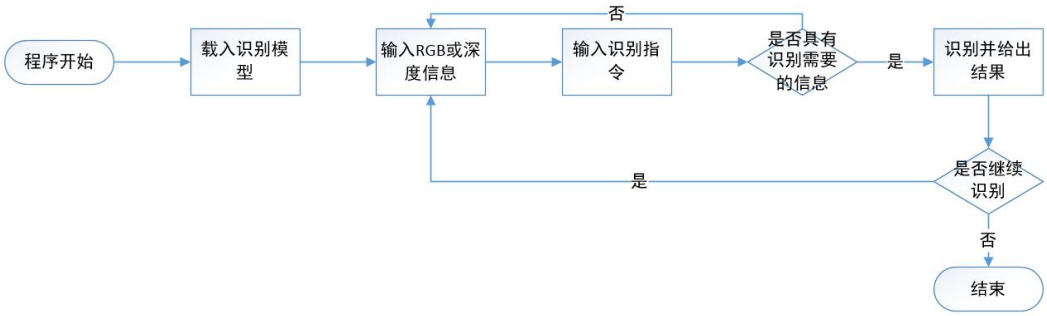


图 5.4　多模态识别程序流程图

试验平台：

```
Windows 8.1
Qt Creator 2.8.1
```

编译器：

```
Qt5.1.1 (MSVC 2010, 32 bit)
GNU Make 3.82.90
Built for i686-w64-mingw32
```

图 5.5　多模态识别程序界面

第 6 章　总结与展望

6.1　本文总结

本文研究了多方法多模态信息融合策略以及深度学习在多模态信息融合中的应用，具体的工作有：

1. 首先，我们利用传统的分层匹配追踪，以及随机权值卷积递归神经网络来进行颜色和深度信息的提取和融合，并探究不同模态的信息之间的互补性。然后我们在这两种方法的基础上提出了一种决策层的信息融合策略，并探究不同方法之间的互补性。实验表明，不仅仅不同模态的信息之间具有互补性，不同信息提取方法之间也具有很强的互补性。将不同模态的信息或不同信息提取方法进行融合都可以使得我们对目标物体特征的感知变得更加全面，可以获得更好的识别效果。大量已有工作都是仅限于利用多传感器或不同模态之间的融合，而我们引入的不同方法之间融合的思路，可以进一步利用数据的互补性，弥补各个方法的不足。
2. 其次，本文在传统信息提取方法之外还引入了深度学习方法。即使用八层的大型卷积神经网络在颜色信息和深度信息中分别进行深度学习，然后将相应的特征层信息提取出来，并进行模态之间的特征层信息融合，最终给出识别结果。此外，本文还提出了一种将深度信息转换为图片信息的正规化方法。这种方法对于深度信息的不同部分采用不同的正规化尺度，从而实现了既保留目标物体与背景之间的间隔，同时又保留目标物体之内形状变化信息的目的。实验证明，所提的深度信息正规化方法可以提升深度学习的识别准确率。
3. 同时，为了解决深度学习中数据量不足的问题，我们使用了预先在海量图片数据集中训练过的模型来进行信息提取。而且由于已有模型是在颜色信息数据集中训练的，我们使用已有深度信息对其进行了微调，使它可以更好地适应深度信息的提取。试验证明，使用 RGB 图片预先训练的深度学习模型可以较好地识别颜色和深度信息。最后，我们基于以上研究提出了综合分层匹配追踪、随机权值递归神经网络以及深度学习三种方法的信息融合识别模型。实验证明，我们的融合模型可以提升识别的准确率和稳定

性，并且降低了深度学习对于数据量的依赖。

6.2 未来工作展望

本文研究了多方法多模态信息融合策略，对不同层次的多种信息融合策略进行了较深入的探讨和实验，提高了分类的准确率和稳定性。不过仍有一些可以进一步完善的方向。具体有：

1. 进一步探究多模态信息融合策略。在本文的探讨与实验中，我们完善了多模态信息融合方面的策略与方法。不过在实验中我们发现，多模态信息融合的潜力还有待开发，应用空间还很广阔。因此，进一步开发多模态信息融合的潜力，以便在各类机器人应用中发挥更大的效果是颇有意义的。
2. 探究深度学习在其他模态识别等任务中的应用。我们知道传统深度学习需要的数据量是非常可观的，因而只在图像识别、分类、分割等应用中获得过较好的结果。但在本文中我们成功的将深度学习应用到了深度信息当中，尽管我们使用深度信息的数据规模并不是很大。这就为我们将深度学习的方法和模型应用到其他模态和领域开辟了道路。在本文中我们将只是将深度信息正规化为图片信息就可以使用预训练的模型取得较好的分类结果，下一步我们可以将其他模态的信息也编码为图片信息，并使用预训练的模型进行微调和分类。
3. 探究更多模态的融合。本文只是将颜色信息和深度信息结合了起来，就获得了识别准确率和稳定性上的显著提升。然而现实中的信息模态远远不止这两种，而且有些模态与 RGB-D 的信息之间具有天然的更大的互补性。例如触觉信息和滑觉信息可以告诉我们目标物体的材质力学信息和摩擦力信息等等。而现实中的机器人操作往往是在错综复杂的环境中进行的，在机器人的感知和操作中考虑更多模态的融合，将会使机器人获得对环境的更加全面的认知，并作出更加智能的决策。

插图索引

表格索引

公式索引

参考文献

[1] Lovchik C, Diftler M A. The robonaut hand: A dexterous robot hand for space. Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on, volume 2. IEEE, 1999. 907–912

[2] Hirzinger G, Brunner B, Dietrich J, et al. Rotex-the first remotely controlled robot in space. Robotics and Automation, 1994. Proceedings., 1994 IEEE International Conference on. IEEE, 1994. 2604–2611

[3] 黄晓瑞, 崔平远, 崔祜涛. 多传感器信息融合技术及其在组合导航系统中的应用. 高技术通讯, 2002, 12(2):107–110

[4] Ogorodnikova O. Safe and reliable human-robot interaction in manufactory, within and beyond the workcell. Robotics in Alpe-Adria-Danube Region (RAAD), 2010 IEEE 19th International Workshop on. IEEE, 2010. 65–70

[5] Burke J L, Murphy R R, Coovert M D, et al. Moonlight in miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. Human–Computer Interaction, 2004, 19(1-2):85–116

[6] Balakirsky S, Carpin S, Kleiner A, et al. Towards heterogeneous robot teams for disaster mitigation: Results and performance metrics from robocup rescue. Journal of Field Robotics, 2007, 24(11-12):943–967

[7] Barnes M, Jentsch F. Human-robot interactions in future military operations. Ashgate Publishing Company, 2010

[8] Pierrot F, Dombre E, Dégoulange E, et al. Hippocrate: a safe robot arm for medical applications with force feedback. Medical Image Analysis, 1999, 3(3):285–300

[9] Lavallee S, Troccaz J, Gaborit L, et al. Image guided operating robot: a clinical application in stereotactic neurosurgery. Robotics and Automation, 1992. Proceedings., 1992 IEEE International Conference on. IEEE, 1992. 618–624

[10] Mahler R P. Statistical multisource-multitarget information fusion. Artech House, Inc., 2007

[11] Akyildiz I F, Su W, Sankarasubramaniam Y, et al. A survey on sensor networks. Communications magazine, IEEE, 2002, 40(8):102–114

[12] Koushanfar F, Slijepcevic S, Potkonjak M, et al. Error-tolerant multimodal sensor fusion. IEEE CAS Workshop on Wireless Communication and Networking, 2002. 5–6

[13] Futagawa M, Iwasaki T, Ishida M, et al. A real-time monitoring system using a multimodal sensor with an electrical conductivity sensor and a temperature sensor for cow health control. Japanese Journal of Applied Physics, 2010, 49(4S):04DL12

[14] Akyildiz I F, Su W, Sankarasubramaniam Y, et al. Wireless sensor networks: a survey. Computer networks, 2002, 38(4):393–422

[15] Choi W, Pantofaru C, Savarese S. Detecting and tracking people using an rgb-d camera via multiple detector fusion. Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011. 1076–1083

[16] Guyon I, Elisseeff A. An introduction to feature extraction. Feature extraction. Springer, 2006: 1–25

[17] Guyon I, Gunn S, Nikravesh M, et al. Feature extraction: foundations and applications, volume 207. Springer, 2008

[18] Weigel V B. Deep Learning for a Digital Age: Technology's Untapped Potential To Enrich Higher Education. ERIC, 2002

[19] Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems, 2009. 1096–1104

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

[21] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification. Advances in Neural Information Processing Systems, 2014. 1988–1996

[22] Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from rgb-d images for object detection and segmentation. Computer Vision–ECCV 2014. Springer, 2014: 345–360

[23] Spinello L, Arras K O. Leveraging rgb-d data: Adaptive fusion and domain adaptation for object detection. Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, 2012. 4469–4474

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[25] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. Computer vision–ECCV 2014. Springer, 2014: 818–833

[26] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[27] 王欣. 多传感器数据融合问题的研究 [d][D]. 长春: 吉林大学, 2006

[28] Hall D L, Llinas J. An introduction to multisensor data fusion. Proceedings of the IEEE, 1997, 85(1):6–23

[29] 何友, 彭应宁. 多传感器数据融合算法综述. 火力与指挥控制, 1996, 21(1):12–21

[30] Luo R C, Su K L. A review of high-level multisensor fusion: approaches and applications. Multisensor Fusion and Integration for Intelligent Systems, 1999. MFI'99. Proceedings. 1999 IEEE/SICE/RSJ International Conference on. IEEE, 1999. 25–31

[31] Neal D R, Copland J, Neal D A. Shack-hartmann wavefront sensor precision and accuracy. International Symposium on Optical Science and Technology. International Society for Optics and Photonics, 2002. 148–160

[32] Waltz E, Llinas J, et al. Multisensor data fusion, volume 685. Artech house Boston, 1990

[33] Shekhar S, Khatib O, Shimojo M. Sensor fusion and object localization. Robotics and Automation. Proceedings. 1986 IEEE International Conference on, volume 3. IEEE, 1986. 1623–1628

[34] Luo R C, Kay M G. Multisensor integration and fusion in intelligent systems. Systems, Man and Cybernetics, IEEE Transactions on, 1989, 19(5):901–931

[35] Luo R C, Kay M G. Multisensor integration and fusion for intelligent machines and systems. Intellect Books, 1995

[36] Luo R C, Yih C C, Su K L. Multisensor fusion and integration: approaches, applications, and future research directions. Sensors Journal, IEEE, 2002, 2(2):107–119

[37] Dai J, Au O C, Fang L, et al. Multichannel nonlocal means fusion for color image denoising. Circuits and Systems for Video Technology, IEEE Transactions on, 2013, 23(11):1873–1886

[38] Khaleghi B, Khamis A, Karray F O, et al. Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 2013, 14(1):28–44

[39] 范新南, 苏丽媛, 郭建甲. 多传感器信息融合综述. 河海大学常州分校学报, 2005, 19(1):1–4

[40] Waske B, Benediktsson J A. Fusion of support vector machines for classification of multisensor data. Geoscience and Remote Sensing, IEEE Transactions on, 2007, 45(12):3858–3866

[41] Welch G, Bishop G. An introduction to the kalman filter. university of north carolina at chapel hill, department of computer science. Technical report, TR 95-041, 2004

[42] Atry P, Hossain M, Saddik A, et al. Multimodal fusion for multimedia analysis. Multimedia systems, 2010, 16(6):345–379

[43] 易正俊. 多源信息智能融合算法. 重庆大学, 2002.

[44] Pawlak Z. Rough sets: theoretical aspects of reasoning about data, system theory, knowledge engineering and problem solving, vol. 9, 1991

[45] Goodman I R, Mahler R P, Nguyen H T. Mathematics of data fusion, volume 37. Springer Science & Business Media, 2013

[46] Stone L D, Streit R L, Corwin T L, et al. Bayesian multiple target tracking. Artech House, 2013

[47] Johansson F, Falkman G. A bayesian network approach to threat evaluation with application to an air defense scenario. Information Fusion, 2008 11th International Conference on. IEEE, 2008. 1–7

[48] Li C, Heinemann P, Sherry R. Neural network and bayesian network fusion models to fuse electronic nose and surface acoustic wave sensor data for apple defect detection. Sensors and Actuators B: Chemical, 2007, 125(1):301–310

[49] Wong Y C, Sundareshan M K. Data fusion and tracking of complex target maneuvers with a simplex-trained neural network-based architecture. Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, volume 2. IEEE, 1998. 1024–1029

[50] Dornfeld D A, DeVries M. Neural network sensor fusion for tool condition monitoring. CIRP Annals-Manufacturing Technology, 1990, 39(1):101–105

[51] Manduchi R. Bayesian fusion of color and texture segmentations. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 2. IEEE, 1999. 956–962

[52] Boutell M, Luo J. Bayesian fusion of camera metadata cues in semantic scene classification. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2. IEEE, 2004. II–623

[53] Senoo T, Yamakawa Y, Mizusawa S, et al. Skillful manipulation based on high-speed sensory-motor fusion. Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. IEEE, 2009. 1611–1612

[54] Fay D A, Waxman A M, Aguilar M, et al. Fusion of multi-sensor imagery for night vision: color visualization, target learning and search. Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on, volume 1. IEEE, 2000. TUD3–3

[55] Han J, Bhanu B. Statistical feature fusion for gait-based human recognition. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2. IEEE, 2004. II–842

[56] Zhang Y, Zhang H, Nasrabadi N M, et al. Multi-metric learning for multi-sensor fusion based classification. Information Fusion, 2013, 14(4):431–440

[57] 高方伟. 多传感器融合的技术研究 [D]. 西安电子科技大学, 2007

[58] 郝润泽, 杨瑞朋. 多传感器数据融合技术研究现状及军事应用. 兵工自动化, 2007, 26(4):16–17

[59] 赵丹丹. 多传感器数据融合在目标识别中的应用研究 [d][D]. 太原: 太原理工大学, 2007

[60] Bishop C M. Pattern recognition. Machine Learning, 2006.

[61] Filliat D. A visual bag of words method for interactive qualitative localization and mapping. Robotics and Automation, 2007 IEEE International Conference on. IEEE, 2007. 3921–3926

[62] Cortes C, Vapnik V. Support vector machine. Machine learning, 1995, 20(3):273–297

[63] Wang Q, Garrity G M, Tiedje J M, et al. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and environmental microbiology, 2007, 73(16):5261–5267

[64] Haykin S, Network N. A comprehensive foundation. Neural Networks, 2004, 2(2004)

[65] Hagan M T, Demuth H B, Beale M H, et al. Neural network design, volume 20. PWS publishing company Boston, 1996

[66] Jaitly N, Nguyen P, Senior A W, et al. Application of pretrained deep neural networks to large vocabulary speech recognition. INTERSPEECH, 2012. 2578–2581

[67] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. Proceedings of the 28th international conference on machine learning (ICML-11), 2011. 689–696

[68] Lai K, Bo L, Ren X, et al. A large-scale hierarchical multi-view rgb-d object dataset. Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011. 1817–1824

[69] Lai K, Bo L, Ren X, et al. Sparse distance learning for object recognition combining rgb and depth information. Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011. 4007–4013

[70] Bo L, Ren X, Fox D. Depth kernel descriptors for object recognition. Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. IEEE, 2011. 821–826

[71] Blum M, Springenberg J T, Wülfing J, et al. A learned feature descriptor for object recognition in rgb-d data. Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, 2012. 1298–1303

[72] Bo L, Ren X, Fox D. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. Advances in neural information processing systems, 2011. 2115–2123

[73] Bo L, Ren X, Fox D. Unsupervised feature learning for rgb-d based object recognition. Experimental Robotics. Springer, 2013. 387–402

[74] Socher R, Huval B, Bath B, et al. Convolutional-recursive deep learning for 3d object classification. Advances in Neural Information Processing Systems, 2012. 665–673

[75] Cheng Y, Zhao X, Huang K, et al. Semi-supervised learning for rgb-d object recognition. 2014 22nd International Conference on Pattern Recognition (ICPR). IEEE, 2014. 2377–2382

[76] Asif U, Bennamoun M, Sohel F. Efficient rgb-d object categorization using cascaded ensembles of randomized decision trees. Robotics and Automation (ICRA), 2015 IEEE International Conference on. IEEE, 2015. 1295–1302

[77] Schwarz M, Schulz H, Behnke S. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. Robotics and Automation (ICRA), 2015 IEEE International Conference on. IEEE, 2015. 1329–1335

[78] Eitel A, Springenberg J T, Spinello L, et al. Multimodal deep learning for robust rgb-d object recognition. Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015. 681–687

[79] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009. 1794–1801

[80] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009. 248–255

[81] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

[82] Shannon C E. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1):3–55

[83] Shannon C E, Weaver W. The mathematical theory of information. 1949.

[84] Wiener N. Cybernetics or Control and Communication in the Animal and the Machine, volume 25. MIT press, 1961

[85] Raman T. Multimodal information presentation system, May 5, 1998. US Patent 5,748,186

[86] Mansoorizadeh M, Charkari N M. Multimodal information fusion application to human emotion recognition from face and speech. Multimedia Tools and Applications, 2010, 49(2):277–297

[87] White F E. Data fusion lexicon. Technical report, DTIC Document, 1991

[88] Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey. Multimedia systems, 2010, 16(6):345–379

[89] Shiffrin R M, Schneider W. Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. Psychological review, 1977, 84(2):127

[90] Xiong N, Svensson P. Multi-sensor management for information fusion: issues and approaches. Information fusion, 2002, 3(2):163–186

[91] Caudill M, Butler C. Understanding neural networks: computer explorations: a workbook in two volumes with software for the macintosh and pc compatibles. MIT press, 1994

[92] Prabhakar S, Jain A K. Decision-level fusion in fingerprint verification. Pattern Recognition, 2002, 35(4):861–874

[93] Chatzis V, Borş A G, Pitas I. Multimodal decision-level fusion for person authentication. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 1999, 29(6):674–680

[94] Clancey W J. The epistemology of a rule-based expert system—a framework for explanation. Artificial intelligence, 1983, 20(3):215–251

[95] Ng P C, Henikoff S. Sift: Predicting amino acid changes that affect protein function. Nucleic acids research, 2003, 31(13):3812–3814

[96] Aharon M, Elad M, Bruckstein A. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. Signal Processing, IEEE Transactions on, 2006, 54(11):4311–4322

[97] Goldberg D E, Holland J H. Genetic algorithms and machine learning. Machine learning, 1988, 3(2):95–99

[98] Metropolis N, Ulam S. The monte carlo method. Journal of the American statistical association, 1949, 44(247):335–341

[99] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 2008, 9:1871–1874

[100] Townsend J T. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 1971, 9(1):40–50

[101] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012. 1097–1105

[102] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 2015, 115(3):211–252

致　谢

衷心感谢导师李洪波老师对本人的精心指导。他不仅是我的毕设指导老师，还是我的班主任和任课老师，在我的整个大学期间对我的学业，学术，乃至日常生活都起了极大的帮助作用，他平易近人的作风，关心学生的热情以及对待工作的忱将使我终生受益。

感谢清华大学智能技术与系统国家重点实验室，以及实验室全体老师和同学们在我毕设工作中的热情帮助和支持。

感谢清华大学计 25 班的各位大神，你们的存在使我的本科轻松快乐了许多。

感谢 ThuThesis，它让我可以专心于论文的内容而非格式，使我的毕设论文工作轻松了许多。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

附录 A　外文资料原文

Multimodal Deep Learning for Robust RGB-D Object Recognition

Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, Wolfram Burgard

***Abstract*— Robust object recognition is a crucial ingredient of many, if not all, real-world robotics applications. This paper leverages recent progress on Convolutional Neural Networks (CNNs) and proposes a novel RGB-D architecture for object recognition. Our architecture is composed of two separate CNN processing streams – one for each modality – which are consecutively combined with a late fusion network. We focus on learning with imperfect sensor data, a typical problem in real-world robotics tasks. For accurate learning, we introduce a multi-stage training methodology and two crucial ingredients for handling depth data with CNNs. The first, an effective encoding of depth information for CNNs that enables learning without the need for large depth datasets. The second, a data augmentation scheme for robust learning with depth images by corrupting them with realistic noise patterns. We present state-of-the-art results on the RGB-D object dataset [15] and show recognition in challenging RGB-D real-world noisy settings.**

I. Introduction

RGB-D object recognition is a challenging task that is at the core of many applications in robotics, indoor and outdoor. Nowadays, RGB-D sensors are ubiquitous in many robotic systems. They are inexpensive, widely supported by open source software, do not require complicated hardware and provide unique sensing capabilities. Compared to RGB data, which provides information about appearance and texture, depth data contains additional information about object shape and it is invariant to lighting or color variations.

In this paper, we propose a new method for object recognition from RGB-D data. In particular, we focus on making recognition robust to imperfect sensor data. A scenario typical for many robotics tasks. Our approach builds on recent advances from the machine learning and computer vision community. Specifically, we extend classical convolutional neural network networks (CNNs), which have recently been shown to be remarkably successful for recognition on RGB images [13], to the domain of RGB-D data. Our architecture, which is depicted in Fig. 1, consists of two convolutional network streams operating on color and depth information respectively. The network automatically learns to combine these two processing streams in a late fusion approach. This architecture bears similarity to other recent multi-stream approaches [21], [23], [11]. Training of the individual stream networks as well as the combined architecture follows a stage-wise approach. We start by separately training the networks for each modality, followed by a third training stage in which the two streams are jointly fine-tuned, together with a fusion network that performs the final classification. We initialize both the RGB and depth stream network with weights from a network pre-trained on the ImageNet dataset [19]. While initializing an RGB network from a pre-trained ImageNet network is straight-forward, using such a network for processing depth data is not. Ideally, one would want to directly train a network for recognition from depth data without pre-training on a different modality which, however, is infeasible due to lack of large scale labeled depth datasets. Due to this lack of labeled training data, a pre-training phase for the depth-modality – leveraging RGB data – becomes of key importance. We therefore propose a depth data encoding to enable re-use of CNNs trained on ImageNet for recognition from depth data. The intuition – proved experimentally – is to simply encode a depth image as a rendered RGB image, spreading the information contained in the depth data over all three RGB channels and then using a standard (pre-trained) CNN for recongition.

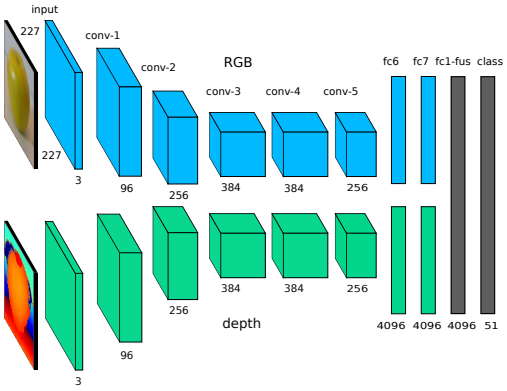


Fig. 1: Two-stream convolutional neural network for RGB-D object recognition. The input of the network is an RGB and depth image pair of size 227 × 227 × 3. Each stream (blue, green) consists of five convolutional layers and two fully connected layers. Both streams converge in one fully connected layer and a softmax classifier (gray).

In real-world environments, objects are often subject to occlusions and sensor noise. In this paper, we propose a data augmentation technique for depth data that can be used for robust training. We augment the available training examples by corrupting the depth data with missing data patterns sampled from real-world environments. Using these two techniques, our system can both learn robust depth features and implicitly weight the importance of the two modalities.

All authors are with the Department of Computer Science, University of Freiburg, Germany. This work was partially funded by the DFG under the priority programm "Autonomous Learning" (SPP 1527). {eitel, springj, riedmiller, spinello, burgard}@cs.uni-freiburg.de

We tested our method to support our claims: first, we report on RGB-D recognition accuracy, then on robustness with respect to real-world noise. For the first, we show that our work outperforms the current state of the art on the RGB-D Object dataset of Lai *et al.* [15]. For the second, we show that our data augmentation approach improves object recognition accuracy in a challenging real-world and noisy environment using the RGB-D Scenes dataset [16].

II. Related Work

Our approach is related to a large body of work on both convolutional neural networks (CNNs) for object recognition as well as applications of computer vision techniques to the problem of recognition from RGB-D data. Although a comprehensive review of the literature on CNNs and object recognition is out of the scope of this paper, we will briefly highlight connections and differences between our approach and existing work with a focus on recent literature.

Among the many successful algorithms for RGB-D object recognition a large portion still relies on hand designed features such as SIFT in combination with multiple shape features on the depth channel [15], [14]. However, following their success in many computer vision problems, unsupervised feature learning methods have recently been extended to RGB-D recognition settings. Blum *et al.* [3] proposed an RGB-D descriptor that relies on a K-Means based feature learning approach. More recently Bo *et al.* [5] proposed hierarchical matching pursuit (HMP), a hierarchical sparse-coding method that can learn features from multiple channel input. A different approach pursued by Socher *et al.* [22] relies on combining convolutional filters with a recursive neural network (a specialized form of recurrent neural network) as the recognition architecture. Asif *et al.* [1] report improved recognition performance using a cascade of Random Forest classifiers that are fused in a hierarchical manner. Finally, in recent independent work Schwarz *et al.* [20] proposed to use features extracted from CNNs pre-trained on ImageNet for RGB-D object recognition. While they also make use of a two-stream network they do not fine-tune the CNN for RGB-D recognition, but rather just use the pre-trained network as is. Interestingly, they also discovered that simple colorization methods for depth are competitive to more involved preprocessing techniques. In contrast to their work, ours achieves higher accuracy by training our fusion CNN end-to-end: mapping from raw pixels to object classes in a supervised manner (with pre-training on a related recognition task). The features learned in our CNN are therefore by construction discriminative for the task at hand. Using CNNs trained for object recognition has a long history in computer vision and machine learning. While they have been known to yield good results on supervised image classification tasks such as MNIST for a long time [17], recently they were not only shown to outperform classical methods in large scale image classification tasks [13], object detection [9] and semantic segmentation [8] but also to produce features that transfer between tasks [7], [2]. This recent success story has been made possible through optimized implementations for high-performance computing systems, as well as the availability of large amounts of labeled image data through, e.g., the ImageNet dataset [19].

While the majority of work in deep learning has focused on 2D images, recent research has also been directed towards using depth information for improving scene labeling and object detection [6], [10]. Among them, the work most similar to ours is the one on object detection by Gupta *et al.* [10] who introduces a generalized method of the R-CNN detector [9] that can be applied to depth data. Specifically, they use large CNNs already trained on RGB images to also extract features from depth data, encoding depth information into three channels (HHA encoding). Specifically, they encode for each pixel the height above ground, the horizontal disparity and the pixelwise angle between a surface normal and the gravity direction. Our fusion network architecture shares similarities with their work in the usage of pre-trained networks on RGB images. Our method differs in both the encoding of depth into color image data and in the fusion approach taken to combine information from both modalities. For the encoding step, we propose an encoding method for depth images ('colorizing' depth) that does not rely on complicated preprocessing and results in improved performance when compared to the HHA encoding. To accomplish sensor fusion we introduce additional layers to our CNN pipeline (see Fig. 1) allowing us to automatically learn a fusion strategy for the recognition task – in contrast to simply training a linear classifier on top of features extracted from both modalities. Multi-stream architectures have also been used for tasks such as action recognition [21], detection [11] and image retrieval [23]. An interesting recent overview of different network architectures for fusing depth and image information is given in Saxena *et al.* [18]. There, the authors compared different models for multimodal learning: (1) early fusion, in which the input image is concatenated to the existing image RGB channels and processed alongside; (2) an approach we denote as late fusion, where features are trained separately for each modality and then merged at higher layers; (3) combining early and late fusion; concluding that late fusion (2) and the combined approach perform best for the problem of grasp detection. Compared to their work, our model is similar to the late fusion approach but widely differs in training – Saxena *et al.* [18] use a layer-wise unsupervised training approach – and scale (the size of both their networks and input images is an order of magnitude smaller than in our settings).

III. Multimodal architecture for RGB-D object recognition

An overview of the architecture is given in Fig. 1. Our network consists of two streams (top-blue and bottom-green part in the figure) – processing RGB and depth data independently – which are combined in a late fusion approach. Each stream consists of a deep CNN that has been pre-trained for object classification on the ImageNet database (we use the CaffeNet [12] implementation of the CNN from Krizhevsky *et al.* [13]). The key reason behind starting from

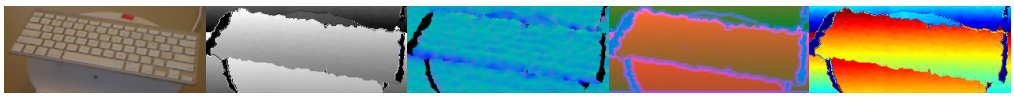


Fig. 2: Different approaches for color encoding of depth images. From left to right: RGB, depth-gray, surface normals [5], HHA [10], our method.

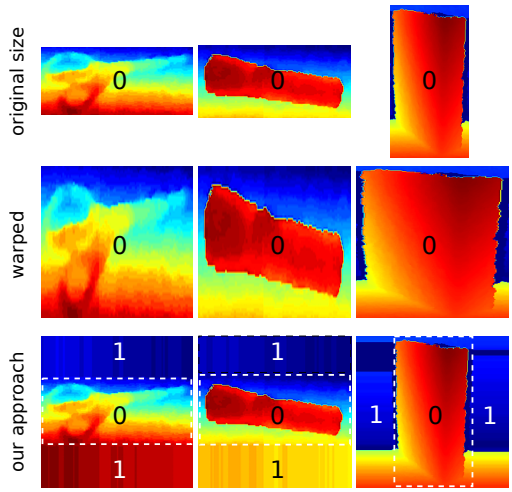


Fig. 3: CNNs require a fixed size input. Instead of the widely used image warping approach (middle), our method (bottom) preserves shape information and ratio of the objects. We rescale the longer side and create additional image context, by tiling the pixels at the border of the longer side, e.g., 1. We assume that the depth image is already transformed to three channels using our colorization method.

a pre-trained network is to enable training a large CNN with millions of parameters using the limited training data available from the Washington RGB-D Object dataset (see, e.g., Yosinski *et al.* [25] for a recent discussion). We first pre-process data from both modalities to fully leverage the ImageNet pre-training. Then, we train our multimodal CNN in a stage-wise manner. We fine-tune the parameters of each individual stream network for classification of the target data and proceed with the final training stage in which we jointly train the parameters of the fusion network. The different steps will be outlined in the following sections.

A. Input preprocessing

To fully leverage the power of CNNs pre-trained on ImageNet, we pre-process the RGB and depth input data such that it is compatible with the kind of original ImageNet input. Specifically, we use the reference implementation of the CaffeNet [12] that expects 227 $\times$ 227 pixel RGB images as input which are typically randomly cropped from larger 256 $\times$ 256 RGB images (see implementation details on data augmentation). The first processing step consists of scaling the images to the appropriate image size. The simplest approach to achieve this is to use image warping by directly rescaling the original image to the required image dimensions, disregarding the original object ratio. This is depicted in Fig. 3 (middle). We found in our experiments that this process is detrimental to object recognition performance – an effect that we attribute to a loss of shape information (see also Section IV-C). We therefore devise a different preprocessing approach: we scale the longest side of the original image to 256 pixels, resulting in a 256 $\times$ N or an N $\times$ 256 sized image. We then tile the borders of the longest side along the axis of the shorter side. The resulting RGB or depth image shows an artificial context around the object borders (see Fig. 3). The same scaling operation is applied to both RGB and depth images.

While the RGB images can be directly used as inputs for the CNNs after this processing step, the rescaled depth data requires additional steps. To realize this, recall that a network trained on ImageNet has been trained to recognize objects in images that follow a specific input distribution (that of natural camera images) that is incompatible with data coming from a depth sensor – which essentially encodes distance of objects from the sensor. Nonetheless, by looking at a typical depth image from a household object scene (c.f., Fig. 4) one can conclude that many features that qualitatively appear in RGB images – such as edges, corners, shaded regions – are also visible in, e.g., a grayscale rendering of depth data. This realization has previously led to the idea of simply using a rendered version of the recorded depth data as an input for CNNs trained on ImageNet [10]. We compare different such encoding strategies for rendering depth to images in our experiments. The two most prevalent such encodings are (1) rendering of depth data into grayscale and replicating the grayscale values to the three channels required as network input; (2) using surface normals where each dimension of a normal vector corresponds to one channel in the resulting image. A more involved method, called HHA encoding [10], encodes in the three channels the height above ground, horizontal disparity and the pixelwise angle between a surface normal and the gravity direction.

We propose a fourth, effective and computationally inexpensive, encoding of depth to color images, which we found to outperform the HHA encoding for object recognition. Our method first normalizes all depth values to lie between 0 and 255. Then, we apply a jet colormap on the given image that transforms the input from a single to a three channel image (colorizing the depth). For each pixel (i, j) in the depth image d of size $W \times H$, we map the distance to color values ranging from red (near) over green to blue (far), essen-

tially distributing the depth information over all three RGB channels. Edges in these three channels often correspond to interesting object boundaries. Since the network is designed for RGB images, the colorization procedure provides enough common structure between a depth and an RGB image to learn suitable feature representations (see Fig. 2 for a comparison between different depth preprocessing methods).

B. Network training

Let $\mathcal{D} = \{(\mathbf{x}^1, \mathbf{d}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^N, \mathbf{d}^N, \mathbf{y}^N)\}$ be the labeled data available for training our multimodal CNN; with $\mathbf{x}^i, \mathbf{d}^i$ denoting the RGB and pre-processed depth image respectively and $\mathbf{y}^i$ corresponding to the image label in one-hot encoding – i.e., $\mathbf{y}^i \in \mathbb{R}^M$ is a vector of dimensionality M (the number of labels) with $y^i_k = 1$ for the position k denoting the image label. We train our model using a three-stage approach, first training the two stream networks individually followed by a joint fine-tuning stage.

1) Training the stream networks: We first proceed by training the two individual stream networks (c.f., the blue and green streams in Fig. 1). Let $g^I(\mathbf{x}^i; \theta^I)$ be the representation extracted from the last fully connected layer (fc7) of the CaffeNet – with parameters θ^I – when applied to an RGB image $\mathbf{x}^i$. Analogously, let $g^D(\mathbf{d}^i; \theta^D)$ be the representation for the depth image. We will assume that all parameters θ^I and θ^D (the network weights and biases) are initialized by copying the parameters of a CaffeNet trained on the ImageNet dataset. We can then train an individual stream network by placing a randomly initialized softmax classification layer on top of f^D and f^I and minimizing the negative log likelihood $\mathcal{L}$ of the training data. That is, for the depth image stream network we solve

$$\min_{\mathbf{W}^D, \theta^D} \sum_{i=1}^{N} \mathcal{L}\left(\text{softmax}\left(\mathbf{W}^D g^D(\mathbf{d}^i; \theta^D)\right), y^i\right), \quad (1)$$

where W^D are the weights of the softmax layer mapping from $g(\cdot)$ to $\mathbb{R}^M$, the softmax function is given by $\text{softmax}(\mathbf{z}) = \exp(\mathbf{z})/\|\mathbf{z}\|_1$ and the loss is computed as $\mathcal{L}(s, y) = -\sum_k y_k \log s_k$. Training the RGB stream network then can be performed by an analogous optimization. After training, the resulting networks can be used to perform separate classification of each modality.

2) Training the fusion network: Once the two individual stream networks are trained we discard their softmax weights, concatenate their – now fine-tuned – last layer responses $g^I(\mathbf{x}^i; \theta^I)$ and $g^D(\mathbf{d}^i; \theta^D)$ and feed them through an additional fusion stream $f([g^I(\mathbf{x}^i; \theta^I), g^D(\mathbf{d}^i; \theta^D)]; \theta^F)$ with parameters θ^F. This fusion network again ends in a softmax classification layer. The complete setup is depicted in Fig. 1, where the two fc7 layers (blue and green) are concatenated and merge into the fusion network (here the inner product layer fc1-fus depicted in gray). Analogous to Eq. (1) the fusion network can therefore be trained by jointly optimizing all parameters to minimize the negative log likelihood

$$\min_{\mathbf{W}^f, \theta^I, \theta^D, \theta^F} \sum_{i=1}^{N} \mathcal{L}\left(\text{softmax}\left(\mathbf{W}^f f([\mathbf{g}^I, \mathbf{g}^D]; \theta^F)\right), y^i\right), \quad (2)$$

where $\mathbf{g}^I = g^I(\mathbf{x}^i; \theta^I)$, $\mathbf{g}^D = g^D(\mathbf{d}^i; \theta^D)$. Note that in this stage training can also be performed by optimizing only the weights of the fusion network (effectively keeping the weights from the individual stream training intact).

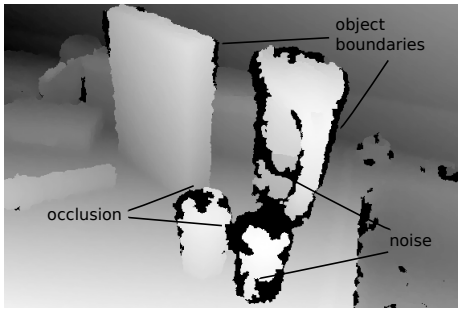


Fig. 4: Kitchen scene in the RGB-D Scenes dataset showing objects subjected to noise and occlusions.

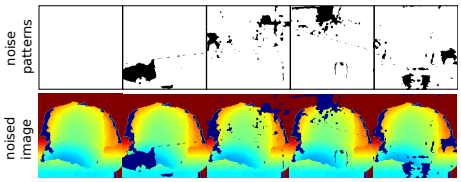


Fig. 5: We create synthetic training data by inducing artificial patterns of missing depth information in the encoded image.

C. Robust classification from depth images

Finally, we are interested in using our approach in real world robotics scenarios. Robots are supposed to perform object recognition in cluttered scenes where the perceived sensor data is subject to changing external conditions (such as lighting) and sensor noise. Depth sensors are especially affected by a non-negligible amount of noise in such setups. This is mainly due to the fact that reflective properties of materials as well as their coating, often result in missing depth information. An example of noisy depth data is depicted in Fig. 4. In contrast to the relatively clean training data from the Washington RGB-D Object dataset, the depicted scene contains considerable amounts of missing depth values and partial occlusions (the black pixels in the figure). To achieve robustness against such unpredictable factors, we propose a new data augmentation scheme that generates new, noised training examples for training and is tailored specifically to robust classification from depth data.

Our approach utilizes the observation that noise in depth data often shows a characteristic pattern and appears at object boundaries or object surfaces. Concretely, we sampled

a representative set of noise patterns $\mathcal{P} = \{P_1, \ldots, P_K\}$ that occur when recording typical indoor scenes through a Kinect sensor. For sampling the noise patterns we used the RGB-D SLAM dataset [24]. First, we extract 33,000 random noise patches of size 256 $\times$ 256 from different sequences at varying positions and divide them into five groups, based on the number of missing depth readings they contain. Those noise patches are 2D binary masks patterns. We randomly sample pairs of noise patches from two different groups that are randomly added or subtracted and optionally inverted to produce a final noise mask pattern. We repeat this process until we have collected $K = 50,000$ noise patterns in total. Examples of the resulting noise patterns and their application to training examples are shown in Fig. 5.

Training the depth network with artificial noise patterns then proceeds by minimizing the objective from Equation Eq. (1) in which each depth sample $\mathbf{d}^i$ is randomly replaced with a noised variant with probability 50%. Formally,

$$\mathbf{d}^i = \begin{cases} \mathbf{d}^i & \text{if } p = 1 \\ P_k \circ \mathbf{d}^i & \text{else} \end{cases} \text{ with } \begin{array}{l} p \sim \mathcal{B}\{0.5\} \\ k \sim \mathcal{U}\{1, K\}, \end{array} \quad (3)$$

where $\circ$ denotes the Hadamard product, $\mathcal{B}$ the Bernoulli distribution and $\mathcal{U}$ the discrete uniform distribution.

IV. Experiments

We evaluate our multimodal network architecture on the Washington RGB-D Object Dataset [15] which consists of household objects belonging to 51 different classes. As an additional experiment – to evaluate the robustness of our approach for classification in real-world environments – we considered classification of objects from the RGB-D Scenes dataset whose class distribution partially overlaps with the RGB-D Object Dataset.

A. Experimental setup

All experiments were performed using the publicly available Caffe framework [12]. As described previously we use the CaffeNet as the basis for our fusion network. It consists of five convolutional layers (with max-pooling after the first, second and fifth convolution layer) followed by two fully connected layers and a softmax classification layer. Rectified linear units are used in all but the final classification layer. We initialized both stream networks with the weights and biases of the first eight layers from this pre-trained network, discarding the softmax layer. We then proceeded with our stage-wise training. In the first stage (training the RGB and depth streams independently) the parameters of all layers were adapted using a fixed learning rate schedule (with initial learning rate of 0.01 that is reduced to 0.001 after 20K iterations and training is stopped after 30K iterations). In the second stage (training the fusion network, 20k iterations, mini-batch size of 50) we experimented with fine-tuning all weights but found that fixing the individual stream networks (by setting their learning rate to zero) and only training the fusion part of the network resulted in the best performance. The number of training iterations were chosen based on the validation performance on a training validation split in a preliminary experiment. A fixed momentum value of 0.9 and a mini-batch size of 128 was used for all experiments if not stated otherwise. We also adopted the common data augmentation practices of randomly cropping 227 $\times$ 227 sub-images from the larger 256 $\times$ 256 input examples and perform random horizontal flipping. Training of a single network stream takes ten hours, using a NVIDIA 780 graphics card.

TABLE I: Comparisons of our fusion network with other approaches reported for the RGB-D dataset. Results are recognition accuracy in percent. Our multi-modal CNN outperforms all the previous approaches.

Method	RGB	Depth	RGB-D
Nonlinear SVM [15]	74.5 $\pm$ 3.1	64.7 $\pm$ 2.2	83.9 $\pm$ 3.5
HKDES [4]	76.1 $\pm$ 2.2	75.7 $\pm$ 2.6	84.1 $\pm$ 2.2
Kernel Desc. [14]	77.7 $\pm$ 1.9	78.8 $\pm$ 2.7	86.2 $\pm$ 2.1
CKM Desc. [3]	N/A	N/A	86.4 $\pm$ 2.3
CNN-RNN [22]	80.8 $\pm$ 4.2	78.9 $\pm$ 3.8	86.8 $\pm$ 3.3
Upgraded HMP [5]	82.4 $\pm$ 3.1	81.2 $\pm$ 2.3	87.5 $\pm$ 2.9
CaRFs [1]	N/A	N/A	88.1 $\pm$ 2.4
CNN Features [20]	83.1 $\pm$ 2.0	N/A	89.4 $\pm$ 1.3
Ours, Fus-CNN (HHA)	**84.1** $\pm$ **2.7**	**83.0** $\pm$ **2.7**	**91.0** $\pm$ **1.9**
Ours, Fus-CNN (jet)	**84.1** $\pm$ **2.7**	**83.8** $\pm$ **2.7**	**91.3** $\pm$ **1.4**

B. RGB-D Object dataset

The Washington RGB-D Object Dataset consists of 41,877 RGB-D images containing household objects organized into 51 different classes and a total of 300 instances of these classes which are captured under three different viewpoint angles. For the evaluation every 5th frame is subsampled. We evaluate our method on the challenging category recognition task, using the same ten cross-validation splits as in Lai *et al.* [15]. Each split consists of roughly 35,000 training images and 7,000 images for testing. From each object class one instance is left out for testing and training is performed on the remaining $300-51 = 249$ instances. At test time the task of the CNN is to assign the correct class label to a previously unseen object instance.

Table I shows the average accuracy of our multi-modal CNN in comparison to the best results reported in the literature. Our best multi-modal CNN, using the jet-colorization, (Fus-CNN jet) yields an overall accuracy of $91.3 \pm 1.4\%$ when using RGB and depth ($84.1 \pm 2.7\%$ and $83.8 \pm 2.7\%$ when only the RGB or depth modality is used respectively), which – to the best of our knowledge – is the highest accuracy reported for this dataset to date. We also report results for combining the more computationally intensive HHA with our network (Fus-CNN HHA). As can be seen in the table, this did not result in an increased performance. The depth colorization method slightly outperforms the HHA fusion network (Fus-CNN HHA) while being computationally cheaper. Overall our experiments show that a pre-trained CNN can be adapted for recognition from depth data using our depth colorization method. Apart from the results reported in the table, we also experimented with different fusion architectures. Specifically, performance slightly drops to 91% when the intermediate fusion layer (fc1-fus) is removed from the network. Adding additional fusion layers

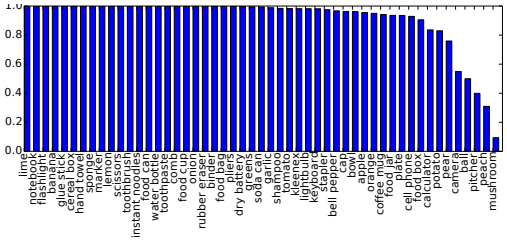


Fig. 6: Per-class recall of our trained model on all test-splits. The worst class recall belongs to mushrooms and peaches.

also did not yield an improvement. Finally, Fig. 6 shows the per-class recall, where roughly half of the objects achieve a recall of $\approx 99\%$.

C. Depth domain adaptation for RGB-D Scenes

To test the effectiveness of our depth augmentation technique in real world scenes, we performed additional recognition experiments on the more challenging RGB-D Scenes dataset. This dataset consists of six object classes (which overlap with the RGB-D Object Dataset) and a large amount of depth images subjected to noise.

For this experiment we trained two single-stream depth-only networks using the Object dataset and used the Scenes dataset for testing. Further, we assume that the groundtruth bounding box is given in order to report only on recognition performance. The first "baseline" network is trained by following the procedure described in Section III-B.1, with the total number of labels $M = 6$. The second network is trained by making use of the depth augmentation outlined in III-C. The results of this experiment are shown in Table II (middle and right column) that reports the recognition accuracy for each object class averaged over all eight video sequences. As is evident from the table, the adapted network (right column) trained with data augmentation outperforms the baseline model for all classes, clearly indicating that additional domain adaptation is necessary for robust recognition in real world scenes. However, some classes (e.g., cap, bowl, soda can) benefit more from noise aware training than others (e.g., flashlight, coffe mug). The kitchen scene depicted in Fig. 4 gives a visual intuition for this result. On the one hand, some objects (e.g., soda cans) often present very noisy object boundaries and surfaces, thus they show improved recognition performance using the adapted approach. On the other hand, small objects (e.g. a flashlight), which are often captured lying on a table, are either less noisy or just small, hence susceptible to be completely erased by the noise from our data augmentation approach. Fig. 7 shows several exemplary noisy depth images from the test set that are correctly classified by the domain-adapted network while the baseline network labels them incorrectly. We also tested the effect of different input image rescaling techniques – previously described in Fig. 3 – in this setting. As shown in the left column of Table II, standard image warping performs poorly, which supports our intuition that shape information gets lost during preprocessing.

TABLE II: Comparison of the domain adapted depth network with the baseline: six-class recognition results (in percent) on the RGB-D Scenes dataset [16] that contains everyday objects in real-world environments.

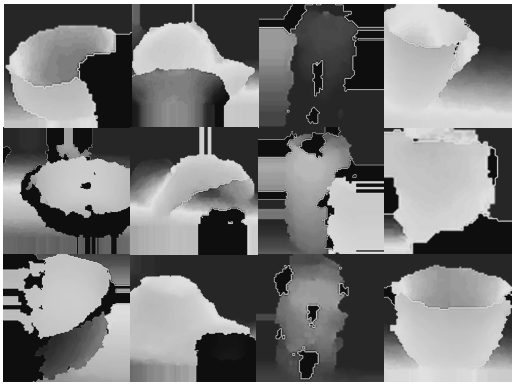
Class	Ours, warp.	Ours, no adapt.	Ours, adapt.
flashlight	93.4	**97.5**	96.4
cap	62.1	68.5	**77.4**
bowl	57.4	66.5	**69.8**
soda can	64.5	66.6	**71.8**
cereal box	**98.3**	96.2	97.6
coffee mug	61.9	79.1	**79.8**
class avg.	73.6 $\pm$ 17.9	79.1 $\pm$ 14.5	**82.1 $\pm$ 12.0**

D. Comparison of depth encoding methods

Finally, we conducted experiments to compare the different depth encoding methods described in Fig. 2. For rescaling the images, we use our proposed preprocessing method described in Fig. 3 and tested the different depth encoding. Two scenarios are considered: 1) training from scratch using single channel depth images 2) for each encoding method, only fine-tuning the network by using the procedure described in Section III-B.1. When training from scratch, the initial learning rate is set to 0.01, then changed to 0.001 after 40K iterations thus stopped after 60K iterations. Training with more iterations did not further improve the accuracy. From the results, presented in Table III, it is clear that training the network from scratch – solely on the RGB-D Dataset – is inferior to fine-tuning. In the latter setting, the results suggest that the simplest encoding method (depth-gray) performs considerably worse than the other three methods. Among these other encodings (which all produce colorized images), surface normals and HHA encoding require additional image preprocessing – meanwhile colorizing depth using our depth-jet encoding has negligible computational overhead. One potential reason why the HHA encoding underperforms in this setup is that all objects are captured on a turntable with the same height above the ground. The height channel used in the HHA encoding therefore does not encode any additional information for solving the classification task. In this experiment, using surface normals yields slightly better performance than the depth-jet encoding. Therefore, we tested the fusion architecture on the ten splits of the RGB-D Object Dataset using the surface normals encoding but this did not further improve the performance. Specifically, the recognition accuracy on the test-set was 91.1 $\pm$ 1.6 which is comparable to our reported results in Table I.

V. Conclusion

We introduce a novel multimodal neural network architecture for RGB-D object recognition, which achieves state of the art performance on the RGB-D Object dataset [15]. Our method consists of a two-stream convolutional neural network that can learn to fuse information from both RGB and depth automatically before classification. We make use of an effective encoding method from depth to image data that allows us to leverage large CNNs trained for object recognition on the ImageNet dataset. We present a novel



a) bowl b) cap c) soda can d) coffee mug

Fig. 7: Objects from the RGB-D Scenes test-set for which the domain adapted CNN predicts the correct label, while the baseline (no adapt.) CNN fails. Most of these examples are subject to noise or partial occlusion.

TABLE III: Comparison of different depth encoding methods on the ten test-splits of the RGB-D Object dataset.

Depth Encoding	Accuracy
Depth-gray (single channel), from scratch	80.1 $\pm$ 2.6
Depth-gray	82.0 $\pm$ 2.8
Surface normals	**84.7** $\pm$ **2.3**
HHA	83.0 $\pm$ 2.7
Depth-jet encoding	83.8 $\pm$ 2.7

depth data augmentation that aims at improving recognition in noisy real-world setups, situations typical of many robotics scenarios. We present extensive experimental results and confirm that our method is accurate and it is able to learn rich features from both domains. We also show robust object recognition in real- world environments and prove that noise-aware training is effective and improves recognition accuracy on the RGB-D Scenes dataset [16].

REFERENCES

[1] U. Asif, M. Bennamoun, and F. Sohel, "Efficient rgb-d object categorization using cascaded ensembles of randomized decision trees," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2015.

[2] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," *arXiv preprint arxiv:1406.5774*, 2014.

[3] M. Blum, J. T. Springenberg, J. Wuelfing, and M. Riedmiller, "A learned feature descriptor for object recognition in rgb-d data." in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2012.

[4] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[5] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for rgb-d based object recognition," in *Proc. of the Int. Symposium on Experimental Robotics (ISER)*, 2012.

[6] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," in *Int. Conf. on Learning Representations (ICLR)*, 2013.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *TPAMI*, pp. 1915–1929, 2013.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision (ECCV)*, 2014.

[11] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision (ECCV)*, 2014.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[14] L. Bo, X. Ren and D.Fox, "Depth kernel descriptors for object recognition," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2011.

[15] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2011.

[16] K. Lai, L. Bo, X. R. Ren, and D. Fox, "Detection-based object labeling in 3d scenes," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2012.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. of the IEEE*, 1998.

[18] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," in *Proc. of Robotics: Science and Systems (RSS)*, 2013.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," 2014.

[20] M. Schwarz, H. Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2015.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems(NIPS)*, 2014.

[22] R. Socher, B. Huval, B. Bhat, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[23] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[24] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[25] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

附录 B　外文资料的调研阅读报告或书面翻译

多模态深度学习与稳定的 RGB-D 物体识别

摘要： 稳定的物体识别是许多现实世界中机器人应用问题中的关键因素。本文基于卷积神经网络（CNN）的最新进展，提出了一种新型的 RGB-D 物体识别架构。我们的体系结构是由两个独立的 CNN 处理流组成的——每个模态对应一个处理流——并继续与后期的融合网络相结合。我们关注不完美的但是在现实世界经常遇到的有噪音的传感器数据。为了获得准确的学习，我们引入了一个分级的训练方法和两个适用于 CNN 的用来处理深度数据的技巧。第一个技巧是，使用一种对 CNN 有效的深度信息编码，使学习不需要很大的深度数据集。第二个技巧是，使用一种适用于深度信息的数据增强方法，用现实的噪音模式破坏它们。我们提出的识别架构在 RGB-D 物体数据集 [15] 上获得了最先进的识别成果，并且在富有挑战性的现实世界 RGB-D 数据中也获得了较好的识别结果。

B.1　引言

RGB-D 物体识别是一项具有挑战性的任务。它在机器人领域以及室内和室外的许多应用中都是核心技术。如今，RGB-D 传感器无处不在。许多机器人系统都可以使用价格便宜，广受支持的开源软件，而不需要使用复杂的硬件来提供独特的感知能力。RGB 数据，提供了有关外观和纹理的信息；而深度数据则包含有关物体形状的附加信息，它是不随亮度或颜色而改变的，所以对物体识别具有特殊的意义。

在本文中，我们提出了使用 RGB-D 数据的新的对象识别方法。特别是，我们关注在有噪声的情况下——一个典型的机器人任务的背景下——稳健地识别不完善的传感器数据。我们的方法建立在机器学习和计算机视觉的最新研究成果上。具体地说，我们扩展了经典的卷积神经网络（CNN），CNN 最近被证明在 RGB 图像 [13]，以及 RGB-D 数据识别方面获得了显着的成功。我们的工作的结构如图 1 所示。由两个关于颜色和深度操作卷积网络流分别处理两种信息。由网络自动学习这两种处理流在后期的融合策略。这种架构与其他最近的多流的方法 [21]，[23]，[11] 相比有着比较大的相似度。将训练出来的单个网络流以及

将合并的部分使用分阶段训练的方式是很有效果的。我们启动训练网络分别对于每个模态，以及两个模态融合的阶段训练微调，与执行网络连接，最终的融合网络会做出最总的分类结果预测。我们初始化 RGB 和深度流网络所用的权值是从一个在 ImageNet 数据集 [19] 上预先训练好的神经网络模型继承来的。尽管我们可以直接使用这个在颜色信息中训练好的模型来分类 RGB 信息，但是并不能直接使用它来分类深度信息。理想情况下，我们可以直接使用深度数据训练出一个适合的 CNN 模型，而不用使用其它模态的数据。然而，实际中是不可行的，因为缺乏大量已经标记的深度数据集。由于缺乏标记的训练的数据，深度模态在训练的初始状态需要借鉴使用 RGB 数据训练出的模型，这是至关重要的。因此，我们提出了一种深度数据编码，以便再利用在 ImageNet 中训练好的 CNN 来识别深度数据。这个已经被试验验证的直觉是简单地将深度图像编码作为 RGB 图像，在三个 RGB 信道中都填上深度数据的信息，然后使用标准（预训练的）CNN 做识别。

在现实环境中，对象是经常受到遮挡和传感器噪声的影响。在本文中，我们提出了一种深度数据增强技术，该技术可用于有噪声场景的稳定的训练。我们增加了训练样例，通过使用从现实环境采样的缺失数据型态来破坏深度数据。使用这两种技术，我们的系统既可以学习稳定的的深度识别和隐式加权两种模式的重要性。ă 我们测试了我们的方法来验证我的想法：我们在 RGB-D 的数据集中使用我们的方法做分类以测试我们方法的精度，然后在有现实世界噪音的数据集中测试我们方法的稳健性。对于第一个，我们的实验结果表明，我们的工作性能优于所有现有技术在 Lai 等人的 RGB-D 对象数据集 [15] 中的分类结果。第二，我们表明我们的数据增强方法提高了在充满挑战的现实世界和有噪声条件下的 RGB-D 场景的数据集 [16] 中的识别准确率。

B.2 相关工作

我们的做事涉及到两块重要的工作，分别是卷积神经网络（CNN）用来做物体识别，和应用计算机视觉的方法去解决使用 RGB-D 数据做识别中遇到的问题。尽管本文不会做一个包含 CNN 和物体识别的综合广泛的文献综述，我们还是会快速的点明我们的方法和近期已有的工作有什么区别。

在众多成功的 RGB-D 物体识别算法中，有很大一部分都是使用手工设计的

信息，比如 SIFT 与深度信道的多种形状特征结合 [15]，[16]。然而，随着它们在许多计算机视觉的问题中都获得了成功，非监督学习的特征学习方法也在最近被拓展到了 RGB-D 识别的领域。Blum 等人 [3] 提出了一种基于 K-Means 的 RGB-D 描述符。更加近期的工作包括 Bo 等人 [5] 提出的分层匹配追踪，一种分层级的稀疏编码方法，可以从多个输入信道学习特征。Socher 等人 [22] 提出了一种不同的方法，这种方法依靠将卷积过滤器和递归神经网络（一种特殊的再现神经网络）结合起来作为识别层级。Asif 等人 [1] 报告了一种可以使用随机森林瀑布来提高识别准确率的分类器，这种分类器也是分层融合的。最后，在最近 Schwarz 等人的独立工作中 [20]，提出了使用 ImageNet 上预先训练出的 CNN 提取出的 RGB-D 特征来做 RGB-D 识别。尽管他们也使用有两个网络流组成的分类架构，他们并没有对 CNN 使用 RGB-D 做微调，而只是使用原有的神经网络。有趣的是，他们还发现了简单的上色方法可以使深度信息获得较好的识别效果，甚至可以与许多复杂的预处理的效果相当。与他们的工作不同，我们获得了更高的准确率，并且训练出一个端对端的融合 CNN：输入未加工的像素信息，直接输出分类信息，整个过程是一个监督学习的过程，并且使用了在相关任务中训练的模型。所以，我们的 CNN 学习到的特征从结构上与现有的其他任务中的特征是可区分的。使用 CNN 做物体识别在计算机视觉和机器学习领域有着很长的历史。尽管人们都知道 CNN 在监督的图像识别、分类任务中（比如 MNIST [17]）很早就可以获得较高的准确率，最近 CNN 方法已经不仅仅可以在例如大规模图例如分类任务 [13]、物体识别 [9]、语义分割 [8] 等方面中超过经典的方法，还可以产生在多种任务中转换的特征 [7]，[2]。计算能力强大的高性能计算系统使得这个最近的成功的故事成为可能，而大规模图像数据集的出现也起了不可或缺的作用，比如 ImageNet[19]。

虽然大多数的深度学习工作的重点是二维图像，最近的研究也已经包含使用深度信息用于改进场景分类和物体检测 [6]，[10]。其中，和我们最相似的是 Gupta 等人 [10] 提出的比较普适的使用 R-CNN 检测器适用于深度信息的方法 [9]。详细地说，他们使用已经在 RGB 图片数据集上训练好的大型 CNN 去提取深度信息中的特征，将深度信息编码到三个信道（使用 HHA 编码）。更加具体地说，他们对每一个像素点，编码了离地面的高度，水平的距离和像素层面上的表面法向和重力之间的角度。我们的融合网络架构与他们的使用预先训练的神经网络有着相似之处。不过我们的方法与他们的方法的不同之处在于，将深度

信息编码为颜色信息的过程和融合两种模态的融合方法。对于编码过程，我们提出了一种对于深度图片的编码方法，我们的这种方法不依赖于复杂的预处理，但是与 HHA 相比获得了性能上的改进。为了实现模态融合，我们在原有神经网络的基础上有加了一层融合层，使得我们的识别架构可以自动的学习出对于识别的融合的策略，这和仅仅在提取出的信息的上面训练出一个线性分类器先比有着更好的表现。多刘的架构已经被人用于多种任务，包括识别 [21]，检测 [11]，图片提取 [23]。一个最近的有关深度和图片信息融合的不同网络结构在 Saxena 等人的研究中给出 [18]。在哪里，作者比较了不同的多模态学习的模型：

1. 早期融合，其中输入图像被级联到现有的图像 RGB 通道并一起处理；
2. 我们定义为后期的融合，其中的功能是为每个单独训练的情态，然后在更高层融合；
3. 将早期和晚期的方法结合；

得出的结论是后期融合（2）和早期和后期相结合的办法最适合把握问题的关键特征，有利于检测。相比与他们的工作，我们的模型是使用类似于后期融合的方法，但又有着很大的不同。[18] 使用逐层无人监督的训练方法，他们的规模（包括其网络和输入图像数据集的大小）都比我们要小一个数量级。

B.3 用于 RGB-D 物体识别的多模态架构

该架构的概述在图 1 中给出。我们的网络由两个流（在图 1 中的顶部蓝色和底部绿色）——分别处理 RGB 和深度数据——然后在更晚的阶段被融合在一起的方法。每个流由一个已经在 ImageNet 数据集中训练好的深度 CNN 组成（我们使用 Krizhevsky 等人 [13] 训练的 CaffeNet[12] 作为预训练模型）。使用预先训练的神经网络的关键原因是我们希望在优先的数据量上使用一个包含着百万级别的参数的大型 CNN。（例如，见 Yosinski 等人 [25] 在最近对于此的讨论）我们首先对两个模态的信息进行预处理，然后使用分阶段的方法训练我们的多模态 CNN 网络。我们首先对每个模态的 CNN 进行微调，然后使用两个模态微调出的模型，将它们结合在一起，联合的训练出融合网络的参数。不同的步骤将会在一下小节被详细说明。

B.3.1 输入预处理

要充分利用在 ImageNet 上预先训练的 CNN 的力量，我们预处理 RGB 和深度模态的输入数据，使得它与原 ImageNet 的输入兼容。具体地说，我们使用在 CaffeNet[12] 中的参考实现，这个模型接受 227 × 227 像素的 RGB 作为输出。这 227 × 227 个像素点通常是用 256 × 256 个像素点的 RGB 图片缩放而成的。处理的第一步要缩放图片到合适的大小。而缩放最简单的方法就是使用图片翘曲，放弃原始图片的长宽比例直接缩放。这在图片 3（中）中有所阐释。我们发现在我们的试验中，这个处理过程会降低识别的准确率——我们把这个问题归因于目标物体形状信息的损失（见 IV-C）。所以我们设计了一种不同的预处理方法：将长边缩放到 256 个像素，得到一个 256 × N 或者 N × 256 的图片。然后我们将缩放后的图片沿短边像瓷片一样平铺到 256 × 256 的空间上。得到的最终的颜色图片和深度图片在物体边界处展示出一种人造的环境（见图 3）。我们对颜色信息和深度信息都是用这种方法进行预处理。

对于 RGB 图像，在预处理缩放之后可直接使用作为 CNN 的输入，而缩放之后的深度数据需要额外的步骤。为了实现这一点，我们可以回想一下是用 ImageNet 训练的神经网络已经被训练去识别某个特别的图像输入分布（即天然相机图像），它与深度传感器得到的标志物体距传感器的距离的信息是不相容的。尽管如此，通过查看一个典型的家用物品的深度图片（参考图 4），可以得出结论，许多定性出现在 RGB 图像中的特征，比如边、角、阴影区域等模式，也会出现在深度图片中。这种认识引出了这种想法：简单地使用深度数据渲染出的三通道的 RGB 图片作为用于在 ImageNet [10] 上训练的 CNN 的输入。我们将比较这些深度图片渲染的方法与我们的做法的不同之处。两种最流行的深度图片编码是：

1. 深度数据渲染成灰度值，并复制灰度值到所要求的三个通道作为神经网络输入；
2. 利用表面法线，其中每个法线矢量的维度对应于一个信道所产生的图像。
3. 一个更复杂的方法，称为 HHA 编码 [10]，对每一个像素点的三个信道，编码了离地面的高度，水平的距离和像素层面上的表面法向和重力之间的角度。

我们提出第四种有效并且计算成本较低的，将信息编码为彩色图像的方法，我们发现此种方法在物体识别中的表现要比 HHA 编码更好。我们的方法首先要

标准化所有的深度值到 0 和 255。然后，我们应用一个喷射颜色表将深度信息从单通道输入转换到一个三通道图像（给深度着色）。对于尺寸为 $W \times D$ 的深度图像中的每个像素（i，j）中，我们将距离信息映射为颜色值，从红色（附近）到绿色再到蓝色（远），并将这些信息分散到三个信道。这三个信道的边缘往往对应着有用的对象边界。因为我们使用的卷积神经网络是用 RGB 图像训练的，着色过程提供了深度信息和 RGB 图像之间足够多的共同结构，所以可以使用 CNN 来处理深度信息（对不同深度预处理方法之间的比较参照图 2）。

B.4 实验

我们在华盛顿大学的 RGB-D 物体数据集 [15] 上评估对我们的多模态神经网络体系结构。这个数据集包括属于 51 个不同类别的日常家用物品。作为额外实验，我们使用了 RGB-D 场景数据集中的被部分遮挡的物体来评估的算法的可靠性以及对于现实世界数据的分类能力。

B.4.1 实验设置

所有实验都是使用开源的 Caffe 框架 [12]。如先前所描述，我们使用的 CaffeNet 作为我们融合网络的基础。它由五个卷积层（在第一、第二和第五之后需要最大池化），其次是两个完全连接层和 SOFTMAX 分类层。除了最后一层之外所有的层都用到了修正线性单元。我们使用预先训练的模型的前八层的权重和偏差来初始化我们网络的前八层，丢掉了最后的 SOFTMAX 分类层。然后我们继续使用分阶段训练。第一阶段（分别训练 RGB 和深度信息）当中所有信息层的参数都按照一种固定学习率（learning rate）的模式在改变（最开始学习率取为 0.01，经过 20K 次迭代后变为 0.001，并在 30K 次迭代后停止训练）。在第二阶段（训练融合网络，20K 次迭代，使用大小为 50 的 mini-batch），我们尝试微调了所有的权值，但是最后发现将上一阶段训练出的两个模态的各自的 CNN 网络权值固定（设定它们的学习率为 0），只微调融合部分的网络效果更好。我们从此任务的十种数据分割方法中随机选出一种分割作为确认分割，而训练的迭代次数是基于在确认分割上的测试结果所选定的。如果没有特殊说明，我们使用固定为 0.9 的 momentum 和固定为 128 的 mini-batch。我们还使用了常见的数据增强的做法：在较大的 256×256 的图片中随机选取大小为 227×227 子图像，

然后实行随机水平翻转。我们使用一块 NVIDIA 780 显卡，训练单支神经网络流需要 10 个小时。

B.4.2 RGB-D 物体数据集

华盛顿 RGB-D 数据集的物体由 41877 个含家用物品 RGB-D 图像组织成，它们分别属于 51 个不同的种类的一共 300 个实例。这些图像是在三个不同视点捕捉的。我们使用间隔 5 帧的采样来评估我们的算法。我们使用和 Lai 等人 [15] 相同的十折交叉验证分割，来评估我们的方法在具有挑战性的类别识别任务中的表现。每个分割包括大约 35000 组训练图像和 7000 组测试图像。每一个对象类中，一个是被留下做测试的。所以我们在剩下的 300-51 = 249 个实例中进行训练。在测试时 CNN 的任务是给每个之前没见过的实例分类。

表 I 显示了我们的多模态 CNN 识别的平均准确率与以往工作中的最好结果的对比。我们得到的最佳结果，使用喷射着色（FUS-CNN JET）与 RGB 和深度信息时（84.1 ± 2.7% 和 83.8 ± 2.7% 时，当分别使用 RGB 或深度模式时），得到的 91.3 ± 1.4% 的整体准确率，比就我们所知的所有前人的工作都要高。我们还展示了使用需要更多计算的 HHA 编码的结果（FUS-CNN HHA）。从表中可以看出，HHA 编码并没有带来表现的提升。我们提出的着色方案要比 HHA 的表现（FUS-CNN HHA）稍好一点，同时需要更少的计算资源。整体来说我们的试验结果表明，一个预先训练的 CNN 可以用来识别深度数据，只需要预先使用我们的着色方法对深度信息进行预处理即可。除了表中展示的结果，我们还做了有关不不同的融合网络结构的试验。具体来说，如果将融合中间层（fc1-fus）去掉，识别准确率会稍微下降至 91%。增加融合层也不能得到更好的结果。最终，图 6 展示了每个类的召回率（recall），其中大约一半物体实现了约等于 99% 的召回率。

B.4.3 深度域适应 RGB-D 场景

为了测试我们的深度信息增强方法在真实世界中的效果，我们进行了在更具挑战性的 RGB-D 场景数据集中的附加实验。此数据集包括六个对象类（与 RGB-D 物体数据集重叠）和大量受噪声影响的深度图像。在这个实验中，我们训练的两个单流深度神经网络，使用物体数据训练，并使用场景数据集进行测试。此外，我们假设一定正确的边框已经被给出，这种就可以只测试识别的性

能。第一个“基准”神经网络由第 III-B.1 中描述的方法所训练出来，标签的总数目 M = 6。第二神经网络是使用 III-C 中提出的深度信息增强方法训练出来。该实验的结果展示于表 II（中间和右栏），报告了每个对象类在所有八个视频序列中的识别精度。从该表中可以看出，经过适应的网络是明显的（右边使用数据增强方法训练的列）优于基准模型中的所有类，这显然表明额外的领域适应性对于在现实世界中的场景的稳健识别是有必要的。然而，一些类（例如，盖，碗，苏打罐）从噪音训练中获得提升比其他类更高（例如，手电筒和咖啡杯）。图 4 中描绘的厨房场景给出了一些有关此结果的视觉直觉。另一方面，一些对象（例如，汽水罐）往往会出现非常嘈杂对象边界和表面，因此它们使用适应方法会显示出较大的识别性能改进。另一方面，小物体（例如，手电筒），这些物体经常是在桌子上拍的，要么不受噪音影响或只受轻微的影响，因此这些噪音容易被我们的数据增强方法完全擦除。图 7 显示了几个被基准网络错误分类，但是可以使用我们的数据增强方法正确分类的例子。我们还测试了如图 3 所描述的不同的图像缩放技术的效果。如表 II 所示，标准图像变形表现的不是很好，这支持了我们的直觉，即形状信息会在预处理过程中丢失。

B.4.4 深度编码方法的比较

最后，我们进行了实验，以比较图 2 所描述的不同的深度编码及方法。对于图像的缩放，我们使用了图 3 描述的需处理方法，并使用不同的深度信息编码方法进行了测试。我们考虑到两种场景：

1. 使用单通道深度图像做训练
2. 对每一种编码方法，只使用第三节-B.1 的方法来预处理深度信息，然后作为训练集。

当从头训练时，初始学习率设定为 0.01，然后在 40K 次迭代后改变到 0.001。经过反复迭代 60K 次后停止。训练更多的迭代次数并不能进一步提高精度。通过查看表 III 中的结果，很显然训练从头训练的表现不如使用已有模型微调的效果好。在后一种情景下，结果显示最简单的编码方式（将深度信息转换为灰度图片）的识别效果要显著差于其他的方法。在其他的编码方法中（所有这些方法都会给深度信息着色），平面法向和 HHA 编码需要额外的图片图像预处理，而使用我们提出的深度-喷射着色的方法几乎不需要任何计算资源。一个 HHA 编码在此情景下表现的不尽人意的可能原因是此数据库中的图片都是放在一个可

以旋转的托盘中，他们的高度都是相同的。所以 HHA 中用到的高度信道就无法包含用于分类的其他信息。在本实验中，使用平面法线要比深度-jet 编码的表现稍好一点。所以我们使用平面法线的编码来试验了我们的融合网络，不过这并没有进一步提高识别表现。具体地说，测试集上的识别准确率是 91.1×1.6，这与我们在表 I 中报告的结果相差不大。

B.5 结论

我们提出了一种可以用于 RGB-D 物体识别，并且在 RGB-D 物体数据集 [15] 中实现了当前最好表现的新型多模态神经网络结构。我们的方法是由二个卷积神经网络流组成，这两个 CNN 流在分类前可以从 RGB 和深度信息中自动地提取和融合信息。我们利用一种有效的编码方式，把深度数据编码为图像，使我们能够充分利用在 ImageNet 数据集中训练的大型神经网络进行对象识别。我们提出了一种新的深度数据增强方法，旨在提高对嘈杂的现实世界深度数据的识别准确率，以便适用于典型的机器人场景。目前大量的实验结果证明了我们方法的正确性，以及它能够从两个模态学习丰富的特征。我们的方法还可以在现实世界环境中稳定地识别物体，并证明了噪声感知训练是有效的，并且可以被用来改进 RGB-D 场景数据集 [16] 的识别精度。

在学期间参加课题的研究成果

个人简历

1993 年 06 月 10 日出生于黑龙江省大庆市。

2012 年 9 月考入清华大学计算机科学与技术系攻读计算机科学与技术工学学士学位至今。

发表的学术论文

[1] **Gao B**, Li H, Li W, et al. 3D Moth-inspired chemical plume tracking and adaptive step control strategy[J]. Adaptive Behavior, 2016: 1059712315623998.

[2] **Gao B**, Li H, Sun F. 3D moth-inspired chemical plume tracking[C]//2015 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, 2015: 1786-1791.