

中国科学技术大学

博士学位论文



社交多媒体数据

语义理解和关联表达

作者姓名 : 吴岳

学科专业 : 信号与信息处理

导师姓名 : 俞能海 教授 李世鹏 教授

完成时间 : 二〇一七年五月

University of Science and Technology of China
A dissertation for doctor's degree



**Social Media Data
Semantic Understanding and
Associative Expression**

Author's Name: Yue Wu
Speciality: Information and Communication Engineering
Supervisor: Prof. Nenghai Yu Prof. Shipeng Li
Finished Time: May, 2017

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____

签字日期：_____

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

公开 保密（____ 年）

作者签名：_____

导师签名：_____

签字日期：_____

签字日期：_____

摘要

近年来，智能手机及其它移动智能设备呈现出了爆发式的增长与普及。高清摄像头、大容量存储和高速的网络连接为用户创造了极其便利的拍摄和分享条件，用户几乎可以在任意时间、任意地点拍摄照片或视频，并将它们分享到社交网络，产生了海量的社交多媒体数据。然而，这些数据都以碎片化的形式存在，当前的社交多媒体系统缺乏智能的工具或服务将它们组织起来，并选取符合用户个性化需求的数据呈现给用户，用户也很难快速准确地检索到他们需要的数据。因此，如何充分挖掘和利用社交多媒体数据成为了当前重要的研究问题。

本论文对社交多媒体数据的语义理解和关联表达做了深入研究，目标是实现一个能够理解社交多媒体数据、根据用户需求选取有关联的数据、并以丰富的表达形式呈现给用户的关联表达系统。由于社交多媒体数据的语义内容丰富多样，对每个语义收集并标注数据的难度和成本很高，语义理解首先需要解决标注难的问题。其次，由于社交多媒体数据的规模庞大，语义理解需要解决处理慢的问题。社交多媒体数据的关联表达是基于对社交多媒体数据语义理解的结果，根据用户个性化的需求选取有关联的数据，并以丰富的表达形式呈现给用户。本论文分别从照片和视频两个角度研究了关联表达的具体应用。语义理解和关联表达构成了挖掘和利用社交多媒体数据相对完整的框架。

针对上述问题，本论文的主要研究工作和创新成果包括：

1. 对于语义理解标注难的问题，提出了一种直接从社交多媒体数据学习语义理解模型的弱监督相关反馈深度学习算法。传统深度学习算法对于训练数据中的标注噪音十分敏感，本论文基于感知连续性，利用数据在特征空间的相互关系，使得不同数据在训练过程中有不同的贡献加权，从而抑制噪音标注的影响。实验结果表明，与已有算法相比，本论文提出的弱监督相关反馈深度学习算法具有更好的噪声鲁棒性。
2. 对于语义理解处理慢的问题，提出了一种从大规模高维数据中选取特征的高效算法。本论文基于二阶在线学习算法，利用特征的置信度进行特征选取，并利用最大/最小堆结构提出了快速在线特征选取算法。由于置信度的单调递增特性，本论文进一步改进了快速二阶在线特征选取算法，将算法的复杂度降低为于非零特征数目成正比。实验结果表明，该算法能够极大减少特征选取的计算时间，并达到当前最好特征选取算法的准确率。
3. 对于语义理解处理慢的问题，提出了基于二阶在线特征选取算法的深度卷

摘要

积神经网络模型简化算法。算法增加了对应卷积层输出特征图每个通道的权重层，并在权重层上进行特征选取，从而将三维卷积核的组稀疏优化问题转化为一维特征选取问题。实验结果表明，该模型简化算法在不损失模型准确率的情况下极大减少了模型的参数个数。

4. 对于照片关联表达问题，提出了一个基于主题的照片集故事化表达系统——Monet。系统首先根据照片的时间和位置信息对照片集进行事件检测。其次，根据照片的质量、多样性和均衡性选取一部分代表性的图片。然后，系统利用弱监督深度学习算法对代表性照片进行内容表征，并利用在线特征选取算法选取最能代表照片内容的特征子集。系统根据照片的特征赋予不同照片不同的主题。最后，系统运用可计算的视频编辑语法对照片进行动画特效处理以及音乐的匹配，生成具有关联表达能力的不同主题的视频。
5. 对于视频关联表达问题，提出了全自动移动多摄像头视频自动剪辑系统——MoVieUp。本论文从音频剪辑和视频剪辑两个角度解决自动剪辑问题。音频剪辑对音频流进行质量评估，在最少切换次数准则下选取高质量的音频流片段，并拼接成单一音频流。视频剪辑首先根据音频的节奏以及语义特性选取镜头切换点。其次，系统对多摄像头视频进行语义分割，得到子镜头，并评估子镜头的视觉质量、相机运动以及相互之间的多样性，最终在保证镜头运动一致性的条件下最大化镜头质量和多样性，完成镜头选取，并拼接镜头得到单一视频流。系统最后对剪辑后的单一音频流和单一视频流进行混流，得到最终剪辑好的视频呈现给用户。

关键词：弱监督深度学习 特征选取 模型简化 照片集故事化表达 视频自动剪辑

ABSTRACT

The recent years have witnessed the explosive growth and ubiquity of mobile smart devices. The high resolution cameras, large storage, and high speed network connection of mobile devices have founded the superior conditions for capturing and sharing. Users can capture and share photos or videos at almost anytime and anywhere. Up to now, the scale of social media data has increased to a huge scale. However, these data exist in a fragmented way on social media, lacking intelligent services to organize them. Neither can social media provide data according to personalized user needs, nor can users search for their required data efficiently and effectively. As a result, how to exploit and utilize the large scale social media data has become an important problem.

This dissertation probes into the semantic understanding and associative expression of social media data. The aim is to implement an intelligent system that can understand, select, and express social media data in an associative way. Due to the wide range of semantics, it's hard to collect and label data for every semantic tag. Semantic understanding should solve the difficulty of labeling. Besides, accelerating the processing speed is essential due to the large scale of social media data. Based on the semantic understanding, this dissertation studies associative expression from photo and video aspects. Semantic understanding and associative expression compose a relatively complete framework for mining and utilizing social media data.

This dissertation conducts deep research on social media data semantic understanding and associative expression related problems with the following achievements:

1. For the difficulty of labeling, we propose a weakly supervised relevance feedback deep learning algorithm to learn from weakly labeled social media data directly. Traditional deep learning algorithms are sensitive to label noises. Our algorithm utilizes the perceptual consistency to attenuate the sensitiveness, which uses the correlation in the feature space to make different samples contribute differently during training. Empirical evaluation with comparison to existing algorithms shows that the relevance feedback algorithm has better robustness to label noises.
2. For the processing speed, we propose a large scale high dimensional second-order online feature selection algorithm. Based on second-order online learning algorithms, the algorithm selects features according to the confidence of features. We

propose fast algorithms with the Max/Min heap. Due to the monotonous increasing property of confidence, we further reduce the complexity of the proposed algorithm to be linear to the number of non-zero features. Empirical evaluation shows that the algorithm can significantly reduce the training time while achieving comparable accuracy to state-of-the-art feature selection algorithms.

3. For the processing speed, we also propose a model simplification algorithm for deep convolutional neural networks based on online feature selection. The algorithm adds a weighting layer corresponding to each channel of the output feature maps of convolutional layers. The group sparsity problem on the three dimensional convolutional kernels is then transformed into the online feature selection problem on the one dimensional weighting vector. Empirical evaluation shows that model parameters are reduced significantly with little impact on the accuracy.
4. For associative photo expression, we propose a theme-based photo storytelling system—Monet. First, the system detects events in photos according to the time and location information. It then selects a representative photo subset according to photo quality, diversity, and uniformity. After that, the system uses the weakly supervised relevance feedback algorithm to analyze the content of the representative photos. Online feature selection algorithm is applied to extract the most distinctive features. Based on the features, each photo is assigned with a theme. Finally, a fancy video with animation and motion effects is generated and aligned with a music according to the computational filming grammars of each theme.
5. For associative video expression, we propose an automatic mobile multi-camera video mashup system—MoVieUp. We solve the mashup problem from audio and video aspects. For audio mashup, the system assesses audio quality, selects high quality audio segments, and stitches them into a single audio stream. For video mashup, the system detects the switching points according to the tempo and semantics of audio. After that, videos are structured into subshots. The system evaluates the quality, motion, and diversity of the subshots. Video shots are selected by maximizing the quality and diversity under the constraint of motion consistency. Finally, the system multiplexes the audio and video streams to generate the well-edited video.

Key Words: weakly supervised deep learning, feature selection, model simplification, photo storytelling, video mashup

目 录

摘要	I
Abstract	III
第 1 章 绪论	1
1.1 研究背景和意义	1
1.2 社交多媒体数据研究的关键问题	3
1.3 本文主要工作	4
1.4 本文主要创新点	7
1.5 本文结构安排	8
第 2 章 国内外研究现状和工作基础	9
2.1 弱监督学习	9
2.1.1 数据去噪	9
2.1.2 噪音鲁棒模型	10
2.2 特征选取	12
2.2.1 批处理方法	12
2.2.2 在线特征选取	12
2.3 模型简化	13
2.4 社交多媒体数据的关联表达	15
2.4.1 基于主题的照片集故事化表达	15
2.4.2 移动多摄像头视频自动剪辑	16
第 3 章 弱监督社交多媒体数据语义理解	19
3.1 弱监督目标识别问题建模	19
3.2 弱监督相关反馈深度神经网络	20
3.2.1 经典深度卷积神经网络	20
3.2.2 相关反馈深度卷积神经网络	21
3.2.3 相关反馈分析	25
3.3 实验结果和评估	26
3.3.1 目标识别	26
3.3.2 社交图片标注	30
3.4 本章小结	31

目 录

第 4 章 大规模社交多媒体数据快速处理	33
4.1 在线特征选取问题建模	33
4.2 置信度加权二阶在线特征选取	34
4.3 快速在线特征选取算法	37
4.3.1 一阶快速在线特征选取算法	38
4.3.2 二阶快速在线特征选取算法	38
4.3.3 复杂度分析	40
4.4 置信度加权二阶多类在线特征选取	41
4.5 实验结果和评估	42
4.5.1 实验设置	43
4.5.2 合成数据集实验评估	43
4.5.3 中等规模真实数据集实验评估	46
4.5.4 物体识别实验评估	47
4.5.5 大规模真实数据集实验评估	49
4.6 深度卷积神经网络模型简化	50
4.6.1 深度卷积神经网络模型简化	51
4.6.2 基于在线特征选取的模型简化	52
4.6.3 实验结果和评估	53
4.7 本章小结	54
第 5 章 基于主题的照片集故事化表达	57
5.1 主要问题与系统框架	57
5.2 照片集分析与梳理	58
5.2.1 事件检测	58
5.2.2 照片筛选	59
5.3 照片集故事合成	61
5.3.1 语义理解	61
5.3.2 风格选取	61
5.3.3 生成视频片段	62
5.3.4 音乐分析	63
5.3.5 故事合成	63
5.4 实验结果和评估	66
5.4.1 事件检测和关键照片选取评估	66
5.4.2 照片集故事合成评估	67
5.5 本章小结	69

目 录

第 6 章 移动多摄像头视频自动剪辑 ······	71
6.1 主要问题 ······	71
6.2 可计算视频剪辑语法 ······	72
6.2.1 用户调研 ······	73
6.2.2 视频剪辑调研结果 ······	73
6.2.3 音频剪辑调研结果 ······	74
6.2.4 可计算视频剪辑语法 ······	74
6.3 移动多摄像头视频自动剪辑系统 ······	75
6.3.1 系统框架 ······	75
6.3.2 音频剪辑 ······	77
6.3.3 镜头切换点检测 ······	78
6.3.4 视频镜头选取 ······	80
6.4 实验结果和评估 ······	83
6.4.1 数据集 ······	83
6.4.2 实验设置 ······	84
6.4.3 音频剪辑评价 ······	85
6.4.4 切换点检测评估 ······	86
6.4.5 视频剪辑评估 ······	87
6.5 本章小结 ······	90
第 7 章 总结与展望 ······	93
7.1 本文总结 ······	93
7.2 研究工作展望 ······	95
参考文献 ······	97
致谢 ······	109
在读期间发表的学术论文与取得的研究成果 ······	111

目 录

图目录

1.1 社交多媒体数据产生和利用现状	2
1.2 社交多媒体数据研究的关键问题	3
1.3 社交多媒体数据的研究内容和相互关联	5
3.1 数据标注噪音类型: (a) 完全随机噪音; (b) 随机噪音; (c) 非随机噪音 ..	20
3.2 深度卷积神经网络结构图	21
3.3 感知连续性示例	22
3.4 弱监督相关反馈深度卷积神经网络	23
3.5 训练数据对于梯度的贡献曲线, 横坐标为数据与其他数据特征之间的 距离	25
3.6 参数 α 对于神经网络分类性能的影响	27
3.7 不同算法在不同噪音比例下的准确率相对于无噪音准确率的下降程 度比较	28
3.8 社交图片标注结果示例	30
4.1 合成数据集 \mathcal{X}_1 和 \mathcal{X}_2 上测试准确率和特征数目之间的关系	44
4.2 合成数据集 \mathcal{X}_1 和 \mathcal{X}_2 上训练时间和特征数目之间的关系	45
4.3 中等规模数据集上不同算法测试准确率比较	47
4.4 中等规模数据集上不同算法训练时间比较	48
4.5 VOC2007 数据集上不同算法在不同特征数目下的测试准确率和训练 时间比较	49
4.6 “news” 和 “rcv1” 数据集上测试准确率和训练时间与特征数目之间 的关系	50
4.7 移除卷积核对深度卷积神经网络的影响示意图	51
4.8 深度卷积神经网络辅助权重层模型简化	52
4.9 模型简化对模型准确率的影响	54
5.1 基于主题的照片集故事化表达系统框架	58
5.2 风格选取流程	62
5.3 视频特效、形状、颜色过滤器和转场样例	63
6.1 移动多摄像头视频自动剪辑示意图	72

图目录

6.2 移动多摄像头视频自动剪辑系统框架	75
6.3 移动多摄像头视频自动剪辑系统符号表示	76
6.4 音频质量评估结果示例	78
6.5 不同 δ 取值对视频切换点检测的影响	80
6.6 移动多摄像头视频自动剪辑示例	83
6.7 视频剪辑评价页面	85
6.8 视频切换频率比较	87
6.9 视频切换点位置比较	87
6.10 MoVieUp 系统与 Virtual Director 系统视频剪辑结果比较	88
6.11 MoVieUp 系统与 Jiku Director 系统视频剪辑结果比较	90

表目录

2.1 照片集关联表达系统比较	16
2.2 移动多摄像头视频自动剪辑系统比较	17
3.1 Cifar10 数据集上不同算法在不同噪音比例下的准确率比较	28
3.2 VOC2007 数据集上不同算法在不同噪音比例下不同类别的平均准确率比较	29
3.3 社交图片标注结果比较	30
4.1 合成数据信息 (“K”,“M”,“B” 分别代表千, 百万, 十亿)	44
4.2 SOFS 算法可伸缩性评测	45
4.3 中等规模特征选取数据集详情	46
4.4 大规模真实数据集信息	49
4.5 大规模高维数据集上不同特征选取算法比较 (ρ 是选取的特征比例) ..	50
4.6 主流深度卷积神经网络的深度和参数规模	51
4.7 VGG-BN 网络结构	53
4.8 VGG-BN 网络不同卷积层在不同准确率下的稀疏度 (ρ 为下降百分比)	54
5.1 用户照片集详细信息	66
5.2 事件检测结果比较	67
5.3 故事生成主观评测结果	68
6.1 移动多摄像头视频数据集及算法优化时间	84
6.2 音频剪辑切换次数比较	86
6.3 音频剪辑主观评价 (MOS) 比较	86
6.4 MoVieUp 和 Virtual Director 系统存在质量问题的镜头数量比较	89
6.5 MoVieUp 和 Jiku Director 系统存在质量问题的镜头数量比较	89

表目录

算法索引

4.1 PET——截断感知机算法	35
4.2 Truncate——截断函数	35
4.3 FOFS——一阶在线特征选取算法	36
4.4 SOFS——二阶在线特征选取算法	37
4.5 一阶快速在线特征选取算法	39
4.6 二阶快速在线特征选取算法	40

第1章 绪论

本章首先介绍社交多媒体数据的含义以及社交多媒体数据的研究背景和意义，并由此引出社交多媒体数据的研究中存在的关键问题，然后介绍本论文的主要工作及创新点，最后介绍全文的结构安排。

1.1 研究背景和意义

近年来，智能手机及其它智能移动设备呈现出了爆发式的增长与普及。高清摄像头、大容量存储和高速的网络连接为用户创造了极其便利的拍摄和分享条件，从而创造了海量的多媒体数据。用户几乎可以在任意时间、任意地点拍摄照片或者视频，并将它们分享到社交网络。据统计，截止到2014年产生了大约2.7万亿的用户照片。到2017年，该数字将会增长到约4.9万亿^[1]。这类由用户产生的多媒体数据具有明显的社交性特点：

1. 用户通常在社交活动中拍摄照片或者视频，同一时间段内同一事件中的其他用户也会拍摄与之相关的内容；
2. 大量拍摄的照片和视频被用户分享到Flickr、Instagram、YouTube、美拍、优酷等社交网站，这些数据包含的时间、地点、内容等信息与其他用户分享的多媒体数据产生关联。

因此，本文将这类用户多媒体数据称为**社交多媒体数据**。

然而，由于缺乏智能的社交多媒体数据语义理解和关联表达服务，这些海量的数据所包含的信息并没有得到充分的挖掘和利用。如图1.1所示，一方面，用户、拍摄设备以及社交多媒体数据的规模都呈现出并保持着爆发式的增长趋势；另一方面，这些记录了珍贵记忆的用户数据被大量的存储在本地或者云端磁盘上，却很少再次被用户浏览和利用。海量的社交多媒体数据并没有带来更好的用户体验。随着人们对于多媒体的内容品质和个性化需求的提高，越来越多的研究人员和工业界研究机构投入更多的精力到社交多媒体数据中。

相比于传统多媒体数据，社交多媒体数据具有以下特点：

1. **质量不确定**。由于普通用户的拍摄技巧比较业余、拍摄时的光线限制、相机的快速移动或者抖动、场景快速变换等原因，社交多媒体数据通常伴有抖动、散焦、过度曝光、欠曝光、模糊、遮挡以及拍摄出无意义的照片或视频等问题，影响了社交多媒体数据的后期浏览体验。



图 1.1 社交多媒体数据产生和利用现状

2. **内容冗余。** 用户在拍照时通常采取多次拍摄的方式来获得最理想的拍照效果，使得社交多媒体数据中存在大量冗余。这些冗余数据带来三个方面的主要问题：(1) 需要耗费大量的时间和精力去整理；(2) 占据了大量的存储资源；(3) 增加了用户查找数据的难度。
3. **多样性。** 用户拍摄的时间、地点和环境比较随机，拍摄的角度、内容具有很大的不确定性，使得社交多媒体数据的内容表现出多样化的特点。社交多媒体数据相关的服务，不仅需要去除冗余的用户数据，也需要最大程度的保留用户数据的多样性。
4. **故事性。** 用户的拍摄行为并不是随机的，而是选择性地记录对他们有意义的时刻和场景。相同时间段内相近地点拍摄的多媒体数据，浓缩了用户在一段时间内的足迹以及经历的事件。完整的多媒体数据则记录了很长一段时间内发生在用户身上的故事。然而，当前社交多媒体数据相应的服务还不能很好地将这些原始的记录片段整理成高观赏性的纪录片。
5. **具有位置和时间信息。** 现代移动设备拍摄时通常都能检测到拍摄的时间、地点等信息，并将这些信息存储在图像或者视频的文件头中。这些时间和地点信息记录了拍摄的时空上下文关系，使得数据之间能够产生关联。基于用户历史拍摄的时间、地点和具体内容，可以更好地挖掘出发生在用户身上的故事。

社交多媒体数据具有数量庞大、质量不确定、内容冗余、多样、带有丰富的故事性和时空上下文信息等特点，对相关的处理算法提出了更高更复杂的要求。因此，对于社交多媒体数据的研究已经成为计算机视觉领域的热点问题，相关的成果不仅有利于推动计算机视觉以及多媒体领域相关课题的创新，对于用户体验的提升以及工业界的发展也具有重要的应用价值和现实意义。

1.2 社交多媒体数据研究的关键问题

社交多媒体数据的研究涉及图 1.2 所示内容分析层的语义理解和应用层的关联表达两个大的方面。

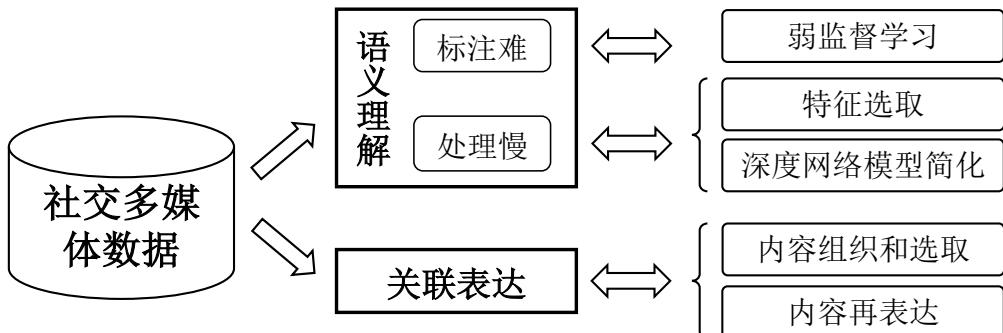


图 1.2 社交多媒体数据研究的关键问题

其中，语义理解面临的主要问题包括：

- 标注难的问题。机器学习算法可以归纳为有监督学习、半监督学习和无监督学习三个类别。有监督学习通常能够获得最好的学习效果。然而，它依赖大量准确的标注数据，在当前大数据的背景下有监督学习的应用成本十分高昂。据了解，目前最大的有标注图片数据集 ImageNet^[2] 花费了大约 25,000 名用户一年左右的时间完成标注。尽管如此，ImageNet 仅包含 22,000 个类别，与现实应用中的语料库相差甚远（如 WordNet¹）。与之相反，半监督学习和无监督学习不需要大量的标注数据，但是学习的效果与有监督学习还有明显的差距。基于以上问题，越来越多的研究人员将注意力投向了弱监督学习上 (Weakly Supervised Learning)。弱监督学习是指从标注不完备、不准确的噪音数据中，充分挖掘有价值的信息，滤除或抑制错误信息，达到学习模型的目的。因此，弱监督学习能够充分利用有噪音的标注数据，解决有监督学习标注难和无监督学习效果差的问题。
- 处理慢的问题。社交多媒体数据的规模十分庞大，对于模型的复杂度以及硬件的计算能力都提出了很高的要求。此外，移动设备的计算能力、存储空间以及电池容量依然有限，提高社交多媒体数据的处理速度对于提升移动端的用户体验也至关重要。为了解决处理慢的问题，一方面可以利用特征选取减少特征提取的种类和数目。图片和视频的内容既包括传统的全局特征、局部特征，也包括近年来提出的深度神经网络产生的特征。对于不同的任务，某些特征具有很强的表征能力，某些特征则十分冗余。因此，选

¹<http://wordnet.princeton.edu/>

取对具体任务最紧凑、最具有表征能力的特征作为数据内容的表达，既可以减少特征提取的种类和数目，也可以减少后续模型学习的计算量。另一方面，可以对特征提取过程中用到的模型进行简化，减少每种特征提取的时间开销。例如，近年来深度神经网络在目标识别、物体检测等领域取得了非常好的效果，但是网络的深度和参数数目也在不断增加，如何在不影响模型准确度的情况下简化深度神经网络成为了当前研究的热点问题。

社交多媒体数据关联表达是指根据用户个性化的需求，从社交多媒体数据中选取有关联的数据，并以一定的表达形式将这些关联数据呈现给用户。它面临的主要问题包括：

- 内容组织和选取。当前，社交多媒体数据以碎片化的形式，按照用户拍摄或上传的时间顺序存储在云端服务器上，社交网站以及搜索引擎根据用户提供的标签对数据进行索引、查询和检索。然而，社交多媒体数据存在质量不确定、内容冗余多样以及故事性等特点，高效的内容组织和选取需要理解数据之间的关联性和故事性，从时间、位置、用户、内容、关联性等多个维度对数据进行组织和选取。
- 内容再表达。内容组织和选取是将数据高效地组织在一起，并选取最能满足用户需求的数据。在实际应用中，需要在原始社交多媒体数据的基础上以一种新的、富有艺术美感的形式将数据重新呈现给用户。例如，Magisto² 系统能够给用户的照片和视频加上丰富的特效，并将视频和音乐的节奏进行匹配，生成类似专业编辑人员编辑的具有丰富表现力的音乐视频。专业的编辑人员在视频编辑中根据素材的内容以及需要表达的效果选取与之相适应的素材和特效对视频进行编辑。对于计算机，如何将专业编辑人员在编辑中运用到的规则和语法转化成可计算的规则和算法是社交多媒体关联表达面临的一个主要问题。由于表现形式以及编辑语法的多样性和主观性，挖掘和应用可计算的编辑语法也具有非常大的挑战性。

1.3 本文主要工作

针对1.2节中提到的关键问题，本文分别对社交多媒体数据的语义理解和关联表达做了深入的研究，构成了社交多媒体数据挖掘和利用的一个相对完整的框架。图 1.3给出了本文研究的具体内容以及相互之间的关联。针对语义理解标注难的问题，弱监督深度学习直接从不准确标注的社交多媒体数据中学习语义

²<http://www.magisto.com>

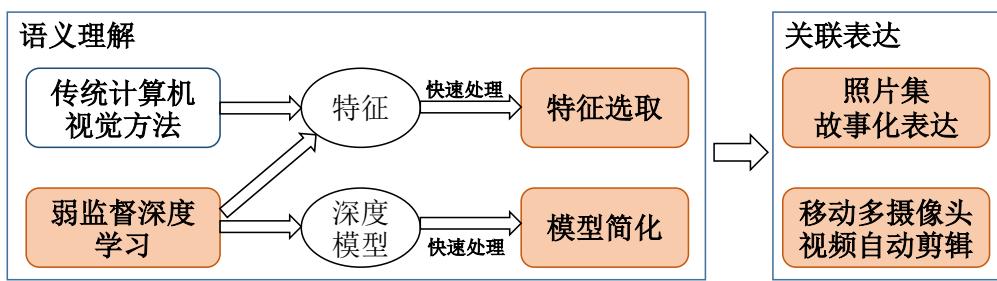


图 1.3 社交多媒体数据的研究内容和相互关联

注：橙色部分表示本论文的研究工作

模型，理解数据的内容。结合传统计算机视觉方法和弱监督深度学习得到的特征，特征选取针对具体的任务选取最紧凑、最有代表性的特征来表征数据内容，减少特征提取的种类和数目，加快大规模社交多媒体数据处理的速度。此外，弱监督深度学习需要耗费大量的计算资源，本论文结合特征提取算法，提出了深度卷积神经网路模型简化算法，减少网络的参数数目和计算时间。对于社交多媒体数据的关联表达，本论文从基于主题的照片集故事化表达和移动多摄像头视频自动剪辑两个方面做了具体的应用研究。其中，基于主题的照片集故事化表达分析检测照片集中的事件，根据照片的质量和关联选取有代表性的照片、并通过可计算的视频编辑语法，对照片集进行故事化的表达。移动多摄像头视频自动剪辑将同一时间段内同一地点不同用户拍摄的多摄像头视频在时间上进行同步，通过可计算的视频编辑语法选取镜头和录音，将多摄像头视频剪辑成单一高质量的音视频流。

基于以上研究点，本论文具体的研究内容包括：

1. 弱监督深度学习算法研究。针对任意语义类别，从社交多媒体获取弱标注数据，提出具有抗噪效果的语义理解模型。传统深度学习算法对于噪声的敏感性是由于所有的数据在学习过程中具有相同的权重。本论文提出方法的基本假设是：同一类别下，正确标注的数据由于语义上关联，在特征空间上比较接近，而错误标注的数据则与其他数据有很大差异。因此，可以利用数据在特征空间上的关系，使得不同数据在训练过程中有不同的贡献加权，使得特征空间上的“孤立”数据具有较小的权重，特征空间上密集区域的数据具有较大的权重，我们称之为相关反馈。该思想可以在学习过程中通过特征空间的低秩近似实现。为了加速模型的训练速度，我们进一步对模型进行了简化和近似，降低了模型的复杂度，用于学习大规模社交多媒体数据的语义理解模型。
2. 大规模高维特征选取算法研究。多媒体数据的特征表示不仅包括高层次的神经网络特征，还包含低层次的全局特征（如颜色特征、纹理特征），局部

特征（如 SIFT^[3]、SURF^[4]）以及通过局部特征描述整体视觉信息的词袋特征等。实际应用中需要根据需求选取对目标任务最有用的特征子集，这对于大规模社交多媒体数据的处理速度以及移动设备有限的计算能力和内存空间尤其重要。此外，去除特定任务不相干的特征，还可以提高特征的表达能力。本论文利用二阶在线学习算法，基于特征的置信度进行特征选取，并利用最大/最小堆结构提出从大规模高维数据中选取特征的高效算法。已有的稀疏在线特征选取算法的复杂度与特征的维度成正比，本论文进一步利用置信度的单调递减特性提出了快速二阶在线特征选择算法，将算法的复杂度降为与非零特征数目成正比。

3. 深度卷积神经网络模型简化算法研究。深度卷积神经网络的深度和模型参数通常比较大，例如经典的 VGG16 网络包含超过 138M 的模型参数。大量的模型参数意味着在实际应用中需要大量的计算资源和时间，极大限制了深度神经网络在大规模社交多媒体数据相关任务上的应用。此外，深度网络在移动设备上的应用已经成为一种趋势。由于移动设备计算能力、存储空间和电池容量的限制，在不影响模型准确率的条件下简化深度网络模型已经成为迫切的需要。本论文提出一种基于在线特征选取的模型简化算法。算法主要针对卷积层进行简化，对卷积层输出的 e 特征图增加对应每个通道的权重层，并利用在线特征选取算法对每个通道对应的权重进行更新和选取。区别于传统方法，在线特征选取方法可以在训练过程中动态调整需要保留的卷积核，减小模型简化对网络性能的影响。
4. 基于主题的照片集故事化表达算法研究。照片集故事化表达首先对照片集进行事件检测，找出用户所拍摄的不同事件，并选取一部分照片子集代表整个照片集。其次，利用弱监督深度学习算法进行照片内容分析。由于深度网络的每一层都是对照片不同层次的语义表达，本论文利用在线特征选取算法选取最能表征照片语义的特征子集，构成照片内容的最终表达。为了达到更好的关联表达效果，不同主题的照片需要采用不同的编辑风格。本论文从网络上抓取训练图片，利用弱监督深度神经网络得到主题风格的分类器，为照片集中的照片选取合适的编辑风格。最后，通过可计算的视频编辑语法，对照片进行动画特效处理以及音乐的匹配，生成具有关联表达能力的视频呈现给用户。
5. 移动多摄像头视频自动剪辑算法研究。多摄像头视频是指在同一时间段、同一地点由不同摄像头拍摄的时间上有重叠的一组视频。多摄像头视频是不同用户从不同的角度对相同事件的记录。本文提出一个全自动移动多摄像头视频自动剪辑系统。我们首先邀请专业的视频编辑人员探讨可计算的

视频编辑语法。根据这些语法，自动剪辑系统首先对音频流进行质量评估，在保证尽可能减少音频切换次数的条件下选取高质量的音频片段，形成单一音频流。对于视频流，系统首先根据音频的节奏以及语义特性选取视频镜头切换点。其次，对多摄像头视频进行语义分割，得到视频子镜头，并对这些子镜头的视觉质量、相机运动以及相互之间的多样性进行评估。在保证镜头运动一致性的前提下最大化质量和多样性，选取视频镜头，得到单一视频流。单一的音频流和频流通过混流得到剪辑好的视频呈现给用户。

1.4 本文主要创新点

本论文的主要创新有以下几点：

1. 针对社交多媒体数据标注难的问题，提出了一种可以从社交多媒体数据中学习目标识别模型的弱监督相关反馈深度神经网络学习算法，摆脱了对大量标注数据的依赖。该算法利用数据之间的关联，使得不同数据在模型训练中有不同的贡献加权。同时，通过对模型的简化和近似，降低了模型复杂度，对于训练大规模社交多媒体数据具有重要意义。
2. 针对大规模社交多媒体数据处理慢的问题，提出了从大规模高维数据中选取特征的高效算法。相比于已有的批处理算法和在线特征选取算法，该算法显著降低了特征选取的时间复杂度，并能达到与已有特征选取算法差别不大甚至更好的准确率。
3. 针对深度网络处理慢、耗费计算资源的问题，提出了简化深度卷积神经网络的高效算法，将传统的多维卷积核组稀疏优化问题转化为一维特征选取问题，在不影响模型准确率的情况下极大减少了模型参数。
4. 提出了基于主题的照片集故事化表达系统——Monet，从照片集中自动检测事件，选取有代表意义的照片，选取合适的视频编辑风格，并将可计算的频编辑语法应用到照片集的编辑中。
5. 提出了移动多摄像头视频自动剪辑系统——MoVieUp，自动剪辑移动多摄像头视频并生成高质量的单一音频视频流。该方法首次考虑了音频流的剪辑，并且首次系统地讨论了视频编辑理论在移动多摄像头视频自动剪辑中的应用。

1.5 本文结构安排

本论文主要研究社交多媒体数据语义理解和关联表达中的几个关键问题：弱监督深度学习、特征选取、模型简化、基于主题的照片集故事化表达和移动多摄像头视频自动剪辑。各章节内容安排如下：

- 第2章从弱监督学习、特征选取、模型简化和关联表达四个方面介绍本论文相关工作的研究现状和工作基础。在弱监督学习方面，主要介绍了传统的弱监督学习方法和近年来热点研究的弱监督深度学习方法；在特征选取方面，主要回顾了传统的批处理方法，用于解决大规模流数据的在线学习方法和在线特征选取算法；在模型简化方面，主要介绍与深度神经网络相关的模型简化工作；在关联表达方面，主要介绍学术界和工业界关于社交多媒体数据的研究成果和代表性应用产品。
- 第3章重点介绍弱监督相关反馈深度神经网络的设计思路和实现方法、网络模型的简化和近似策略以及相应的实验结果。
- 第4章介绍大规模社交多媒体数据快速处理方法，包括在线特征选取的基本模型、本文提出的置信度加权二阶在线特征选取方法以及相应的快速算法。此外，该部分还介绍深度卷积神经网络模型简化的问题建模和具体算法，以及相应的实验结果。
- 第5章介绍基于主题的照片集故事化表达系统，包括照片集的事件检测、代表性照片选取以及照片的质量评估、多样性评价和均衡性考量。系统为选取的代表性照片上分配不同的主题风格，并利用针对不同风格设计的可计算的编辑语法，对照片赋予丰富的特效，生成具有表现力的视频，重现照片集中的场景和故事。
- 第6章介绍移动多摄像头视频自动剪辑系统。首先介绍针对移动多摄像头视频自动剪辑的可计算的视频编辑语法，并提出相应的系统框架。其次，该部分详细介绍了音频的质量评估和剪辑方法、基于音频节奏和语义内容的视频切换点检测算法和基于镜头质量、多样性和运动连续性的镜头选取算法。最后，通过实验证明了系统各个部分的有效性和整体的用户体验。
- 第7章对全文进行总结，并展望未来可以进一步开展和改进的工作。

第2章 国内外研究现状和工作基础

本章对社交多媒体数据语义理解和关联表达涉及的关键问题的研究现状和工作基础做详细回顾。首先回顾弱监督学习，总结近年来弱监督学习和弱监督深度神经网络的发展情况；然后针对社交多媒体数据的特征选取，回顾传统的批处理方法、解决大规模流数据的在线学习方法、在线特征提取算法以及它们在解决大规模社交多媒体数据特征选取问题中的不足；在模型简化方面，主要介绍与深度神经网络相关的模型简化工作；对于关联表达，本章从照片集和移动多摄像头视频两个角度分别介绍学术界和工业界的研究成果和代表性应用产品，指出它们在社交多媒体数据关联表达中的不足，引出本论文提出的基于主题的照片集故事化表达系统和移动多摄像头视频自动剪辑系统。

2.1 弱监督学习

弱监督学习是指训练数据标注不完备或包含噪音条件下的学习问题。本论文主要针对弱监督目标识别问题做工作总结和算法创新。具体的方法可以分为两个方面：数据去噪和噪音鲁棒模型。

2.1.1 数据去噪

数据去噪是指找出并移除可能错误标注的数据。数据去噪的优点在于不依赖目标任务的模型和训练方法，但同时也面临区别噪音数据和异常数据的难题^[5]。数据去噪方法会导致两种类型的错误：正确标注的样本被误判为错误标注并被丢弃，以及错误标注的样本被漏判。

经典的数据去噪方法基于从数据本身提取的特征进行判别。Brodley 等人提出的交叉过滤法^[6]采用类似于交叉验证的思路，将训练数据分成 n 等份，并选择 m 种分类算法（称为过滤算法）。对于每份数据，在剩下的 $n - 1$ 份数据上训练 m 个分类模型，然后用得到的 m 个分类器对这部分数据进行预测，最后判定错误标注的数据并将之移除。此外，Zhou 等人提出最大边界难分类样本学习算法迭代地学习正负类别分界面，保留难分类数据，移除易分类数据，通过更新正负样本集的方法达到去除噪音数据的目的^[7]。

交叉过滤法和最大边界难分类样本学习算法属于有监督学习方法。研究人员也提出了半监督噪音数据去除方法——核平均算法^[8,9]。半监督学习方法首先

人工标注部分数据作为种子数据，记为 $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 。剩余数据记为 $\mathcal{U} = \{\mathbf{z}_j\}_{j=1}^n$ 。通常情况下， $m \ll n$ 。核平均算法的核心思想是对未标注的数据加权，使得加权后未标注数据的分布与有标注数据的分布相同。

此外，社交多媒体数据具有丰富的上下文信息，这些上下文信息从侧面反映了数据内容，因此可以用来辅助数据去噪。Schroff 等人提出了一种基于图片上下文信息进行重排序的算法^[10]。算法首先利用图片附近的文本上下文学习标注是否正确的后验概率并进行重排序，然后从基于文本重排序的结果中选择前 n_+ 个图片作为正样本，再从所有其他类别的图片中选择 n_- 个图片作为负样本，再对图片提取视觉特征，训练 SVM 分类器。社交主动学习（Social Active Learning）首先用视觉特征训练分类器，并评价数据对于提升分类效果的信息量，同时利用文本特征评价每个数据标注可靠的置信度，综合考虑信息量和置信度选取数据，迭代地训练模型并选取正确标注的数据^[11]。

可以看出，数据去噪方法分成两步进行：特征提取和数据去噪的模型训练，对于噪声的判别也依赖于人为假设和定义的准则，这些因素导致在真实的社交多媒体数据上，数据去噪方法的效果受多重因素影响，很难达到理想的效果。

2.1.2 噪音鲁棒模型

噪音鲁棒模型是指对数据的标注噪音不敏感或具有抑制作用的模型。在经验风险最小化（Empirical Risk Minimization, ERM）规则和特定的损失函数下，如果模型发生分类错误的概率保持不变并且与标注噪音无关，则认为模型是对标注噪音鲁棒的。研究人员讨论了在特定条件下，理论上是否存在对标注噪音完全鲁棒的模型^[12]。例如，0 – 1 损失函数在均匀标注噪音或者能达到 0 错误率的情况下是噪音鲁棒的^[13,14]。Beigman 等人讨论了非随机噪音模型下的鲁棒模型^[15]，最小平方差损失函数在均匀标注噪音下同样是噪音鲁棒的。其它常见的损失函数，如指数损失函数，对数损失函数，以及 Hinge Loss 都不是噪音鲁棒的。换句话说，大部分常见的机器学习算法都不是完全噪音鲁棒的，但可以在一定程度上减小错误标注的影响，提升模型的鲁棒性。

对于经典的 SVM 分类器，Bunescu 等人提出了稀疏多实例学习算法用于解决训练样本中的噪音数据问题^[16,17]。多实例学习将一组包含正样本的有噪音数据称为一个正样本包，将一组负样本数据称为负样本包，同时假设正样本包中至少包含一个真实的正样本，负样本包中全是负样本。通过对正负样本包采用不同的约束条件和惩罚系数，达到抑制噪音影响的目的。

直推式支持向量机 (Transductive Support Vector Machine, TSVM)^[9,18] 是一种半监督学习算法，该算法同时根据有标注的数据 \mathcal{L} 和未标注的数据 \mathcal{U} 寻找分界

面，并约束最多有 r 个未标注数据被判定为正样本来抑制噪音数据的影响。

逻辑回归是一种经典的概率统计分类模型，得到了十分广泛的应用。然而，逻辑回归算法对于训练数据中的标注噪音十分敏感。Feng 等人提出了一种基于次高斯分布的鲁棒逻辑回归算法^[19]。根据理论分析结果，算法首先去除范数大于一定阈值的数据，在剩余数据中仅最大化前 n_1 个数据的标注和预测结果之间的相关度。此外，在随机噪音 (Noise at Random, NAR) 假设下，研究人员提出用隐藏变量对数据的真实类别、错误标注的转移概率以及测到的标注建模，并通过 EM 算法优化目标结果^[20,21]。

近年来，深度卷积神经网络在目标识别上获得了巨大的成功。卷积神经网络最早于 1998 年被 Lecun 等人应用在文字识别上^[22]。得益于显卡设备的快速发展，卷积神经网络的深度、宽度以及结构不断改进，深度卷积神经网络的识别能力和应用范围都得到了巨大的提升^[23-28]。

然而，深度卷积神经网络需要大量的训练数据，对于数据标注的准确性有很大依赖。如何利用容易获得的弱监督数据训练深度卷积神经网络成为了近年来的研究热点之一。自举深度神经网络是 2014 年谷歌研究团队提出的利用图片特征之间的相似性监督网络学习、抑制错误标注影响的学习方法^[29]。文章认为，如果图片的特征之间具有相似性，预测的结果也应该比较相似。这种相似性被称之为感知连续性。在随机噪声的假设下，该方法在神经网络中加入全连接隐藏层表示真实的类别。为了达到感知连续性的约束，论文提出引入类自适应编码器的方法训练网络。此外，还可以通过约束隐藏层输出的熵最小达到软自举优化，以及约束隐藏层概率最大类别的熵最小达到硬自举优化。Sukhbaatar 等人通过约束隐藏层参数的迹达到感知连续性的优化目标^[30]。Xiao 等人综合考虑了标注可能遇到的随机噪音和非随机噪音，通过隐藏变量对不同噪音类型下的概率建模，用两个神经网络分别学习噪音类型和真实的标注，通过 EM 算法学习整个模型的参数^[31]。以上方法都基于特定的噪音模型。Azadi 等人提出一种辅助图片正则项 (Auxiliary Image Regularizer, AIR) 方法^[32]，根据数据的特征结构，通过组约束使得只有部分数据具有响应，在训练数据中识别出有用的辅助数据，从而更好地训练神经网络。某种程度上，可以认为辅助图片正则项方法是在训练数据中寻找最近邻数据，减少深度模型对噪音数据的拟合。

传统的噪音鲁棒模型仍然基于特征提取和模型训练两个步骤，具有2.1.1节提到的局限性。现有的弱监督深度学习方法基于特定的噪音模型，难以处理真实的场景，或建模过于复杂，难以训练。因此，本论文提出一种新的利用数据在特征空间的相关性作为相关反馈的弱监督深度神经网络，并对网络进行简化和近似，提高模型的适用性和实用性。

2.2 特征选取

数据的特征表示不仅包括高层次的神经网络特征，还包含低层次全局特征（如颜色特征^[33]、边缘特征^[33]、纹理特征^[34]）、局部特征（如 SIFT^[3]、SURF^[4]），以及通过局部特征描述整体视觉信息的词袋特征等^[35]。实际应用中需要根据需求选取对目标任务最有用的特征子集，这对于处理大规模社交多媒体数据以及计算能力、内存和电量都十分受限的移动设备尤其重要。此外，去除特定任务不相干的特征，还可以提高特征的表征能力。特征选取在机器学习和数据挖掘领域得到了广泛的研究，从处理数据的方法上可以将特征选取算法分成两大类别：批处理方法和在线特征选取。

2.2.1 批处理方法

批处理方法是指训练过程中每次迭代都需要考虑所有的训练数据，可以分为三个类别：

- 过滤法 (Filter)。过滤法分析特征之间的关联、距离、交互信息熵等，选取最有代表意义的特征子集^[36-38]。Yang 等通过分析指出，传统的过滤法存在单调性的问题，不同大小的特征子集之间存在单调的包含关系，这种包含关系在实际情况中并不成立^[39]。他们对特征之间的联系建模，提出了一种多核学习的方法。
- 包装法 (Wrapper)。包装法使用预先定义的分类器评价特征子集的性能^[40]。这类方法迭代地选取不同的子集，并得到该子集在分类器上评价结果，虽然包装法能够获得该分类器上最好的特征子集，但计算过程过于复杂，因此对于该类方法的研究相对较少。
- 嵌入法 (Embedded)。嵌入法将特征选取与模型训练融合，是一种综合平衡过滤法的高效率和包装法的高准确性的方案^[41,42]。

2.2.2 在线特征选取

批处理方法的缺点在于需要将所有训练数据都加载到内存中。对于大规模高维数据，这类方法的局限性十分明显。此外，批处理方法假设数据预先已经全部存在，而实际场景中数据大多是流媒体数据。因此，随着近年来数据量的增大和维度的增加，大量的工作转向了在线学习。最早的在线学习算法是 1958 年提出的感知机算法^[43]。2006 年，Crammer 等人在感知机算法上约束型每次更新后

在当前数据上能获得正确的预测结果^[44]。考虑到批处理学习算法中，二阶海森矩阵能够显著提高算法的收敛速度，Crammer 等人假设模型参数服从高斯分布，用协方差矩阵表示当前模型对于参数的不确定性，提出了置信度加权的在线学习算法^[45]。该算法每次更新时约束更新后的模型以一定概率在当前的数据上获得正确的预测结果。自适应的置信度加权在线学习算法在此基础上降低了模型对于数据噪音的敏感性^[46]。

在线学习也被应用到特征选取上。Langford 等人提出的稀疏在线学习算法在感知机模型上增加了模型参数的 L_1 范数作为正则项，学习稀疏在线模型^[47]。Duchi 等提出的 FOBOS 算法将稀疏在线学习分成两步，第一步是传统的在线学习，第二步是约束模型的 L_1 范数同时使得模型尽可能接近第一步得到的参数^[48]。另一种稀疏在线学习思路是优化模型在主空间和对偶空间的距离，利用模型的 L_1 范数实现稀疏优化而提出的 RDA 算法^[49]。RDA 算法在高稀疏度下往往能获得更好的效果。受置信度加权等二阶算法的启发，Duchi 等利用梯度的协方差矩阵构建二阶信息，提出了自适应的二阶 FOBOS 算法和二阶 RDA 算法^[50]。

基于 L_1 范数的稀疏在线学习算法是针对特征选取的“软”约束方法。参数设定与目标特征数目之间没有确定的关联。基于 L_0 范数的在线特征选取也得到了研究人员的关注。Wu 等提出的在线特征流学习算法能够在每次迭代后返回一个模型和它选取的特征子集^[51]。该算法每次获得所有数据的某个特征，算法按顺序依次处理所有特征。另一种更普遍的应用场景下，算法每次获取一个数据的部分或者所有特征，算法按时序依次处理所有数据。该场景下，Huang 等提出一种无监督在线特征选取算法^[52]。Wang 等利用有监督的数据，根据权重向量的绝对值在线选取特征^[53]。

批处理方法的主要问题在于不具有可伸缩性，以及对于流数据不具有很好的应对能力。基于 L_1 范式的稀疏在线学习方法不能直接约束特征数目，在实际应用中需要根据不同的数据反复调整参数选取预期数目的特征。当前的在线特征选取算法根据权重向量的绝对值选取特征，其结果与批处理方法还有较大差距，并且算法的复杂度仍然较高。因此，本论文提出了大规模高维特征选取算法，不仅显著减小了计算复杂度，还能达到批处理方法相近甚至更好的准确率，对于处理大规模社交多媒体数据具有非常大的应用价值。

2.3 模型简化

近年来，深度卷积神经网络在目标识别、物体检测等领域获得了巨大的成功。为了进一步提高网络的表征能力，研究人员不断改进网络的深度^[24]、宽度^[54]以及拓扑结构^[25,28,55]。然而，大量的网络参数也要求大量的时间开销和计算资源，

同时也极大地限制了深度网络在计算能力、存储空间和电池续航受限的移动设备上的应用。如何在不影响网络性能的情况下减少网络参数成为了当前研究的热点问题。深度网络模型简化相关的工作可以分为三个类别：矩阵分解、量化以及稀疏优化。

矩阵分解利用参数之间的相关性，对参数矩阵做低秩分解，减少网络参数的个数。Denil 等人提出将参数矩阵分解成两个低秩矩阵的乘积，其中一个矩阵作为特征空间的一组基，并提出了基向量字典的构建方法^[56]。Denton 等人提出在网络的预测阶段，对参数矩阵做奇异值分解，如果参数矩阵的奇异值迅速下降，则参数矩阵能够被前 K 个最大奇异值及对应的奇异向量很好地近似^[57]。Rigamonti 等人提出对每个通道的卷积核用秩为 1 的矩阵近似，减小计算量^[58]。Jaderberg 等人在矩阵分解的基础上进一步利用通道之间的冗余信息，将原始的卷积操作分解成两步卷积运算^[59]。Ioannou 等人和 Tai 等人改进并扩展了低秩分解方法，将其用于更大的深度网络^[60,61]。Mamalet 等人将卷积核分解为秩为 1 的向量乘积，并与后续的池化 (Pooling) 操作融合为一层卷积运算，减少运算量^[62]。

量化是指利用较少的比特数表示网络参数，减少模型的大小和乘法运算的复杂度。当前网络通常采用 32 比特的浮点数表示网络参数。研究表明，网络参数可以用更少的比特数表示。例如，Hwang 等人和 Arora 等人提出仅用 $+1, -1, 0$ 三个数值表示网络参数并训练卷积神经网络^[63,64]。Courbariaux 等人和 Rastegari 等人进一步提出用二个数值表示网络参数^[65,66]。Gong 等人提出用向量量化的方法量化全连接层的参数^[67]。针对卷积层的向量量化在 Wu 等人提出的 Q-CNN 网络中得到研究和应用^[68]。Anwar 等人用最小平方差方法量化网络^[69]。Chen 等人利用哈希函数随机将网络参数分组，达到量化的目的^[70]。

当前网络的参数矩阵是密集矩阵，稀疏优化的目标是使得最终的参数矩阵稀疏，达到模型简化的目的。区别于参数矩阵低秩分解，Liu 等人提出对卷积核做稀疏分解，并提出了高效的稀疏矩阵相乘算法^[71]。受 L_1 范数和 L_2 范数约束的启发，Han 等人提出重复交替进行删除神经元之间连接和重新训练精简后网络的模型简化方法^[72,73]。然而，这些稀疏方法产生的稀疏网络不是结构化的，运算时会导致无规则的内存访问，不能带来实际的运算加速。Li 等人根据卷积核的绝对值之和去除部分卷积核，达到运算的加速^[74]。Murray 和 Chiang 运用结构化稀疏方法约束隐藏层神经元的个数^[75]。Anwar 等提出了卷积核、通道以及卷积核内部的结构化稀疏方法^[76]。他们还提出用粒子滤波器 (Particle Filter) 衡量网络连接的重要性，从而优化网络结构。Wen 等人系统讨论了结构化的稀疏算法，从卷积核、通道、卷积核形状、深度四个方面对网络进行结构化约束，不仅达到了减少网络参数的目的，还获得了实际运算速度上的提升。Hu 等人通过研究发现，大网络部分神经元的响应大部分情况下为 0，且与网络的输入信号无关。因此，

他们通过分析网络神经元在大数据集上的响应去除部分神经元^[77]。Soravit 等人提出了通道稀疏连接方法，在保持与其他结构化方法同样计算速度的情况下获得了很好的效果^[78]。

以上介绍的模型简化方法，虽然取得了一定的效果，但同时也存在很多的问题。矩阵分解方法对于全连接层以及大卷积核操作具有非常好的效果，然而最新的网络更倾向于使用更少的全连接层，并通过级联小卷积核的方法达到大卷积核相同大小的感知野 (Receptive Field)，不仅减少了运算量，还提高了网络的表征能力^[26]。量化方法需要特定硬件或软件库的支持才能显著提高运算的速度。非结构化的稀疏优化方法对于减少参数数目作用比较明显，对于计算速度的提升十分有限。结构化的方法一般基于参数的绝对值决定参数的重要性，具有一阶在线特征选取方法相同的缺陷，组稀疏优化方法增加了网络优化的难度。为此，有必要提出一种新的模型简化方法，既能保证简化后网络的表征能力，减小参数规模，提升网络运算效率，又易于优化，提高模型简化的可操作性。

2.4 社交多媒体数据的关联表达

本节从基于主题的照片集故事化表达和移动多摄像头视频自动剪辑两个方面回顾社交多媒体数据关联表达相关的工作。

2.4.1 基于主题的照片集故事化表达

照片集关联表达涉及事件检测、关键照片选取和故事化表达等多个方面。通常，用户照片包含拍摄的时间和位置信息，可以用来检测照片集中记录的事件。Platt 等人提出用一个小时或者自适应的阈值作为相邻事件之间的时间间隔^[79]。Graham 等人扩展了该方法，使用事件聚类的类内拍照频率和类间时间间隔调整已有的事件划分^[80]。Gargi 提出将拍摄频率急速增加的时间点作为事件的起点，将长时间间隔没有拍摄行为作为事件的终点^[81]。Matthew 等人将可信度、动态规划和贝叶斯信息准则 (Bayes Information Criterion, BIC) 运用到照片的相似度矩阵，检测事件的边缘位置^[82]。一般来说，事件检测问题可以表示为一个聚类问题。Loui 和 Svakis 提出用两类的 K-means 聚类算法将照片分组，并检查照片之间颜色的相似性改进聚类结果^[83]。Gong 等人利用层次聚合聚类算法将照片分配到不同的聚类中心^[84]。Platt 等人用隐马尔科夫模型聚类^[79]。Mei 等人在时间、位置以及内容特征上利用混合高斯模型解决事件检测问题^[85]。Xu 等人进一步利用纹理和深度特征改进了该算法^[86]。

近年来许多研究工作和产品相继出现，用以解决关键照片选取问题。在学术

表2.1 照片集关联表达系统比较

	Magisto	Animoto	Google+		Magisto	Animoto	Google+
物体识别	-	-	+	设计风格	+	-	-
人脸检测	+	-	+	相机运动	+	-	-
音乐分析	+	-	-	场景分析	+	-	+
照片分析梳理	-	-	+	色彩调整	+	+	+

界，关键照片选取主要依赖照片的代表性^[82,85-87]。Cooper 等人将事件中第一张照片作为关键照片。Mei 等人选择具有最大后验概率的照片^[85]。Chu 等人提出在照片的聚类中，根据相似图片对之间的相互关系选取关键照片^[87]。Xu 等人根据事件的重要性引入了照片的受欢迎程度 (popularity) 以及事件内部的相似度决定关键图片^[86]。工业界的在线服务，如 Microsoft Onedrive¹、Google²可以在一定程度上对照片集进行事件检测和照片选取，但缺乏对数据的故事化表达。

照片集故事化表达一直以来受到了工业界和学术界的共同关注。例如，Magisto³和 Animoto⁴是两个可以根据用户提供的照片生成音乐视频的在线服务。然而，它们依赖用户主动选取和提供的照片，不能直接从照片集中总结并整理出故事呈现给用户。此外，用户需要手动指定音乐视频编辑的风格。在学术界，Hua 等人提出的 Photo2Video 系统是从照片生成音乐视频的先驱性工作^[88]。该系统利用相机运动将静态照片转换成运动片段，并通过转场效果以及与音乐节奏的匹配生成最终视频。然而该系统采用的编辑风格和编辑效果比较单一。其他系统如 Tiling SlideShow 将照片和背景音乐同步，并以贴片幻灯片的形式播放^[89]。Kuo 等人提出的 Sewing Photos 系统专注于解决播放照片幻灯片时平滑的转场效果^[90]。Sewing Photos 和 Tiling SlideShow 也存在 Photo2Video 同样的编辑风格单一的问题。

以上工作没有系统地对照片集进行总结整理，在表达时很少运用丰富的视频制作特效、编辑风格和编辑语法，因而对于照片集的故事化表达能力十分有限。表格 2.1总结了现有照片集关联表达系统存在的问题。因此，我们需要提出一个能够对照片集进行事件挖掘和关键图片选取，并运用专业的编辑语法按照照片的主题风格故事化表达的系统。

2.4.2 移动多摄像头视频自动剪辑

移动多摄像头视频是指在同一个事件中，由多个移动摄像头从多个角度拍摄的，时间上有重叠的一组视频^[91]。随着智能设备的普及与性能的提升，移动

¹<https://onedrive.live.com>

²<https://plus.google.com>

³<http://magisto.com>

⁴<http://animoto.com>

表 2.2 移动多摄像头视频自动剪辑系统比较

	diversity	shakiness	tilt	occlusion	audio mashup	cut point
VD ^[91]	Yes	Yes	No	No	No	Manual
Jiku ^[94]	Yes	Yes	Yes	Yes	No	Learning

注：VD 是 Virtual Director^[91] 的简称。

多摄像头视频自动剪辑成为了近年来的热点问题。

移动多摄像头视频自动剪辑的方法可以总结为三个类别：基于规则的方法^[92]、基于优化的方法^[91] 和基于学习的方法^[93,94]。基于规则的方法模仿专业人员的编辑过程。然而，视频剪辑的过程带有很强的主观性和倾向性，而并非固定的编辑规则。Shrestha 等人提出了从视频质量、多样性和切换点合适度选取镜头的优化算法^[91]。然而实际系统并没有提出切实可行的切换点合适度评价方法，仅仅考虑了视频质量和多样性。在视频质量评估中，该系统也没有考虑移动视频中的倾斜和遮挡问题。此外，该系统通过贪心算法解优化方程，仅获得了目标方程的局部最优解。Nguyen 等人提出 Jiku Director 系统用于解决在线移动多摄像头视频自动剪辑问题^[93,94]。该系统通过学习隐马尔科夫模型完成镜头选取和镜头长度选取。通过这种方式学到的模型与内容无关，而实际中，镜头角度和长度的选取都是和内容密切相关的，并且受运动强度、音乐节奏以及其它因素的影响^[95]。此外，由于不能准确判断视频的角度，该系统不能做到全自动视频剪辑，尤其是在模型的训练阶段该系统需要人工干预。Arev 等人提出的多摄像头视频自动剪辑系统^[96] 依赖场景的三维重建，不适用于移动多摄像头视频。

以上系统均没有考虑音频剪辑，而完整的视频由音频流和视频流两部分构成，高质量的音频流对于提升用户体验具有十分重要的作用。表格 2.2 比较了现有移动多摄像头视频剪辑系统的特点和问题，

移动多摄像头视频自动剪辑还与视频编辑相关，包括视频摘要 (Video Summarization)、镜头选取和家庭音乐视频编辑。视频摘要与视频剪辑的共同点在于它们都要最大化有信息内容的部分。Sundaram 等人提出了从可计算镜头中生成快速概览的实用框架^[97]。可计算镜头的检测通常基于人类记忆的一个仿真模型^[97]。该论文将视觉编辑语法运用到镜头编辑中（选取、缩放、时长、顺序等）^[98]，对于本论文的工作具有很大的借鉴意义。

镜头选取在演讲和会议等诸多特定场景都得到了广泛的研究，通常可以通过识别演讲者或者检测人脸来选取需要展示的镜头内容^[99,100]。Ranjan 等人和 Zhang 等人提出的系统中用跟踪和基于音频定位的方法来选择镜头。以上系统都可以归结为基于音频的方法。在移动多摄像头视频自动剪辑中，音频不是唯一的

关注点，音频定位和人脸检测在嘈杂的拍摄环境和低视觉质量条件下的作用十分有限。

此外，还有大量关于家庭视频或音乐视频编辑的工作。Hua 等人提出了自动家庭视频编辑系统 AVE，从一系列的家庭视频中提取一部分最精彩的镜头^[101]。他们提出了两套规则分别保证对原始视频的代表性，以及音频和视频之间的协调性。类似的方法被拓展到自动音乐视频编辑中，通过分析视频的时序结构匹配音频的节奏^[95]。然而，由于移动多摄像头视频需要进行时间上的同步，并且需要保证内容的质量和多样性，这些系统不能直接用于移动多摄像头视频自动剪辑。

第3章 弱监督社交多媒体数据语义理解

本章研究训练数据标注不准确情况下的弱监督社交多媒体数据语义理解问题。弱监督社交多媒体数据语义理解是分析、挖掘、利用社交多媒体数据的基础。本章主要研究弱监督目标识别问题，首先对弱监督目标识别问题建模，然后依次对本章提出的弱监督相关反馈深度神经网络的构建、简化和近似以及相关反馈分析做详细介绍。最后通过实验验证本章提出的算法对于标注噪音的鲁棒性。

3.1 弱监督目标识别问题建模

在目标识别问题中，给定 N 个训练数据 $X = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in \mathbb{R}^d$ 是训练数据的特征表示， $y \in \mathcal{Y}$ 是数据的类别，通常为离散的整数值， \mathcal{Y} 是类别空间，类别数目为 $|\mathcal{Y}| = K$ 。本章仅讨论单类别分类问题，即每个训练数据有且仅有一个类别。

传统的目标识别问题假设数据的标注 y 是准确无误的，在社交多媒体数据中实际获得的标注 y 和真实的数据类别 z 存在不一致的情况。在统计学上，用一个二值的随机变量 E 表示是否存在标注噪音。数据 X 、真实标注 Z 、实际观测到的标注 Y 和随机变量 E 之间存在图 3.1 所示三种关系^[102]:

- **完全随机噪音 (NCAR)**。噪音 E 独立于其他随机变量，包括真实的标注 Z 。
- **随机噪音 (NAR)**。噪音 E 独立于数据 X ，但依赖真实的标注 Z 。该模型允许非对称噪音，即某些类别的数据更有可能出现标注噪音。随机噪音可等价地表示为一个标注矩阵或转移矩阵：

$$\gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \ddots & \vdots \\ \gamma_{K1} & \cdots & \gamma_{KK} \end{pmatrix} = \begin{pmatrix} P(Y=1|Z=1) & \cdots & P(Y=K|Z=1) \\ \vdots & \ddots & \vdots \\ P(Y=1|Z=K) & \cdots & P(Y=K|Z=K) \end{pmatrix} \quad (3.1)$$

- **非随机噪音 (NNAR)**。上述两种噪音类型均假设标注噪音对于同一类别下的所有样本具有相同的影响。然而在实际场景中，上述假设不一定成立。例如，当样本与其他类别样本之间的距离较近时，更有可能发生错误标注。此外，样本分布密度较低的区域标注的可靠性也比其它区域低。在图 3.1(c) 所示的模型中，标注噪音 E 同时依赖于数据 X 和真实类别 Y ，错误标注

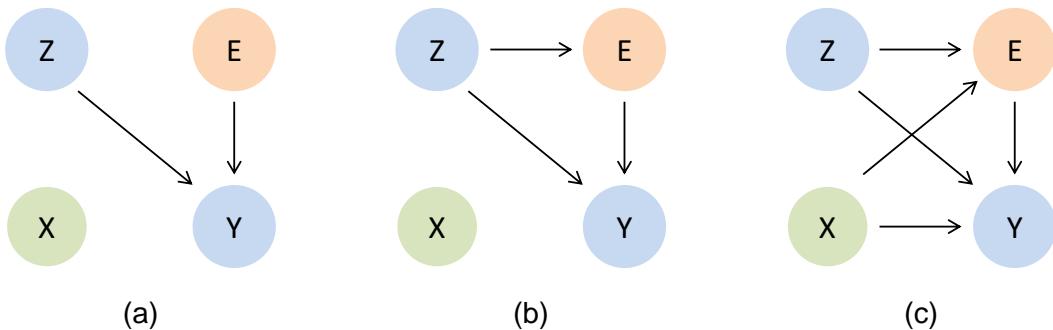


图 3.1 数据标注噪音类型: (a) 完全随机噪音; (b) 随机噪音; (c) 非随机噪音

更有可能出现在某些类别和数据空间的某些区域。非随机噪音是最有普适性的噪音类型。例如，分界面附近和低样本密度分布区域的标注噪音只能用非随机噪音来建模。

3.2 弱监督相关反馈深度神经网络

传统弱监督目标识别算法需要人工设计一系列的特征，如全局特征、局部特征等。这些特征的表征能力直接影响目标识别的效果，不仅提高了研究人员设计特征的难度，也制约了目标识别效果的提高。近年来，深度卷积神经网络在目标识别上获得了巨大的成功。2012年，Alex等人将深度卷积神经网络应用到了百万规模的ImageNet目标识别任务上，提出了AlexNet网络模型，通过5层卷积层和3层全连接层的网络结构，从原始的图片像素中提取从浅层到深层的语义特征并做目标识别^[23]。深度学习的优点在于不需要人工设计特征，网络通过反向传播的方式同时学习特征提取和分类器。然而，经典的深度卷积神经网络对数据标注的准确性有很大的依赖性，如何利用大规模弱监督社交多媒体数据训练深度卷积神经网络成为了近年来的研究热点之一。

3.2.1 经典深度卷积神经网络

如图 3.2 所示，经典深度卷积神经网络的前几层通常是卷积层，最后几层为全连接层，不同的任务可能会采用不同深度的网络。假设训练数据为 $X = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in \mathbb{R}^d$ 表示第 i 个数据, $y_i \in [0, K - 1]$ 是第 i 个数据的标注类别, K 是类别数目 N 是图片总数目。假设网络的层数为 M , 网络参数表示为 $W = \{W^1, \dots, W^M\}$ 。数据在第 m 层的特征图 (Feature Map) 表示为 $Z^m(X) = [\mathbf{z}^m(\mathbf{x}_1), \dots, \mathbf{z}^m(\mathbf{x}_N)]^T \in \mathbb{R}^{N \times d_m}$ 。最后一个全连接层的输出 $\mathbf{z}^M(\mathbf{x}_i)$ 作为 Softmax 分类器的输入，得到所有类别上的概率分布 $\mathbf{p}(\mathbf{z}^M(\mathbf{x}_i))$ 。

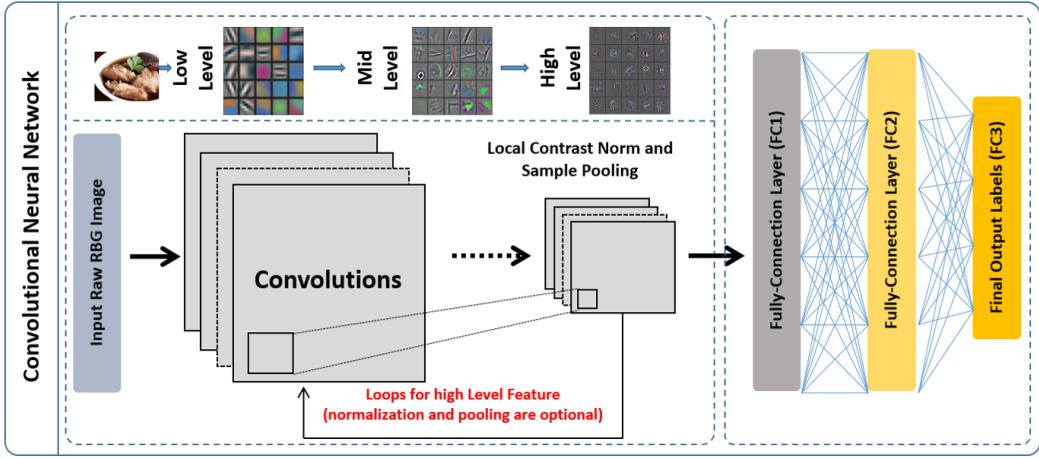


图 3.2 深度卷积神经网络结构图

通常，卷积神经网络的损失函数是 Softmax 函数的负对数似然函数和权重衰减项 (weight decay) 之和：

$$\mathcal{L}(W; X, \mathbf{y}) = -\frac{1}{N} \left[\sum_{i=1}^N \log p(y_i | \mathbf{x}_i; W) \right] + \frac{\beta}{2} \|W\|_F, \quad (3.2)$$

其中， β 是权重衰减项的系数。假设最后一层为全连接层，上述损失函数相对于最后一层特征图 Z^M 以及参数 W^M 的梯度为：

$$Z^M = Z^{M-1} W^M \quad (3.3)$$

$$\frac{\partial \mathcal{L}(W; X, Y)}{\partial \mathbf{z}^M(\mathbf{x}_i)} = -\frac{1}{N} \left(\mathbf{1}_{y_i}(\mathbf{z}^M(\mathbf{x}_i)) - \mathbf{p}(\mathbf{z}^M(\mathbf{x}_i)) \right), \quad (3.4)$$

$$\begin{aligned} \frac{\partial \mathcal{L}(W; X, Y)}{\partial W^M} &= (Z^{M-1})^T \frac{\partial \mathcal{L}(W; X, Y)}{\partial \mathbf{z}^M(\mathbf{x}_i)} \\ &= -\frac{1}{N} \sum_{i=1}^N \mathbf{z}^{M-1}(\mathbf{x}_i) \left(\mathbf{1}_{y_i}(\mathbf{z}^M(\mathbf{x}_i)) - \mathbf{p}(\mathbf{z}^M(\mathbf{x}_i)) \right)^T. \end{aligned} \quad (3.5)$$

其它层参数的梯度通过方向传播 (Back Propagation) 得到^[22]。从上述公式可以看出，错误的标注会导致参数梯度计算错误，并被反向传播，使得经典的深度卷积神经网络对于标注噪音十分敏感。

3.2.2 相关反馈深度卷积神经网络

目前已有部分工作研究弱监督深度学习，然而现有方法大多基于特定的噪音模型，难以处理真实的场景，或建模过于复杂，难以训练。2014 年谷歌研究团队提出了利用数据特征之间的相似性监督网络学习，抑制错误标注影响的学习方法^[29]。文章认为，如果数据的特征之间具有相似性，那么预测的结果也应该比较相似。文章将这种关系称之为感知连续性，并将感知连续性应用到随机噪音

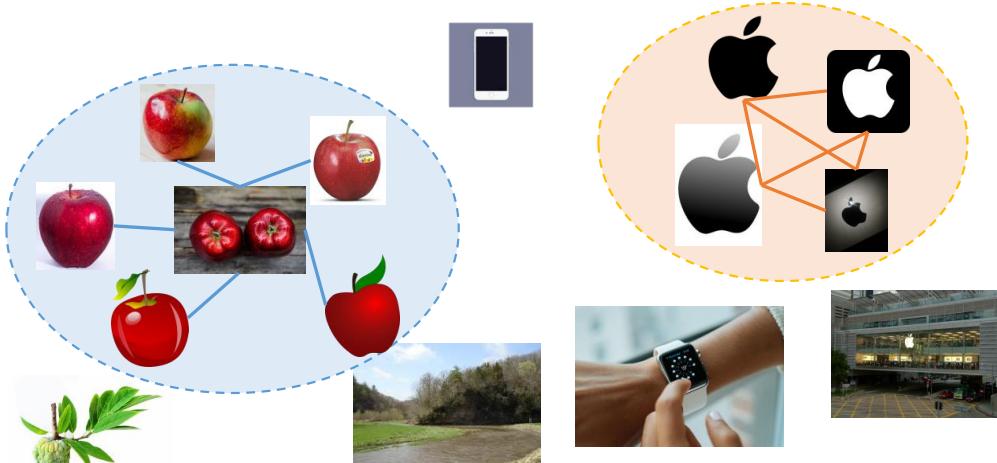


图 3.3 感知连续性示例

假设下的网络学习中。本章在感知连续性的基础上提出了不依赖于特定噪声类型的相关反馈深度卷积神经网络。

基于感知连续性，本章提出方法的基本假设是正确标注样本在数据空间具有相似性，它们的特征在特征空间也具有相似性，而错误标注的样本则不具有这种相似性。因此，可以利用特征之间的相关性作为反馈，使得不同数据在网络训练过程中发挥不同的作用。图 3.3 是在搜索引擎中搜索“apple”的部分结果，虚线框中的数据相互之间比较接近，在训练过程中应该发挥更大的作用，虚线框以外的数据不具有与其他数据的相似性，为噪音标注的可能性较大，在训练过程中应该发挥较小的作用。

为了表示特征之间的关系，我们将网络最后一层的特征转换为能反映特征之间相互关系的关联特征表示 (Affinity Representation)。类似于 Belkin 等人提出的最近邻系统^[103]，我们定义如下相似度矩阵 $S \in R^{N \times N}$ ：

$$S_{ij} = \begin{cases} \exp\left\{-\frac{\|\mathbf{z}^M(\mathbf{x}_i) - \mathbf{z}^M(\mathbf{x}_j)\|^2}{\gamma^2}\right\} & y_i = y_j \\ 0 & \text{otherwise,} \end{cases} \quad (3.6)$$

其中 γ 是尺度因子。为了更好地反映相似度矩阵的局部结构，我们用一个对角矩阵 D 对相似度矩阵进行正则化， $D_{ii} = \sum_{j=1}^N S_{ij}$ 。训练数据最终的特征表示为 $\Psi(X; W) = [\psi(\mathbf{x}_1), \dots, \psi(\mathbf{x}_N)] = D^{-1}S$ ，矩阵 $\Psi(X; W)$ 的每一列包含了数据 \mathbf{x}_i 和其他数据特征之间的关系。

假设理想情况下不受噪音影响的模型参数为 W^* ，噪音鲁棒学习算法应该尽量优化 W 使其逼近 W^* 。该优化目标可以通过最小化特征表示 $\Psi(X; W)$ 和理想情况下的特征表示 $\Psi(X; W^*)$ 之间的差值 E_n 得到。 E_n 是由标注噪音引起的特征

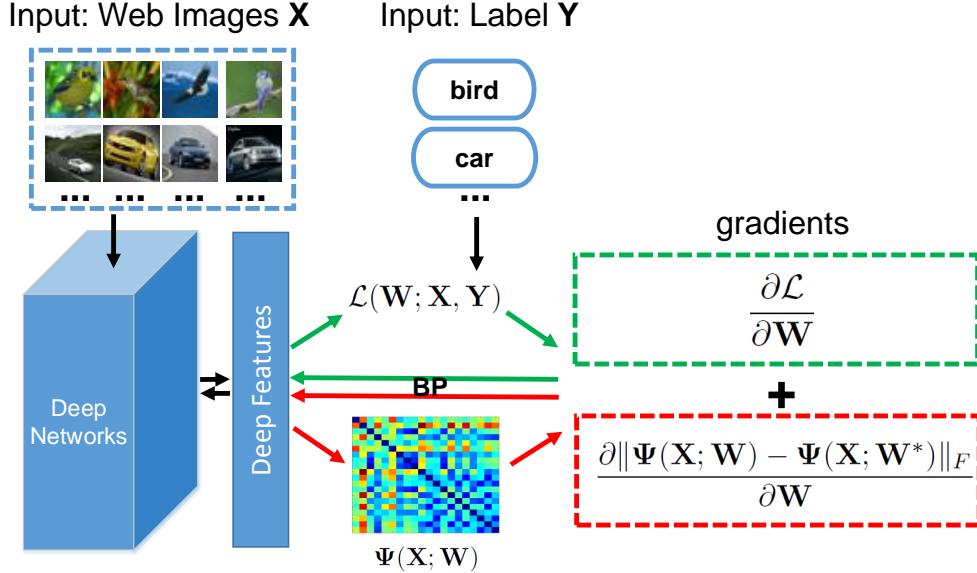


图 3.4 弱监督相关反馈深度卷积神经网络

表示的误差。换句话说，可以认为 $\Psi(X; W)$ 是理想特征与一个加性噪声之和：

$$\Psi(X; W) = \Psi(X; W^*) + E_n \quad (3.7)$$

根据方程(3.7)以及低秩理论^[104]，我们假设 $\Psi(X; W^*)$ 是 $\Psi(X; W)$ 的低秩近似矩阵：

$$\text{rank}(\Psi(X; W)) > \text{rank}(\Psi(X; W^*)) \quad (3.8)$$

在社交多媒体数据上，当类别数目足够多，可以假设错误的标注来自于同一训练数据集上的其他类别，训练数据的特征最多有 K 个模式， $\Psi(X; W^*)$ 的秩等于类别数目 K ，因此， $\Psi(X; W^*)$ 可以通过如下优化方程得到：

$$\min_{\Psi(X; W^*)} \|\Psi(X; W) - \Psi(X; W^*)\|_F, \quad \text{s.t.} \quad \text{rank}(\Psi(X; W^*)) = K. \quad (3.9)$$

如图 3.4 所示，相关反馈神经网络利用训练数据之间的感知连续性，通过特征矩阵的重构误差抑制训练过程中噪音的影响。然而，方程(3.9)带来的计算量极大增加了优化过程的时间开销。此外，该方法优化过程分两步：解优化方程(3.9)和反向传播。为了解决该方法以上缺点，本章进一步提出了改进的简化和近似算法。改进的算法基于下面的命题：

命题 3.1 令 $L = D - S, H^* \in R^{N \times K}$ 由 $\Psi(X; W)$ 的 K 个最大特征值对应的特征向量构成，可以得到：1) 方程(3.9)，即 $\Psi(X; W)$ 的秩为 K 的最佳近似矩阵由特征向量矩阵 H^* 唯一决定；2) H^* 也是下列优化方程的最优解：

$$\min_H \text{tr}[H^T L H] \quad \text{s.t.} \quad H^T H = I. \quad (3.10)$$

由于方程(3.9)和方程(3.10)均在 H^* 取得最优值,可以认为方程(3.9)和方程(3.10)作为惩罚项是等价的。

证明 命题3.1可以通过如下三个定理得到。不失一般性,假设 $\text{rank}(\Psi(X; W)) = r$ 。矩阵的秩为 K 的最小重构误差矩阵可以通过 *Eckart-Young-Mirsky* 定理^[105] 得到:

定理 3.2 (Eckart-Young-Mirsky) 对秩为 r 的矩阵 $P \in \mathbb{R}^{m \times n}$ 进行奇异值分解 (Singular Value Decomposition, SVD) 得到 $P = U\Sigma V^T$, $U^T U = I$, $V^T V = I$, 如果 $K < r$, 则有:

$$\arg \min_{\substack{\hat{P} \in \mathbb{R}^{m \times n} \\ \text{rank}(\hat{P}) = K}} \|P - \hat{P}\|_F = U\hat{\Sigma}V^T,$$

其中 $\hat{\Sigma}$ 是包含 P 的前 K 个最大奇异值的对角矩阵。

此外,如果矩阵 P 是实对称矩阵,它的奇异值和特征值之间具有如下定理所示关系:

定理 3.3 对实对称矩阵 P 特征值分解得到 $P = Q\Lambda Q^T$, $Q^T Q = I$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ 是矩阵 P 的特征值。则有

$$Q = U$$

因此,根据定理3.2和定理3.3,实对称矩阵 $\Psi(X; W)$ 的秩为 K 的最小重构误差矩阵由 $\Psi(X; W)$ 的前 K 个最大特征值对应的特征向量对应的矩阵构成。

定理 3.4 (Rayleigh^[106] 定理) 令 $H^* = [\mathbf{h}_1^*, \dots, \mathbf{h}_K^*] = \arg \min_H \text{tr}[H^T L H]$, 并且 $H^T H = I$, 则最优解 H^* 可以通过求解如下泛化特征值分解问题得到:

$$L\mathbf{h}_i = (1 - \lambda_i)D\mathbf{h}_i,$$

其中 $\{1 - \lambda_i^* | i = 1, \dots, K\}$ 是矩阵 $\Psi(X; W)$ 的前 K 个最大特征值, $H^* = \{\mathbf{h}_i^* | i = 1, \dots, K\}$ 是对应的特征向量。

由于方程(3.9)和方程(3.10)均在 H^* 取得最优值,可以认为方程(3.9)和方程(3.10)作为惩罚项是等价的。 \square

通过以上命题和证明可以发现,方程(3.9)的最优解可以通过优化方程(3.10)中的最小迹得到。因此,我们提出将最小迹的优化目标引入到经典的神经网络中,最终的噪音鲁棒深度网络目标方程为:

$$\tilde{\mathcal{L}} = \mathcal{L}(W; X, Y) + \alpha \text{tr}[H^T L H]. \quad (3.11)$$

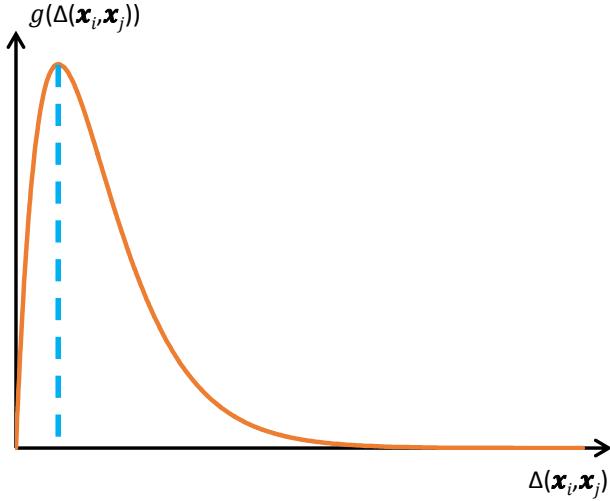


图 3.5 训练数据对于梯度的贡献曲线，横坐标为数据与其他数据特征之间的距离

上述优化方程仍然需要对特征矩阵做特征值分解，本文进一步提出了近似方法。首先构建标注矩阵 $Y = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T \in \{0, 1\}^{N \times K}$ ，每一列 $\mathbf{y}_i \in \{0, 1\}^{K \times 1}$ 表示数据 \mathbf{x}_i 的标注向量，并且只有 y_i 位置为非零值。由于 Y 矩阵是在标注空间上对数据的表述， H 矩阵是在特征空间上的主成分表述，根据感知连续性， H 可以通过如下优化方程近似得到^[107,108]：

$$\min_H \|HH^T - YY^T\|_F^2. \quad (3.12)$$

为了满足 H 矩阵的正交性，一个合理的近似解为 $H = Y(Y^T Y)^{-\frac{1}{2}}$ 。

3.2.3 相关反馈分析

为了验证上述方法有效性，我们从梯度的角度分析相关反馈对于噪声的抑制作用。定义第 M 层特征之间的距离为：

$$\Delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{z}^M(\mathbf{x}_i) - \mathbf{z}^M(\mathbf{x}_j)\|^2 \quad (3.13)$$

当 y_i 与 y_j 相同时， \mathbf{x}_i 和 \mathbf{x}_j 之间的距离 S_{ij} 表示为：

$$S_{ji} = \exp\left\{-\frac{\|\mathbf{z}^M(\mathbf{x}_j) - \mathbf{z}^M(\mathbf{x}_i)\|^2}{\gamma^2}\right\} = \exp\left\{-\frac{\Delta(\mathbf{x}_j, \mathbf{x}_i)}{\gamma^2}\right\} \quad (3.14)$$

S_{ij} 相对于输入 $\mathbf{z}^M(\mathbf{x}_i)$ 的梯度为：

$$\frac{\partial S_{ij}}{\partial \mathbf{z}^M(\mathbf{x}_i)} = \frac{\partial S_{ij}}{\partial \Delta(\mathbf{x}_i, \mathbf{x}_j)} \frac{\partial \Delta(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{z}^M(\mathbf{x}_i)} = \frac{2S_{ij}}{\gamma^2} (\mathbf{z}^M(\mathbf{x}_i) - \mathbf{z}^M(\mathbf{x}_j)) \quad (3.15)$$

定义 $G = HH^T$, 目标方程(3.11)中的相关反馈项对于特征 $\mathbf{z}^M(\mathbf{x}_i)$ 的梯度为:

$$\begin{aligned}
 \frac{\partial \text{tr}(H^T LH)}{\partial \mathbf{z}^M(\mathbf{x}_i)} &= \frac{\partial \text{tr}(HH^T L)}{\partial \mathbf{z}^M(\mathbf{x}_i)} = \frac{\partial \text{tr}(GL)}{\partial \mathbf{z}^M(\mathbf{x}_i)} = \frac{\sum_{k=1}^N \sum_{j=1}^N G_{kj} \partial L_{jk}}{\partial \mathbf{z}^M(\mathbf{x}_i)} \\
 &= \frac{\sum_{k=1}^N \sum_{j=1}^N (G_{kk} - G_{kj}) \partial S_{kj}}{\partial \mathbf{z}^M(\mathbf{x}_i)} \\
 &= \frac{2 \sum_{k=1}^N \sum_{j=k+1}^N (G_{kk} - G_{kj}) \partial S_{kj}}{\partial \mathbf{z}^M(\mathbf{x}_i)} \\
 &= \sum_{j=i+1}^N (G_{ii} - G_{ij}) \frac{2 \partial S_{ij}}{\partial \mathbf{z}^M(\mathbf{x}_i)} \\
 &= \sum_{j=i+1}^N (G_{ii} - G_{ij}) \frac{4S_{ij}}{\gamma^2} (\mathbf{z}^M(\mathbf{x}_i) - \mathbf{z}^M(\mathbf{x}_j)) \quad (3.16)
 \end{aligned}$$

$$\left\| \frac{\partial \text{tr}(H^T LH)}{\partial \mathbf{z}^M(\mathbf{x}_i)} \right\|^2 \propto g(\Delta(\mathbf{x}_i, \mathbf{x}_j)) = \Delta(\mathbf{x}_i, \mathbf{x}_j) \left(\exp \left\{ -\frac{\Delta(\mathbf{x}_i, \mathbf{x}_j)}{\gamma^2} \right\} \right)^2 \quad (3.17)$$

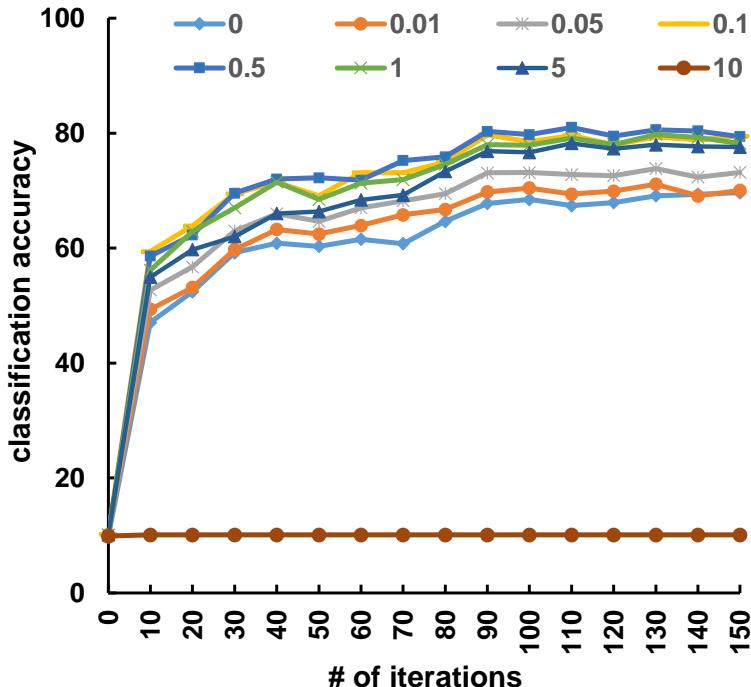
图片 \mathbf{x}_i 对于梯度的贡献是图3.5所示的单峰函数。在峰值的右边，随着 \mathbf{x}_i 和 \mathbf{x}_j 之间距离的增大，反馈的梯度逐渐减小至0，意味着当同一类的样本之间距离过大时，相关反馈的作用较小。在峰值的左边，随着距离减小，反馈的梯度也减小，距离为0时梯度也为0。这是因为深度学习需要多样化的训练样本，距离为0意味着两个训练数据完全相同，反馈的梯度也应该为0。

3.3 实验结果和评估

本节从实验的角度验证本章提出的弱监督相关反馈深度神经网络的有效性。首先在标准数据集上验证该方法对于噪声标注的鲁棒性，其次，我们将该方法用于真实的社交数据集，验证该方法在社交图片标注上的有效性。

3.3.1 目标识别

实验数据：我们在两个公开数据集上分别验证算法对噪声的鲁棒性。一个是Cifar10^[109]，包含10个类别60,000张 32×32 的彩色图片，其中50,000张用于训练，10,000张用于测试。为了产生不同噪声比例的训练数据，在每个类别的训练数据上按照不同比例，随机选取图片，并将他们的类别随机替换为数据集中的其他类别，训练数据集的总图片数目保持不变。在我们的实验设置中，训练数据从无噪声到90%的噪声均匀取10个噪声比例。另一个数据集是PASCAL VOC2007^[110]，包含20个类别总共9,963张图片。我们将数据集随机等分成训练数据和测试数据。

图 3.6 参数 α 对于神经网络分类性能的影响

比较基准: 本章提出的相关反馈神经网络称为 RFCNN, 与以下四个方法进行比较:

- **CNN:** 经典的卷积神经网络。
- **RPCA+CNN:** 训练卷积神经网络之前, 首先用 RPCA^[104] 方法重构训练数据, 并移除重构误差较大的数据, 移除的比例和噪音的比例相同。
- **CAE+CNN:** 用卷积自动编码器预训练卷积神经网络的每一层, 然后微调整一个网络, 减小噪音标注的影响^[111]。
- **NL+CNN:** 用全连接层表示噪音概率矩阵并和卷积神经网络一起训练^[30]。

对于 VOC2007 数据集, 我们与另外两种方法进行比较:

- **Best_VOC:** 在 ImageNet 数据集预训练网络, 并在 VOC2007 上微调^[112]。
- **Web_HOG:** 在网络图片上基于局部模型和人工设计的特征对图片分类^[113]。

参数设置: 首先, 我们调整公式 (3.2) 中权重衰减项的系数 β 。对于 10% 的噪音比例, 该系数取 0.004 时网络能达到最好的效果, 对于 20% 的噪音比例, 取值为 0.008, 其他噪音比例下取值为 0.04。该参数设置对于两个数据集都能取得最好的效果。此外, 我们按照经验将公式 (3.6) 中的 γ 参数设为 0.1, 使得特征相似度在合理的范围。图 3.6 显示了在 Cifar10 数据集 20% 噪音下, 公式 (3.11) 中不同 α 取值对于网络准确率的影响。我们发现, 只有当 α 取值过大时(比如取 10),

表3.1 Cifar10 数据集上不同算法在不同噪音比例下的准确率比较

算法 \ 噪声比例	无噪声	10%	20%	30%	40%	50%	60%	70%	80%	90%
CNN	81.24	77.79	71.97	65.09	55.65	45.60	36.65	25.02	19.46	17.55
RPCA+CNN	81.24	77.94	72.44	65.94	57.82	45.77	36.55	23.68	17.85	15.49
CAE+CNN	81.55	78.54	73.19	67.69	60.83	52.71	44.71	34.39	27.54	18.61
NL+CNN	81.16	78.28	73.36	68.26	61.63	55.83	47.33	37.12	30.81	19.49
RFCNN	81.60	79.39	76.21	72.81	68.79	63.01	54.78	45.48	35.43	20.56

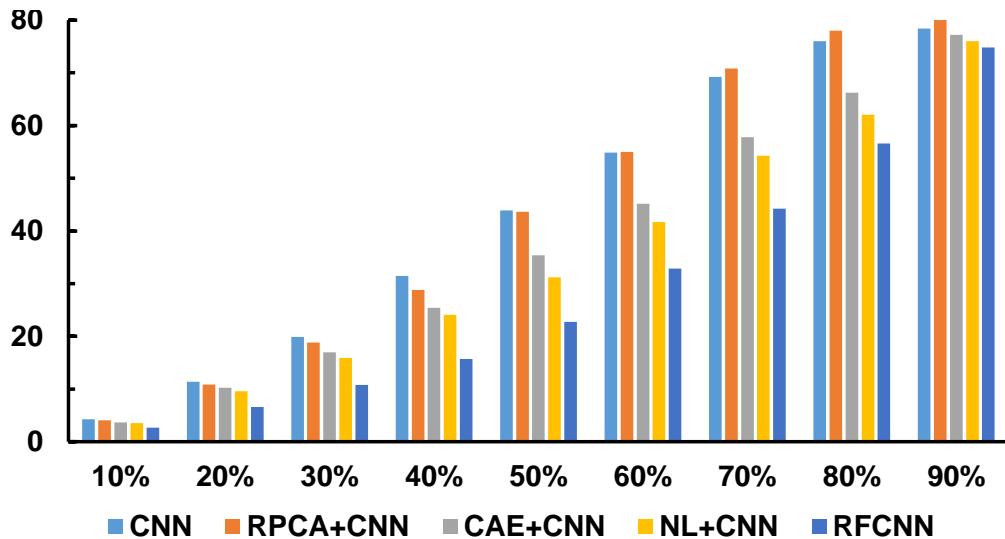


图3.7 不同算法在不同噪音比例下的准确率相对于无噪音准确率的下降程度比较

模型完全丧失了分类能力，对于其他取值，准确率都保持在相对稳定的范围，并在取值为0.5时达到最优。此外，我们发现 α 取0.5在其他噪音条件下也能取得最好的效果。因此，以下实验 α 均取0.5。

实验结果：表格3.1显示了在Cifar10数据集上不同噪音程度下不同算法的分类准确率比较。图3.7反映了在不同噪音程度下不同方法分类准确率相对于无噪音下降程度的比较。本文提出的算法在所有条件下都达到了最好的实验结果，甚至在无噪音数据集上，我们的算法也比经典的卷积神经网络取得了略好的准确率。我们发现，在30%数据噪音下，经典卷积神经网络的准确率下降了将近20%，相比之下，本章提出的算法仅下降了10%，表现出对数据噪音很强的鲁棒性。此外，我们发现数据预处理方法RPCA+CNN在噪音比例小于50%时，准确率要高于经典的卷积神经网络，当有更多的噪音数据时，RPCA+CNN的效果则比经典CNN要差。这个现象的原因在于当噪音数据增多时，数据预处理移除正确数据的风险也随之增大，导致在最终的训练数据中噪音数据的比例增加。CAE+CNN和NL+CNN算法的性能十分接近，在30%噪音比例下，准确率分别下降17.0%和15.9%。CAE+CNN虽然能够解决区域级的噪音（背景噪音），但对于样本集噪音（如标注错误），算法的鲁棒性比较有限。对于NL+CNN，实验证明仅仅在网

表3.2 VOC2007 数据集上不同算法在不同噪音比例下不同类别的平均准确率比较

类别 \ 算法	Best_VOC	Web_HOG	CNN (Web)	CNN (Webx4)	RFCNN (Web)	RFCNN (Webx4)
Aeroplanes	88.5	68.5	84.1	85.4	85.8	91.3
Bicycles	81.5	48.2	68.8	69.4	69.7	75.2
Birds	87.9	47.3	77.1	77.1	77.4	83.3
Boats	82.0	55.7	73.0	74.5	75.1	81.5
Bottles	47.5	40.0	63.0	63.7	63.8	70.2
Buses	75.5	56.3	74.2	74.7	75.8	81.3
Cars	90.1	60.1	74.3	75.0	75.6	80.6
Cats	87.2	64.1	79.2	81.6	82.7	88.3
Chairs	61.6	43.6	61.8	62.3	62.7	67.0
Cows	75.7	59.2	73.8	75.7	76.9	82.5
Dining tables	67.3	32.9	48.9	53.3	53.5	60.0
Dogs	85.5	46.5	79.5	80.2	80.6	86.3
Horses	83.5	56.2	81.0	83.8	84.7	90.0
Motorbikes	80.0	62.4	82.1	84.6	84.9	90.3
People	95.6	41.3	48.4	50.7	49.2	75.8
Potted plants	60.8	29.6	57.9	58.9	59.1	64.8
Sheep	76.8	41.4	72.0	75.9	76.0	81.0
Sofas	58.0	35.6	31.6	41.0	50.8	57.8
Trains	90.4	68.9	83.4	84.5	84.8	89.9
TV/Monitors	77.9	35.5	64.7	69.1	69.2	74.9
mAP	77.7	49.6	68.9	71.1	71.9	78.6

络上增加一层噪音适应层并不能达到很好的噪音鲁棒性。相反，本章所提算法可以抑制噪音在所有层的影响。

在 PASCAL VOC2007 数据集上，我们首先在 ImageNet 数据集上预训练 Alex-Net 网络^[23]，然后在网络数据上微调网络参数。为了获取网络数据，我们将数据集的每个类别作为查询词抓取搜索引擎的返回结果，并滤除重复的图片。我们收集了两个训练数据集，第一个数据集中正负样本的数目和 VOC2007 中相同，在该数据集上的实验我们记为 CNN(Web) 和 RFCNN(Web)。第二个数据集中我们将正样本的数目增加到 VOC2007 的 4 倍，并将该数据集上的方法记为 CNN(Webx4) 和 RFCNN(Webx4)。根据统计的结果，两个数据集上的噪音比例分别是 20% 和 40%。不同方法在 VOC2007 测试数据集上的平均准确率如表 3.2 所示。可以发现：

- CNN(Web) 相比 Web_HOG 具有十分明显的提升，证明了深度学习网络比基于人工设计的特征训练的分类模型具有更强的噪声鲁棒性。
- 相比于在有限的标注好的数据集上训练深度模型，RFCNN(Webx4) 取得了更好的效果。

表 3.3 社交图片标注结果比较

	CNN(Web)	RPCA+CNN(Web)	CAE+CNN(Web)	NL+CNN(Web)	CNN(ImageNet)	DeViSE	RFCNN
NDCG@1	0.08	0.23	0.11	0.24	0.20	0.28	0.32
NDCG@3	0.18	0.32	0.25	0.33	0.29	0.36	0.41
NDCG@5	0.26	0.39	0.34	0.41	0.39	0.43	0.46



图 3.8 社交图片标注结果示例

3.3.2 社交图片标注

本章提出的弱监督相关反馈深度卷积神经网络可以大量的社交图片作为训练数据，可以用于任意类别的社交图片标注。例如，社交图片中广泛存在的“风景”，“家庭”等标签并没有相应的标注好的训练数据。

我们根据 Flickr¹上 100,000 名活跃用户（注册超过 2 年并在过去 6 个月上传超过 500 张照片）提供的标签中选取了 200 个常见标签，其中 50 个标签例如“日落”，“观光”，“生日”等没有在 ImageNet 数据集中出现。实验中的基准方法采用了 ImageNet 数据集。对于这 200 个标签，我们仅用 ImageNet 数据集提供的每类 1,000 张图片作为训练数据训练模型，该方法称为 **CNN(ImageNet)**。此外，我们还采用了视觉语义嵌套方法 DeViSE^[114]，在 150 个已有训练数据类别上训练视觉语义嵌套模型，对 200 个类别进行标注。对于我们的方法，为了获得完整的 200 个类别的数据，我们从搜索引擎上为每个类别抓取 1,000 张图片，移除重复图片并训练模型。网络模型与 VOC2007 实验中相同。

为了测试不同算法的性能，我们从 **MIT-Adobe FiveK** 数据集^[115] 中随机选取了 1,000 张图片作为测试集。每种方法根据预测的分数为每张图生成 5 个有序标签。我们邀请了 25 位标注人员为每个标签打分：2—非常相关；1—相关；0—不相关。我们采用归一化折扣增益值 (Normalized Discounted Cumulative Gain, NDCG) 作为评价指标。NDCG 可以评价不同级别的相关性，相关结果越靠前得

¹<https://www.flickr.com/>

到的评分越高。有序结果中位置 p 的 NDCG 定义为：

$$\text{NDCG}@p = Z_p \sum_{i=1}^p \frac{2^{r^i} - 1}{\log(1+i)} \quad (3.18)$$

其中 2^{r^i} 是第 i 个标签的相关性评价， Z_p 是归一化参数。

表格 3.3 显示了相关性评价结果。相比于其它噪音鲁棒算法，本章提出的方法获得了最好的效果。此外，CNN(ImageNet) 数据集由于训练的标注空间较小，效果较差，证明了利用社交图片的有效性。本章算法的效果也比视觉语义嵌套方法 DeVISE 要好。

图 3.8 显示了本章提出的弱监督相关反馈深度卷积神经网络在社交图片上标注结果的示例，其中有下划线的标注没有在 ImageNet 数据集中出现。

3.4 本章小结

本章提出了弱监督相关反馈深度卷积神经网络，从大量的弱标注社交多媒体数据中学习目标识别模型。算法利用了训练数据之间的相关性使得不同的训练数据在网络训练中有不同的梯度贡献，抑制了训练数据中标注噪音的影响，实现了在不需要人工标注的情况下快速学习任意的语义类别。为了提高算法的效率，本章对弱监督相关反馈深度神经网络做了进一步的简化和近似。本章在混入噪音的标准数据集和真实的社交图片上进行了实验，验证了算法的对于标注噪音的鲁棒性。

本章的主要贡献包括：

- 基于感知连续性提出了端到端的弱监督相关反馈神经网络，直接从弱标注社交多媒体数据中学习任意语义类别。
- 对相关反馈网络做了简化近似，提高了算法的计算速度。
- 本章提出的弱监督相关反馈深度神经网络不依赖特定的网络结构，可以方便地应用到更复杂的网络结构和应用场景。

第4章 大规模社交多媒体数据快速处理

社交多媒体数据由于规模庞大，需要耗费大量的计算资源和处理时间。本章从特征选取和模型简化两个角度讨论大规模社交多媒体数据的快速处理。

本章首先介绍用于大规模高维数据的特征选取算法。社交多媒体数据的特征包括深度网络特征、全局特征、局部特征等，实际应用中需要根据特定的需求选取对目标任务最有用的特征，这对于提高大规模社交多媒体数据的处理速度以及在计算能力、内存空间和电池续航都十分有限的移动设备上处理社交多媒体数据都具有重要意义。此外，去除与特定任务不相干的特征，还可以提高特征的表达能力。

其次，本章介绍用于深度卷积神经网络的模型简化算法。深度卷积神经网络在很多计算机视觉领域都表现出很好的效果，然而大量的模型参数需要大量的计算资源和时间，极大地限制了深度卷积神经网络在实际场景中的应用。此外，深度网络在移动设备上的应用已经成为一种趋势，由于移动设备计算能力的限制，在不影响模型准确率的条件下简化深度网络模型已经成为迫切的需要。

4.1 在线特征选取问题建模

特征选取是指从数据中移除不相关或冗余的特征。在当前大数据的背景下，特征选取已经成为了十分重要的技术，并在多个领域尤其是大规模高维数据的场景下获得了广泛应用^[116,117]。尽管特征选取已经被广泛地研究，大部分的算法都属于批处理学习。批处理学习的主要问题在于需要将整个数据集加载到内存中，对于实际问题中的大规模高维数据不具有可伸缩性。此外，批处理方法假设所有的训练数据和特征在训练前已经确定，而实际场景中需要处理的通常是流数据，并可能伴随有新的特征出现。为了克服批处理算法的这些问题，近年来有部分工作研究在线特征选取^[39,51,53]。在线学习的优点在于算法每次迭代只处理一个数据，因此具有很高的伸缩性，并且可以很好地应对数据中模式和特征的变化。然而，目前已有的在线特征选取算法复杂度仍然过高，算法的准确率与批处理方法也有不小差距。因此，本章提出了快速二阶在线特征选取算法，不仅对于大规模高维数据具有很高的可伸缩性和学习效率，算法的准确率也与批处理方法十分接近。

不失一般性，本章首先研究二分类问题，并在4.4节将算法扩展到多类问题。

令 $\{(\mathbf{x}^t, y^t) | t = 1, \dots, T\}$ 表示训练过程中依次收到的数据，每个数据 $\mathbf{x}^t \in \mathbb{R}^d$ 是一个 d 维的向量，数据类别 $y^t \in \{+1, -1\}$ 。在线学习算法学习一个相同维度的线性分类器 $\mathbf{w} \in \mathbb{R}^d$ 。在时刻 t ，算法接收到数据 \mathbf{x}^t ，基于当前的模型参数 \mathbf{w}^t 预测它的类别 $\hat{y} \in \{+1, -1\}$ 。

$$\hat{y}^t = \text{sign}(\mathbf{w}^t \cdot \mathbf{x}^t). \quad (4.1)$$

预测以后，算法得到真实的类别 y^t ，并衡量在数据 (\mathbf{x}^t, y^t) 上的损失函数 $\ell(\mathbf{w}^t)$ 。损失函数通常为真实类别和预测结果的函数。根据损失函数和特定的更新规则，算法更新模型参数 \mathbf{w}^t 。例如，在线梯度下降算法更新的规则为：

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta^t y^t \mathbf{x}^t, \quad (4.2)$$

η^t 是时刻 t 的学习率 (Learning Rate)。根据不同的更新规则，在线学习算法可以分为两类：

- 一阶算法：本质上是梯度下降算法^[44]；
- 二阶算法：挖掘输入数据几何特点^[46] 或构建目标方程的近似海森矩阵^[50]；

在线特征选取算法对 \mathbf{w} 的 L_0 范数施加公式 (4.3) 中所示约束，选取权重向量 \mathbf{w}^t 中相对较少的一部分元素，并将其他元素设为 0。

$$\|\mathbf{w}\|_0 \leq B, \|\mathbf{w}\|_0 = \sum_{i=1}^d w_i^0, \quad (4.3)$$

其中 B 是预先定义的常数。相应地，只有与 \mathbf{w} 中非零权重对应的 \mathbf{x} 的 B 个特征被选取。

4.2 置信度加权二阶在线特征选取

在线特征选取最直接的算法是截断感知机算法 (Perceptron with Truncation, PET)^[53]。具体来说，分类器在每次迭代时首先根据 \mathbf{w}^t 预测类别 \hat{y}^t 并计算损失函数 $\ell(\mathbf{w}^t)$ 。如果 $\ell(\mathbf{w}^t)$ 大于 0，则 $\mathbf{w}^{t+1} = \mathbf{w}^t$ ；否则，分类器根据感知机规则更新 \mathbf{w}^{t+1} : $\hat{\mathbf{w}}^{t+1} = \mathbf{w}^t + \eta y^t \mathbf{x}^t$ 。算法保留更新后参数绝对值最大的 B 个元素，其他元素设为 0。截断后的分类器参数 \mathbf{w}^{t+1} 被用于下一轮迭代的预测和更新。算法 4.1 显示了 PET 算法的框架，算法 4.2 显示了在线特征选取的截断函数。

根据 Wang 等人的分析，上述算法在实际应用中并不总能取得很好的效果^[53]。它不能保证被截断的参数足够小，因而不能保证很小的错误率。因此，Wang 等人提出了一阶在线特征选区 (First Order Feature Selection, FOFS) 算法，在截断之前对权重向量做稀疏投影。FOFS 算法保证了每次迭代时分类器参数 \mathbf{w}^t 都限制在一个 L_1 范数约束的超体内部。算法 4.3 显示了 FOFS 算法的细节。

input : B - 需要选取的特征个数, η - 学习率

output: 权重向量 \mathbf{w}^T

```

1 初始化  $\mathbf{w}^1 = \mathbf{0}$ ;
2 for  $i \leftarrow 1$  to  $T$  do
3   接收到数据  $\mathbf{x}^t \in \mathbb{R}^d$ , 预测类别  $\hat{y}^t = \text{sign}(\mathbf{w}^t \cdot \mathbf{x}^t)$ ;
4   接收真实类别  $y^t$ ;
5   计算损失函数  $\ell(\mathbf{w}^t)$ ;
6   if  $\ell(\mathbf{w}^t) > 0$  then
7      $\hat{\mathbf{w}}^{t+1} = \mathbf{w}^t + \eta y^t \mathbf{x}^t$ ;
8      $\mathbf{w}^{t+1} = \text{Truncate}(\hat{\mathbf{w}}^{t+1}, B)$ ;
9   end
10 end

```

算法 4.1: PET——截断感知机算法

input : $\hat{\mathbf{w}}$ - 权重向量, B - 需要选取的特征个数

output: 截断的权重向量 \mathbf{w}

```

1  $\mathbf{w} = \hat{\mathbf{w}}$ ;
2 if  $\|\hat{\mathbf{w}}\|_0 > B$  then
3   除了  $\mathbf{w}$  的绝对值最大的  $B$  个元素, 其他元素全部设为 0;
4 end

```

算法 4.2: Truncate——截断函数

一般来说, 一阶在线特征选取算法的复杂度和特征维度成正比。对于超高维度数据, 算法的速度比较慢。此外, 当输入数据的不同维度特征不在同一个尺度时, 一阶算法可能会移除有价值的特征。如公式(4.1)所示, 预测结果不仅依赖于权重向量, 也依赖于输入数据。即使 $|w_i| < |w_j|$, 也不能保证 $|w_i * E(x_i)| < |w_j * E(x_j)|$, $E(x_i)$ 是 x_i 的期望。为了克服一阶算法的局限性, 我们探索了二阶在线学习的最新发展, 提出了置信度加权二阶在线特征选取 (Second Order Feature Selection, SOFS) 算法。

置信度加权二阶在线学习算法^[118] 假设线性分类器的权重向量服从高斯分布 $\mathbf{w} \sim \mathcal{N}(\mu, \Sigma)$ 。权重的置信度通过协方差矩阵 Σ 的对角元素表示, 对角元素 Σ_{jj} 越小, 权重 w_j 的均值的置信度越高。算法在接收到训练数据之前, 所有权重有共同的置信度或不确定性。训练过程中, 给定一个接收到的训练数据 (\mathbf{x}^t, y^t) , 置信度加权算法更新权重使得在当前数据 \mathbf{x}^t 上做出正确预测的概率大于一个阈

input : B - 需要选取的特征个数, η - 学习率, λ - 正则化参数

output: 截断的权重向量 \mathbf{w}

- 1 $\tilde{\mathbf{w}}^{t+1} = (1 - \lambda\eta)\mathbf{w}^t + \eta y^t \mathbf{x}^t;$
- 2 $\hat{\mathbf{w}}^{t+1} = \min\left\{1, \frac{1}{\|\tilde{\mathbf{w}}^{t+1}\|_2} \tilde{\mathbf{w}}^{t+1}\right\};$
- 3 $\mathbf{w}^{t+1} = \text{Truncate}(\hat{\mathbf{w}}^{t+1}, B);$

算法 4.3: FOFS——一阶在线特征选取算法

值 τ 。同时, 算法尽量保持与更新前的权重分布相同。置信度加权算法可以表示为如下所示优化问题:

$$\begin{aligned} (\hat{\boldsymbol{\mu}}^{t+1}, \Sigma^{t+1}) &= \arg \min_{\boldsymbol{\mu}, \Sigma} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) \\ \text{s.t. } &\Pr_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)}[y^t(\mathbf{w} \cdot \mathbf{x}^t) \geq 0] \geq \tau, \end{aligned} \quad (4.4)$$

其中 $D_{\text{KL}}(*, *)$ 是 Kullback-Leibler(KL) 距离。两个高斯分布 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ 和 $\mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)$ 的 KL 距离定义为:

$$\begin{aligned} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) &= \frac{1}{2} \log \frac{\det \Sigma^t}{\det \Sigma} + \frac{1}{2} \text{Tr}((\Sigma^t)^{-1} \Sigma) \\ &+ \frac{1}{2} (\boldsymbol{\mu}^t - \boldsymbol{\mu})^T (\Sigma^t)^{-1} (\boldsymbol{\mu}^t - \boldsymbol{\mu}) - \frac{d}{2}. \end{aligned} \quad (4.5)$$

公式(4.4)中的约束可以重新表达为: $y^t(\boldsymbol{\mu} \cdot \mathbf{x}^t) \geq \phi \sqrt{(\mathbf{x}^t)^T \Sigma \mathbf{x}^t}$, $\phi = \Phi^{-1}(\tau)$ (Φ 是高斯分布的累积函数)。研究人员提出了多种方法解公式 (4.4) 中的优化问题。本章采用能够对每个训练数据的预测函数进行自适应正则化的 AROW 算法^[46]。研究和实验表明, 该算法对于训练数据中的噪音具有更好的鲁棒性。AROW 算法的目标方程为:

$$(\hat{\boldsymbol{\mu}}^{t+1}, \Sigma^{t+1}) = \arg \min_{\boldsymbol{\mu}, \Sigma} \left\{ D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) + \frac{1}{2\gamma} \ell^t(\boldsymbol{\mu}) + \frac{1}{2\gamma} (\mathbf{x}^t)^T \Sigma \mathbf{x}^t \right\}, \quad (4.6)$$

其中 $\gamma > 0$ 是正则化参数。 $\ell^t(\boldsymbol{\mu})$ 是平方铰链损失函数 (Squared Hinge Loss):

$$\ell^t(\boldsymbol{\mu}) = \max(0, 1 - y^t(\boldsymbol{\mu} \cdot \mathbf{x}^t))^2. \quad (4.7)$$

方程 (4.6) 存在如下闭合解:

$$\begin{aligned} \beta^t &= \frac{1}{(\mathbf{x}^t)^T \Sigma^t \mathbf{x}^t + \gamma} \quad \mathbf{g}^t = -2 \max(0, 1 - y^t(\boldsymbol{\mu}^t \cdot \mathbf{x}^t)) y^t \mathbf{x}^t \\ \hat{\boldsymbol{\mu}}^{t+1} &= \boldsymbol{\mu}^t - \frac{1}{2} \beta^t \Sigma^t \mathbf{g}^t \quad (\Sigma^{t+1})^{-1} = (\Sigma^t)^{-1} + \frac{\text{diag}(\mathbf{x}^t (\mathbf{x}^t)^T)}{\gamma} \end{aligned} \quad (4.8)$$

需要注意的是, 本文提出的 SOFS 算法仅仅考虑和计算协方差矩阵 Σ 的对角元素。从效率的角度, 维护完整协方差矩阵需要 $O(d^2)$ 的空间复杂度和 $O(d^2)$ 的计算复杂度, 这对于大规模高维数据是不切实际的。从学习能力的角度, 相关

input : B - 需要选取的特征个数, γ - 正则化参数

output: 权重向量 μ^T 和对角协方差矩阵 Σ^T

```

1 初始化  $\mu^1 = \mathbf{1}$ ,  $\Sigma^1 = I$ ;
2 for  $i \leftarrow 1$  to  $T$  do
3   接收到数据  $\mathbf{x}^t \in \mathbb{R}^d$ , 并预测  $\hat{y}^t = \text{sign}(\mu^t \cdot \mathbf{x}^t)$ ;
4   接收到数据的真实类别  $y^t$ , 计算损失函数  $\ell(\mu^t)$ ;
5   if  $\ell(\mu^t) > 0$  then
6     根据公式 (4.8)计算  $\hat{\mu}^{t+1}, \Sigma^{t+1}$ ;
7     for  $j \leftarrow 1$  to  $d$  do
8       if  $\Sigma_{jj}^{t+1}$  是最小的  $B$  个元素之一 then
9          $\mu_j^{t+1} = \hat{\mu}_j^{t+1}$ ;
10      else
11         $\mu_j^{t+1} = 0$ ;
12      end
13    end
14  end
15 end

```

算法 4.4: SOFS——二阶在线特征选取算法

研究工作也表明在数据量足够的情况下, 对角协方差矩阵可以获得比完整协方差矩阵更好的性能, 原因在于学习的初始阶段, 完整协方差矩阵算法适应数据之间相互依赖的能力, 当数据不可分时同样也使得它在逼近最佳权重向量时过度拟合噪音^[119]。

不同于一阶在线特征选取算法基于权重向量的绝对值大小决定特征的重要性, 本章提出的二阶在线特征选取算法的核心思想是利用二阶信息保留 B 个置信度最高的特征。具体来说, 在线学习过程中, 当训练数据 (\mathbf{x}^t, y^t) 的损失函数不为 0 时, 算法仅更新前 B 个最小协方差 Σ_{jj} 对应的 B 个最确信的权重, 剩余权重设为 0。算法 4.4 显示了本章提出的 SOFS 算法。

4.3 快速在线特征选取算法

当前的在线特征选取算法的一个普遍问题在于计算复杂度过高。具体来说, 在线特征选取的一个主要时间开销在于从 d 维数组 (FOFS 算法中的权重绝对值向量和 SOFS 算法中的协方差矩阵的对角向量) 中选取最大或最小的 B 个元素。本节提出一个基于最小堆的快速 FOFS 算法和快速 PET 算法, 避免了在迭代的

每一步对整个向量排序^[53]。此外，基于类似的最大堆的实现，本节进一步利用置信度单调递减特性提出了复杂度更低的快速 SOFS 算法。

4.3.1 一阶快速在线特征选取算法

从 d 维数组中找出最大的 B 个元素（算法 4.2 中的 Truncate 函数）的直接的做法是对 d 个元素排序，并选取前 B 个元素。为了提高计算效率，我们构建了一个最小堆存储权重向量 \mathbf{w}^t 的 B 个最大绝对值。学习过程中，当分类器的权重向量发生改变以后，通过如下两步更新找出最大的 B 个元素：

- 调整已经存在于堆中的元素的位置，维护最小堆结构。
- 比较不在堆中的每个元素与堆顶元素的大小。如果小于堆顶元素，则将它的值设为 0，否则将堆顶元素替换为当前元素，并调整堆顶元素与子节点的位置，维护最小堆结构，原堆顶元素的值设为 0。

算法 4.5 显示了快速 FOFS 算法的详细步骤。快速 PET 算法的过程与之类似。

证明上述算法正确性的关键在于证明每次迭代以后绝对值最大的 B 个特征仍然在最小堆中。用 h_1, \dots, h_d 表示堆中特征的位置下标，其他不在堆中的特征的下标记为 h_{B+1}, \dots, h_d 。在第一步中， v_{h_1}, \dots, v_{h_B} 被重新组织以满足最小堆的结构，存在如下两个命题成立：

命题 4.1 如果模型更新后 $v_{h_i}, \forall i \in [1, B]$ 仍然在最大的 B 个元素中，则 v_{h_i} 不会被替换出最小堆；

命题 4.2 如果模型更新后 $v_{h_i}, \forall i \in (B, d]$ 在最大的 B 个元素中，则 v_{h_i} 一定会被替换进最小堆。

证明 对于命题 4.1，如果 v_{h_i} 不是 B 个最大元素中最小的，则 v_{h_i} 始终不会成为堆顶元素，因而一定不会被替换出最小堆。如果 v_{h_i} 是 B 个最大元素中最小的，则意味着最小堆中元素已经构成了最大的 B 个特征，剩下的 $d - B$ 个特征权重的绝对值均比 v_{h_i} 小，因此仍然不会在第二步过程中被替换出最小堆。对于命题 4.2，可以得到 v_{h_i} 是最大的 B 个元素之一时，堆顶元素一定小于 v_{h_i} ，因此一定会被替换进最小堆。综上所述，本文提出的最小堆结构和更新方法可以找出权重绝对值最大的 B 个特征。□

4.3.2 二阶快速在线特征选取算法

尽管一阶快速在线特征选取算法避免了对所有元素排序，算法复杂度依然和特征的维度成正比。对于本章提出的二阶在线特征选取算法，可以进一步利用

input : B - 需要选取的特征个数, η - 学习率, λ - 正则化参数

output: 权重向量 μ^T

```

1 初始化  $\mathbf{w}^1 = \mathbf{1}$ ,  $\mathbf{v}^1 = (|w_1^1|, \dots, |w_d^1|) = \mathbf{0}$ ,  $\mathbf{v}^1$  上大小为  $B$  的最小堆  $H$ ;
2 for  $i \leftarrow 1$  to  $T$  do
3   接收到数据  $\mathbf{x}^t \in \mathbb{R}^d$ , 并预测  $\hat{y}^t = \text{sign}(\mathbf{w}^t \cdot \mathbf{x}^t)$ ;
4   接收到真实类别  $y^t$ , 计算损失函数  $\ell(\mathbf{w}^t)$ ;
5   if  $\ell(\mathbf{w}^t) > 0$  then
6      $\tilde{\mathbf{w}}^{t+1} = (1 - \lambda\eta)\mathbf{w}^t + \eta y^t \mathbf{x}^t$ ;
7      $\mathbf{w}^{t+1} = \min\{1, \frac{1}{\|\tilde{\mathbf{w}}^{t+1}\|_2}\} \tilde{\mathbf{w}}^{t+1}$ ;
8      $\mathbf{v}^{t+1} = (|w_1^{t+1}|, \dots, |w_d^{t+1}|)$ ;
9     调整  $H$  中节点的位置, 维护最小堆结构;
10    for  $j \leftarrow 1$  to  $d, v_j^{t+1} \notin H$  do
11      if  $v_j^{t+1} > H_{min}$  then
12        获取堆顶节点  $H_{min}$ , 堆顶对应的特征位置记为  $s$ ;
13         $w_s^{t+1} = 0$ , 并将堆顶  $H_{min}$  替换为  $v_j^{t+1}$ ;
14        调整堆顶元素与子节点的位置, 维护最小堆结构;
15      else
16         $w_j^{t+1} = 0$ ;
17      end
18    end
19  end
20 end

```

算法 4.5: 一阶快速在线特征选取算法

二阶特征的特殊性将算法复杂度降低为和非零特征的个数成正比, 这对于大规模高维稀疏数据具有重大的意义。区别于一阶在线特征选取算法, 本章提出的二阶算法具有如下单调递减特性:

命题 4.3 (单调递减性) 对于 $\forall t$ 和 $\forall j \in [1, d]$, 公式 (4.8)中的对角协方差矩阵 Σ^t 满足 $\Sigma_{jj}^{t+1} \leq \Sigma_{jj}^t$ 。

命题的正确性可以由 $\text{diag}(\mathbf{x}^t(\mathbf{x}^t)^T)/\gamma$ 始终非负得到。基于上述命题, 本章提出快速二阶在线特征选取算法。算法维护一个最大堆结构存储当前协方差矩阵对角元素的最小 B 个元素。由于协方差矩阵每个对角元素的单调递减性质, 对于每个被更新权重的特征, 算法的更新规则为:

- 如果特征已经在最大堆中, 算法仅需要比较当前特征与子节点的大小维护

```

input :  $B$  - 需要选取的特征个数,  $\gamma$  - 正则化参数
output: 权重向量  $\mu^T$  和对角协方差矩阵  $\Sigma^T$ 

1 初始化  $\mu^1 = \mathbf{1}$ ,  $\Sigma^1 = I$ ,  $B$  个  $\Sigma^1$  元素的最大堆  $H$ ;
2 for  $i \leftarrow 1$  to  $T$  do
3   接收到数据  $\mathbf{x}^t \in \mathbb{R}^d$ , 并预测  $\hat{y}^t = \text{sign}(\mu^t \cdot \mathbf{x}^t)$ ;
4   接收到真实类别  $y^t$ , 根据公式 (4.7) 计算损失函数  $\ell(\mu^t)$ ;
5   if  $\ell(\mu^t) > 0$  then
6     for  $j \leftarrow 1$  to  $d$ ,  $x_j^t \neq 0$  do
7       根据公式 (4.8) 计算  $\hat{\mu}_j^{t+1}$ ,  $\Sigma_{jj}^{t+1}$ ;
8       如果  $\Sigma_{jj}^{t+1} \in H$ , 递归调整  $\Sigma_{jj}^{t+1}$  与它子节点的位置, 维护最大
         堆结构;
9     end
10    for  $j \leftarrow 1$  to  $d$ ,  $x_j^t \neq 0$ ,  $\Sigma_{jj}^{t+1} \notin H$  do
11      if  $\Sigma_{jj}^{t+1} < H_{max}$  then
12        获取堆顶节点  $H_{max}$ , 堆顶对应的特征位置记为  $s$ ;
13         $\mu_s^{t+1} = 0$ , 将堆顶  $H_{max}$  替换为  $\Sigma_{jj}^{t+1}$ ;
14        调整堆顶元素与子节点的位置, 维护最大堆结构;
15      else
16         $\mu_j^{t+1} = 0$ ;
17      end
18    end
19  end
20 end

```

算法 4.6: 二阶快速在线特征选取算法

最大堆结构。因为置信度单调递减, 更新后节点元素的值一定小于父节点;

- 如果被更新的特征不在最大堆中, 则比较其与堆顶的大小, 如果小于堆顶, 则替换堆顶, 并将原堆顶对应的特征权重置为 0, 否则将当前特征的权重置为 0。对于没有被更新权重的特征, 不需要进行比较, 因为堆顶具有单调递减的特性, 没有被更新权重的特征的置信度一定大于堆顶。

算法 4.6 显示了快速 SOFS 算法的细节。

4.3.3 复杂度分析

上述快速在线特征选取算法显著提高了在线特征选取的效率。本节分析上述算法的计算复杂度。

记权重向量的维度为 d , 每个数据平均非零特征个数为 m , 最差情况下, PET 算法单步迭代的计算复杂度为 $\{3m + d - B + d \log B\}$:

- $2m$: 计算损失函数, 更新权重向量;
- m : 计算权重向量的绝对值;
- $B \log B$: 维护最小堆;
- $(d - B) \log B$: 找出最大的 B 个元素, 维护最小堆;
- $d - B$: 将相应的权重置为 0。

FOFS 算法单步迭代的计算复杂度远高于 PET 算法, 为 $\{2m + 4d - B + d \log B\}$:

- $2m$: 计算损失函数, 更新模型;
- d : 计算权重向量的范数;
- d : 稀疏投影;
- d : 计算权重向量的绝对向量;
- $B \log B$: 维护最小堆;
- $(d - B) \log B$: 找出最大的 B 个元素, 维护最小堆;
- $d - B$: 将相应的特征值向量置为 0。

SOFS 算法单步迭代的复杂度为 $\{4m + m \log B\}$:

- $3m$: 计算损失函数, 更新模型和对角协方差矩阵;
- $m \log B$: 维护最大堆 (只有 m 个值发生改变);
- m : 将相应的权重置为 0。

当 $m \ll d$ 并且 $B \ll d$ 时, SOFS 算法处理大规模高维稀疏数据具有很高的效率和可伸缩性。在最差情况下, $m \approx d$, SOFS 算法的复杂度与 PET 算法接近, 但仍然小于 FOFS。

对于空间复杂度, 我们只考虑分类器需要的空间占用, 不考虑数据加载和存储的内存开销。在算法的具体实现众, 出于效率考虑, 输入数据存储成键值对的稀疏形式, 模型参数表示成密集向量。PET 算法和 FOFS 算法需要保存权重向量 \mathbf{w} 和它的绝对值向量 \mathbf{v} , 因而空间复杂度为 $O(2d)$ 。SOFS 算法也需要 $O(2d)$ 的空间复杂度用来保存权重向量和协方差矩阵的对角元素。因此, SOFS 算法的空间复杂度和一阶在线特征选取算法的空间复杂度相同。

4.4 置信度加权二阶多类在线特征选取

多类问题中, 假设共有 k 个类别, 每个训练数据的类别为 $y \in \{0, 1, \dots, k-1\}$ 。本节采用一对多的策略 (one-vs-the-rest) 将二阶在线特征选取算法扩展到多类问

题。根据 Crammer 等人的策略^[45], 置信度加权多类模型的分布类似于二分类问题, $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu} \in \mathbb{R}^{kd}$, $\Sigma \in \mathbb{R}^{kd \times kd}$ 。引入新的类别相关的特征:

$$\psi(\mathbf{x}, i) = [\mathbf{0}^T, \dots, \mathbf{x}^T, \dots, \mathbf{0}^T]^T,$$

只有 $\psi(\mathbf{x}, i)$ 的第 i 个位置为 \mathbf{x} , 其他位置为 $\mathbf{0}$ ($\mathbf{0}, \mathbf{x} \in \mathbb{R}^d$)。在每次迭代中, 分类器接收到训练数据 \mathbf{x}^t 并预测类别 $\hat{y}^t = \arg \max_{i=0}^{k-1} \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}, i)$ 。损失函数为:

$$\ell(\boldsymbol{\mu}^t) = \max(0, 1 - \boldsymbol{\mu}^t \cdot \Delta\psi^t)^2, \quad (4.9)$$

其中 $\Delta\psi^t$ 依赖于多分类问题的更新策略。对于最大分数多分类更新:

$$\Delta\psi^t = \psi(\mathbf{x}^t, y^t) - \psi(\mathbf{x}^t, \arg \max_{i=0, i \neq y^t}^{k-1} \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}, i)). \quad (4.10)$$

对于均匀多分类更新:

$$\Delta\psi^t = \sum_{i=0}^{k-1} \alpha_i^t \psi(\mathbf{x}^t, i), \quad \alpha_i^t = \begin{cases} -1/|E^t| & i \in E^t \\ 1 & \text{if } i = y^t, \\ 0 & \text{otherwise} \end{cases} \quad (4.11)$$

$$E^t = \{i \neq y^t : \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}^t, i) \geq \boldsymbol{\mu}^t \cdot \psi(\mathbf{x}^t, y^t)\}. \quad (4.12)$$

多分类更新目标方程为:

$$(\hat{\boldsymbol{\mu}}^{t+1}, \hat{\Sigma}^{t+1}) = \arg \min_{\boldsymbol{\mu}, \Sigma} \{D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma), \mathcal{N}(\boldsymbol{\mu}^t, \Sigma^t)) + \frac{1}{2\gamma} \ell(\boldsymbol{\mu}) + \frac{1}{2\gamma} (\Delta\psi^t)^T \Sigma \Delta\psi^t\}. \quad (4.13)$$

目标方程的闭合解与公式 (4.8) 类似, 区别在于将 $y^t \mathbf{x}^t$ 替换成 $\Delta\psi^t$ 。

多类在线特征选取仍然选取 B 个最确信特征。一对多策略的多类问题中, 特征的置信度依赖 k 个二分类器。第 j 个特征的置信度定义为 $C_j = k - \sum_{i=0}^{k-1} \Sigma_{ik+j, ik+j}$ 。算法仅更新前 B 个最大 C_j 对应的权重参数, 其他权重设为 0。算法细节与算法 4.4 类似, 区别在于将 $y^t \mathbf{x}^t$ 替换成 $\Delta\psi^t$ 。多类 SOFS 算法的时间复杂度是二分类问题的 k 倍。

多类问题中 $\sum_{i=0}^{k-1} \Sigma_{ik+j, ik+j}$ 仍然具有单调递减性:

命题 4.4 (单调递减性) 对于 $\forall t$ 和 $\forall j \in [1, d]$, 公式 (4.13) 中的 Σ^t 满足 $\sum_{i=0}^{k-1} \Sigma_{ik+j, ik+j}^t \leq \sum_{i=0}^{k-1} \Sigma_{ik+j, ik+j}^{t+1}$ 。

因此, 快速二分类二阶在线特征选取算法也适用于多类二阶在线特征选取。

4.5 实验结果和评估

本节在不同规模的合成数据和真实数据上通过实验验证本章提出的二阶在线特征选取算法的高效性和有效性。

4.5.1 实验设置

对于在线特征选取算法，如果没有显式说明，在线学习算法仅在训练数据上学习一轮。实验的比较对象包括当前最好的在线和批处理特征选取算法：

- PET：截断感知机算法，在线特征选取的基准算法^[53]；
- FOFS：前最好的一阶在线稀疏投影特征选取算法^[53]；
- mRMR：最小冗余最大相关特征选取算法^[120]，最好的批处理方法之一，以及它的图形处理器并行版本 (GPU-mRMR)^[121]；
- liblinear：用于大规模线性分类的开源库^[122]，选用其中的 l1-SVM 算法作为 *Embedded* 特征选取的代表算法。
- FGM：当前最好的 *Embedded* 批处理特征选取方法之一^[123]。

本实验中，在线学习方法使用 Hinge Loss 作为损失函数，并通过五重交叉验证找出最优超参数。对于每一个数据集，在线学习方法随机打乱数据顺序 10 次并取平均评测结果作为最终结果。对于 liblinear 中的 l1-SVM 算法，实验中调节 C 参数获得不同的特征个数。对于 FGM，实验遵循 Tan 等人论文中的设定将 C 设为 10^[123]。对于 mRMR，首先用该算法选取特定数目的特征，然后用在线梯度下降算法训练分类器。在具体的算法实现中，我们充分利用了在线学习依次处理单个数据的特点，使用两个并行线程分别加载数据和训练模型。

4.5.2 合成数据集实验评估

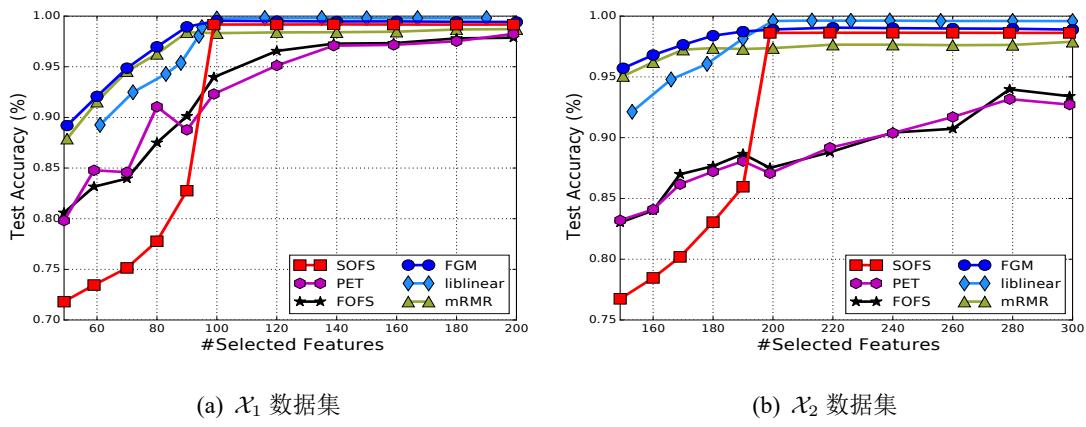
本节仿照 FGM 算法的评估方法，生成了三种类型的合成数据测试算法的性能、效率以及可伸缩性，分别是 $\mathcal{X}_1 \in R^{100K \times 10K}$, $\mathcal{X}_2 \in R^{100K \times 20K}$, $\mathcal{X}_3 \in R^{1M \times 1B}$ 。三个数据集都用于二分类任务。每个数据的每个维度从独立同分布的高斯分布 $\mathcal{N}(0, 1)$ 中采样得到。为了模拟真实的数据，每个采样得到的数据都是稀疏数据，有效特征维度分别为 100, 200, 和 500。每个数据随机选取 \mathcal{X}_1 的 200 维, \mathcal{X}_2 的 400 维, 和 \mathcal{X}_3 的 500 维作为噪声。为了获得数据的类别，从高斯分布 $\mathcal{N}(0, 1)$ 中采样得到权重向量 \mathbf{w}^* 作为分界面基准向量 (*groundtruth*)，每个数据类别为 $y = \text{sign}(\mathbf{w}^* \cdot \mathbf{x}^*)$, \mathbf{x}^* 是没有噪音特征的数据。合成数据集的详细情况如表 4.1 所示。

本节首先在 \mathcal{X}_1 和 \mathcal{X}_2 数据集上评估所有特征选取算法。然后在 \mathcal{X}_3 数据集测试本章算法的效率和可伸缩性。图 4.1 和图 4.2 显示了 \mathcal{X}_1 和 \mathcal{X}_2 上的准确率和训练时间的比较结果。

准确率。根据图 4.1 中的结果，有如下几点发现。首先，当选取的特征数目足够多时，(\mathcal{X}_1 中的 100 维, \mathcal{X}_2 中的 200 维)，SOFS 算法可以达到接近批处理算法的准确率，liblinear 和 FGM 相对于 SOFS 算法的优势十分有限。其次，当

表4.1 合成数据信息 (“K”, “M”, “B” 分别代表千, 百万, 十亿)

DataSet	#Train	#Test	Dim	IDim ^a	NDim ^b	#Feature
X_1	100K	10K	10K	100	200	30M
X_2	100K	10K	20K	200	400	60M
X_3	1M	100K	1B	500	500	1B

^a有效特征维度^b噪音特征维度图4.1 合成数据集 \mathcal{X}_1 和 \mathcal{X}_2 上测试准确率和特征数目之间的关系

选取的特征数目较少时，批处理算法的效果比在线学习算法更好，其中 FGM 和 mRMR 的测试准确率尤其突出。SOFS 算法的准确率虽然在特征不足时准确率不高，但随着更多的特征被选取，它的准确率迅速饱和并达到最佳。两个一阶在线特征选取算法表现最差，尤其是在 \mathcal{X}_2 数据集上。PET 算法和 FOFS 算法的准确率仅在特征数目很少时比 SOFS 算法高。特征数目足够多时，一阶在线特征选取算法的准确率不能达到批处理算法相当的水平。总结起来，本章提出的算法能有效挖掘有意义的特征，并能在特征数目足够多的情况下达到类似于批处理算法的准确率。

训练时间。模型训练的时间开销也是实际问题必须考虑的关键因素。图 4.2 显示了各个算法的训练时间比较。一般来说，批处理算法虽然效果较好，但是训练时间远高于在线学习算法。本章提出的 SOFS 算法只需要几秒钟就可以达到批处理算法的准确率。相反，liblinear 需要大约 10 倍的训练时间，FGM 和 mRMR 在 \mathcal{X}_2 数据集上甚至需要 100 倍的训练时间。并行 mRMR 算法相对于非并行算法减少了大约一半的时间。在线特征选取算法在这两个数据集上的训练时间差别不大，我们将在更大规模和更高维度上进行评估。尽管如此，准确率和训练时间的实验比较结果证明了 SOFS 算法是一个快速有效的在线特征选取算法。

本实验在大规模超高维度的 \mathcal{X}_3 数据集上测试 SOFS 的可伸缩性，由于已有

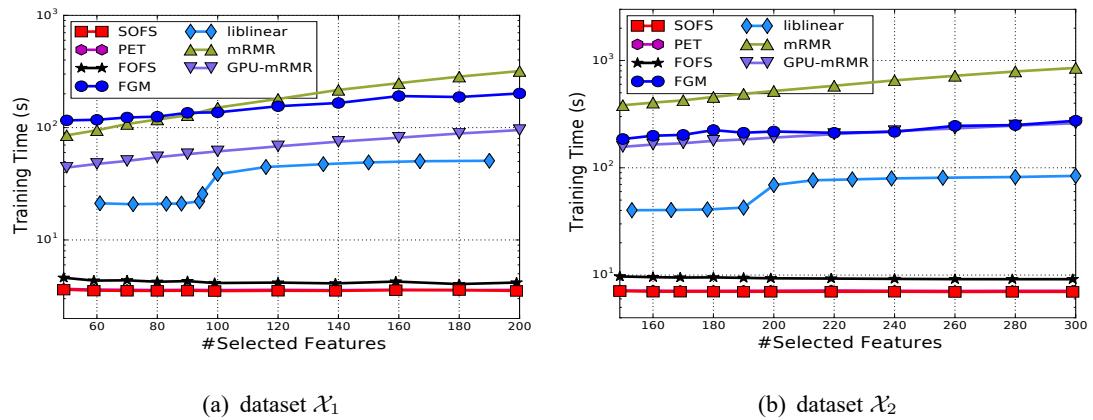


图 4.2 合成数据集 \mathcal{X}_1 和 \mathcal{X}_2 上训练时间和特征数目之间的关系

表 4.2 SOFS 算法可伸缩性评测

	训练时间 (s)			准确率 (%)			模型稀疏度 (%)		
	OGD	AROW	SOFS	OGD	AROW	SOFS	OGD	AROW	SOFS
\mathcal{X}_1	3.58	3.59	3.52	98.44	98.48	99.17	0.00	0.00	99.00
\mathcal{X}_2	7.06	7.02	7.00	97.83	98.52	98.62	0.00	0.00	99.00
\mathcal{X}_3	114.82	130.72	132.94	99.39	99.55	99.56	83.16	72.22	99.99

特征选取算法在 \mathcal{X}_3 上需要耗费几个小时甚至几天才能完成特征选取，实验仅在 \mathcal{X}_1 , \mathcal{X}_2 和 \mathcal{X}_3 上测试 SOFS 算法是否能够处理增长的维度和规模，特征选取的数目分别固定为 $B = 100, 200, 500$ 。此外，实验将 SOFS 算法与两个全部维度上的在线学习算法比较，验证 SOFS 算法的有效性。作为基准算法的两个在线学习算法分别是在线梯度下降算法 (OGD) 和自适应权重向量正则化算法 (AROW)，比较结果如表 4.2 所示。

根据表中的结果可以发现，测试准确率相对于基准算法有所提高，说明移除不相关特征确实可以提高模型的准确率。更重要的是，SOFS 只需要少于 1% 的特征就可以达到这个准确率。快速有效的特征选取有如下三个好处：(1) 当输入特征是密集数据时，稀疏的分类器可以显著减少预测时间；(2) 可以显著减少预测时的内存占用；(3) 可以显著减少特征提取的时间。在该数据集上，OGD 和 AROW 算法需要大约 1GB 内存存储分类器（每个权重需要 4 个字节），而 SOFS 算法仅需要 2 KB。在嵌入式系统等内存空间十分有限的条件下，紧凑的分类器更加具有实际意义和经济价值。

此外，随着数据数目和特征维度的增加，SOFS 训练时间的增加在可接受的范围。在十亿个特征的数据集上，仅需要 2 分多钟的时间就可以完成模型训练和特征选取，其他特征选取算法陷入维度灾难的问题。例如，PET 算法需要至少 10 个小时从 \mathcal{X}_3 中选取 500 个特征，其他复杂度更高的算法的训练时间更久。此外，

表 4.3 中等规模特征选取数据集详情

数据集	relathe	pcmac	basehock	ccat	aut	real-sim
特征维度	4,322	7,510	4,862	47,236	20,072	20,958
训练数据数目	1,000	1,000	1,500	13,149	40,000	50,000
测试数据数目	427	946	493	10,000	22,581	22,309
特征个数	87,352	55,470	101,974	994,133	1,969,407	2,560,340

相比于在线学习算法，SOFS 并没有引入过多额外的时间开销。原因在于在具体实现中数据加载和模型训练分为两个线程同时进行。由于三个算法都比较高效，数据加载占据了主要的时间。总结起来，实验中的低训练时间和高准确率表明本章提出的算法能够快速有效地挖掘大规模超高维度数据中的有效特征。

4.5.3 中等规模真实数据集实验评估

本节在中等规模公开数据集上评测在线特征选取算法。数据集详情如表 4.3 所示。数据集可以从亚利桑那州立大学特征选取网站¹ 或 SVMLin 网站² 下载。

准确率：图 4.3 显示了不同算法测试准确率的比较结果。在线特征选取算法中，除了在某些数据集上选取的特征数目过少时，SOFS 的准确率均高于 PET 和 FOFS 算法。该结果与在合成数据集上的结果相似。相比于批处理算法，当选取的特征数目足够多时，SOFS 能够获得当前最好的 FGM 算法相近甚至更好的准确率。FGM 算法在特征数目很少时效果很好。liblinear 在本实验中表现出了很有趣的现象。测试准确率首先随着特征数目的增多迅速增长，当到达某个阶段以后又迅速的下降。该现象很可能是由于当选取的特征数目较多时， l_1 惩罚项的系数较小， l_1 -SVM 面临过拟合的问题。 $mRMR$ 算法在特征数目较少时效果较好。SOFS 的准确率随着特征数目增多迅速增加并超过 $mRMR$ 算法。然而， $mRMR$ 算法的整体效果比其他特征选取算法要差很多。

一般来说，批处理算法在选取的特征数目较少时效果更好。然而，随着特征数目增多，本章提出的 SOFS 算法迅速达到批处理算法相近甚至更好的准确率。

训练时间：图 4.4 显示了不同算法在中等规模数据集上训练时间的比较。首先，SOFS 算法需要最少的训练时间完成特征选取，尤其是在后三个数据维度较高的数据集上。其次，SOFS 相对于 PET 和 FOFS 有大于 10 倍的训练时间优势。此外，FOFS 的训练时间比 PET 长，尤其是在后三个数据集上。再次，在批处理算法中，liblinear 是最高效的，但仍需要 SOFS 几十倍的训练时间。FGM 的训练

¹<http://featureselection.asu.edu/datasets.php>

²<http://vikas.sindhwani.org/svmlin.html>

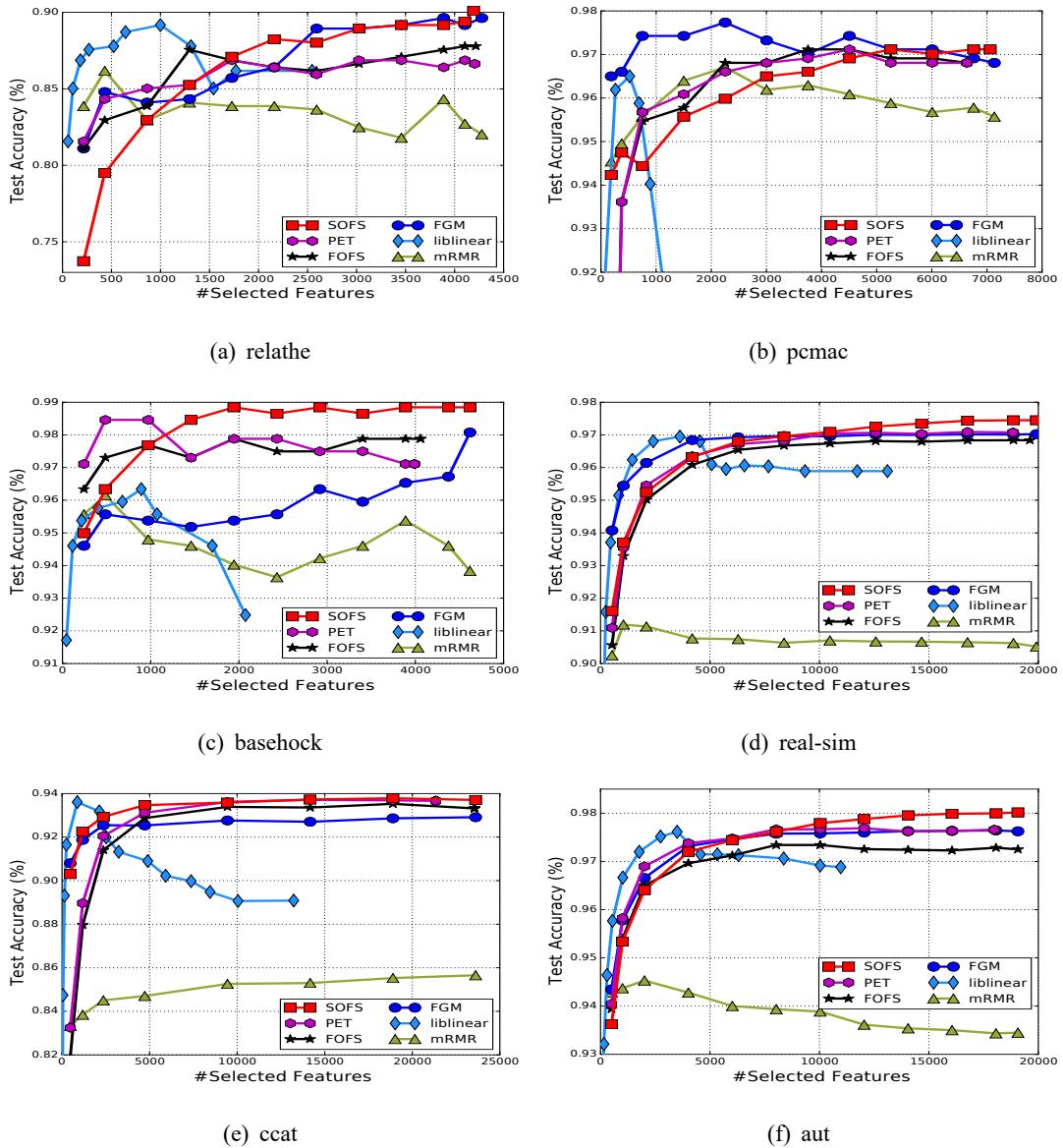


图 4.3 中等规模数据集上不同算法测试准确率比较

时间比 liblinear 大约高一个量级。计算最慢的算法是 mRMR。在“aut”数据集上，即使是并行的 mRMR 算法也需要超过 6,000 秒选取 10,000 个特征法。此外，只有当训练数据的数目超过数据维度时并行算法才能提高计算速度 (“real-sim” 和 “aut” 数据集)。

可以发现，中等规模数据集上的训练时间比较结果与 4.3.3 节分析的计算复杂度一致。SOFS 的复杂度与非零特征的数目成正比。虽然一阶在线特征选取算法的复杂度都与特征维度成正比，PET 仍然比 FOFS 高效。

4.5.4 物体识别实验评估

本节将多类 SOFS 算法应用到物体识别任务中。实验在 VOC2007 数据集^[110]，的 20 个物体类别上学习分类模型。实验首先从图片中将每个物体剪切成单独的

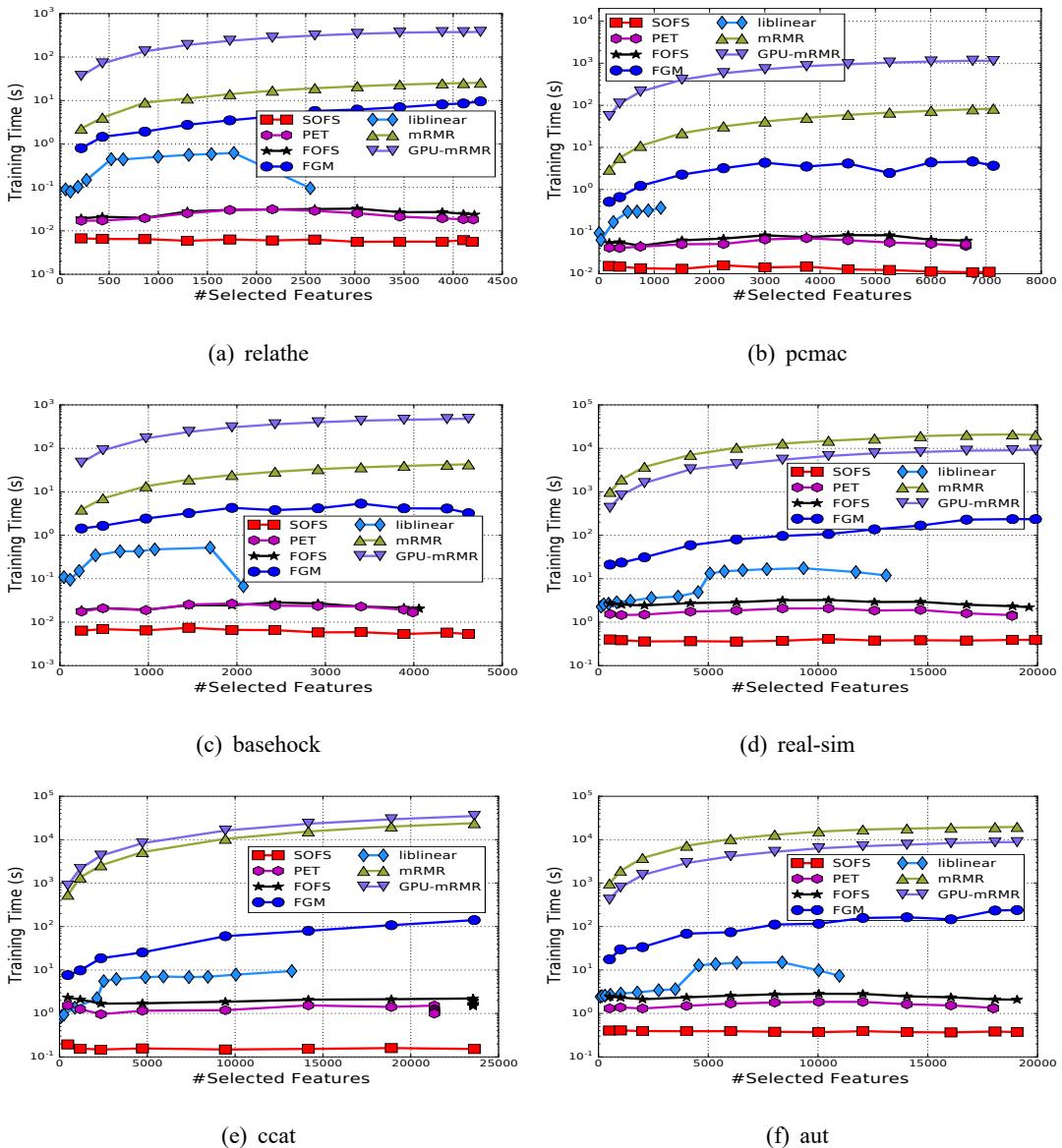


图 4.4 中等规模数据集上不同算法训练时间比较

图片，并将其中一半用作训练，剩余的一半用作测试。为了提取图片特征，我们采用了广泛使用的深度卷积神经网络——在 ImageNet 上预训练的 VGG16 模型^[24]。网络的最后两层全连接层的输出作为图片的特征表示。通过这些处理，我们获得了 12,315 个训练数据和 12,325 个测试数据，每个数据表示成 8,192 维的稀疏特征。特征的稀疏性是由于网络采用了校正线性激活函数^[23]。由于 FGM 算法只适用于二分类问题，这部分实验不考虑 FGM 算法。此外，由于本实验是多分类问题而 liblinear 只能使得多分类问题的权重稀疏而不是特征稀疏，实验也忽略了 liblinear 算法。

图 4.5 显示了 SOFS 算法在物体识别问题上相对于其他在线和批处理特征选取算法的分类的准确率和训练时间。可以发现，在所有算法中 SOFS 达到了最好的分类效果。随着选取特征数目的增多，SOFS 的测试准确率迅速增加，验证了

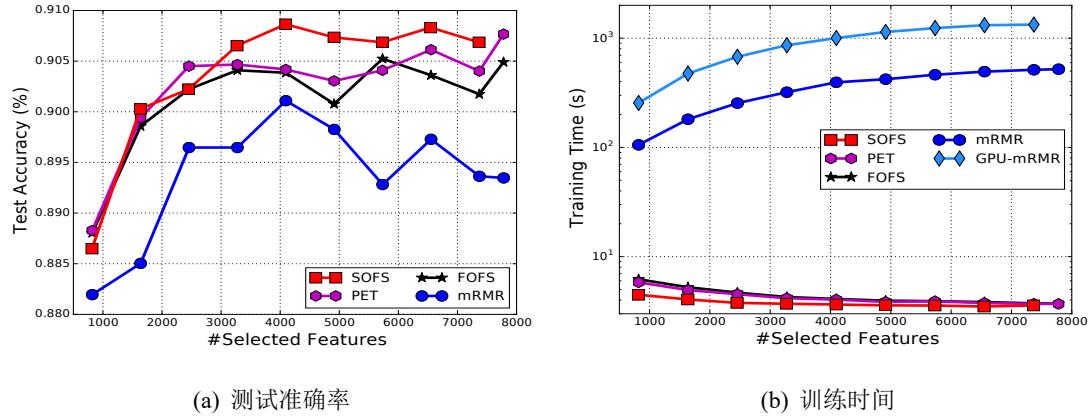


图 4.5 VOC2007 数据集上不同算法在不同特征数目下的测试准确率和训练时间比较

表 4.4 大规模真实数据集信息

数据集	特征维度	训练数据个数	测试数据个数	特征个数
news	1,355,191	10,000	9,996	5,513,533
rcv1	47,152	781,265	23,149	59,155,144
url	3,231,961	2,000,000	396,130	231,249,028

SOFS 算法在计算机视觉任务上的有效性。在 VOC2007 数据集上, mRMR 算法的效果最差。另一个发现是 SOFS 算法在 4,000 个特征左右达到最好的准确率, 大约是所有特征数目的一半。考虑到特征是深度卷积神经网络的全连接层的输出, 减少特征数目也意味着可以减小网络神经元的数目。图 4.5(b) 再次验证了在线特征选取算法的效率。

4.5.5 大规模真实数据集实验评估

本节在大规模真实数据集上评测 SOFS 算法的性能, 采用的数据集如表 4.4 所示。第一个数据集“news”维度较高, 第二个数据集“rcv1”规模较大, 第三个数据集“url”规模和维度均较大。出于训练时间考虑, 本实验仅比较 SOFS 算法, PET 算法(快速)和 FGM 算法(高效)之间的差异。

表 4.5 和图 4.6 显示了三个算法测试准确率和训练时间之间的比较结果。由于 FGM 算法在“url”数据集上训练时间过久, 因此表格中缺少相关实验结果。根据表格结果, 可以发现 SOFS 的效果十分接近甚至优于 FGM 算法, 尤其当足够多的特征被选取时。SOFS 算法和 FGM 算法的准确率都高于 PET 算法。对于训练时间, PET 和 SOFS 在“news”数据集上的比较结果表明 PET 对于维度更敏感。一个有趣的现象在于 PET 算法选取 0.5% 的特征时需要耗费更多的时间, 原因在于此时 PET 算法收敛的速度过慢, 需要反复地更新模型。FGM 算法是计算最复杂的特征选取算法, 训练时间至少比在线特征选取算法高一个量级。此外, 训练

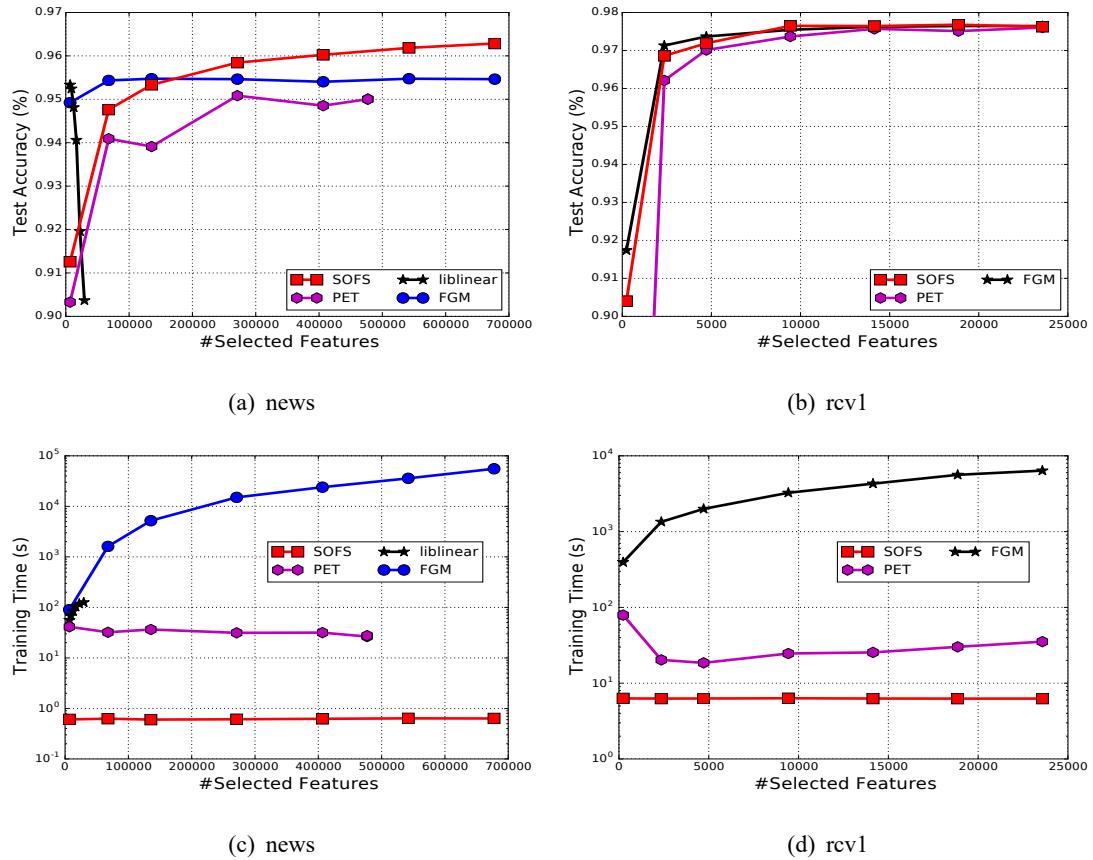


图 4.6 “news” 和 “rcv1” 数据集上测试准确率和训练时间与特征数目之间的关系

表 4.5 大规模高维数据集上不同特征选取算法比较 (ρ 是选取的特征比例)

Dataset	ρ	0.005	0.05	0.1	0.2
news	PET	90.33%(41.34s)	94.09%(32.18s)	93.91%(36.54s)	95.08%(31.37s)
	SOFS	91.26%(0.61s)	94.76%(0.63s)	95.33%(0.60s)	95.84%(0.61s)
	FGM	94.92% (90.10s)	95.43% (1610.53s)	95.47% (5206.20s)	95.46%(15055.28s)
rcv1	PET	73.18%(79.13s)	96.21%(20.30s)	97.01%(18.53s)	97.37%(24.63s)
	SOFS	90.40%(6.29s)	96.86%(6.27s)	97.19%(6.28s)	97.65%(6.32s)
	FGM	91.74% (394.98s)	97.13% (1346.03s)	97.37% (1994.78s)	97.54%(3253.97s)
url	PET	98.15%(1100.28s)	98.38%(1664.15s)	98.21%(1528.01s)	98.21%(1573.35s)
	SOFS	98.32%(6.95s)	98.74%(7.05s)	98.92%(6.94s)	99.18%(6.94s)

时间随着选取特征数目的增加也迅速增加。根据实验结果，可以发现 SOFS 算法在大规模高维数据集上选取特征的巨大优势。实际问题中，往往需要在相同数据集上反复运行多次在线算法使得模型收敛或选取参数，此时 SOFS 算法的优势将更加明显。

4.6 深度卷积神经网络模型简化

近年来深度卷积网络在目标识别、物体检测、物体分割等领域都获得了巨大的进步，同时网络的宽度、深度、参数的规模也在逐渐增加。表 4.6 显示了当前主流的神经网络的深度和参数个数比较，如 VGG16 网络包含超过 138M 的模型

表4.6 主流深度卷积神经网络的深度和参数规模

网络	CaffeNet	VGG16	VGG19	Inception	Inception-v3	Resnet-50	Resnet-101	Resnet-152
深度	8	16	19	22	59	50	101	152
参数个数	60M	138M	144M	11M	24M	26M	45M	60M

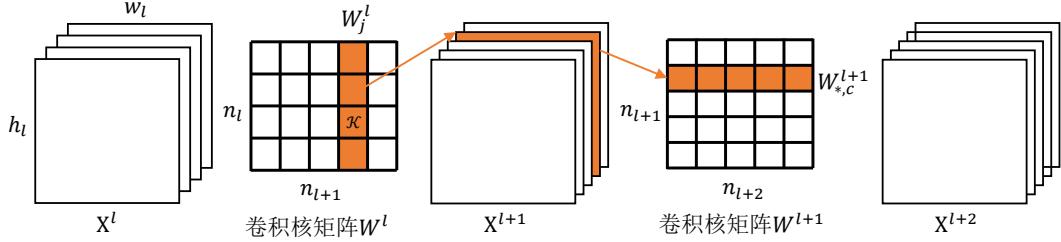


图4.7 移除卷积核对深度卷积神经网络的影响示意图

参数。大量的模型参数意味着在实际应用中需要大量的计算资源和时间，极大地限制了深度神经网络在大规模社交多媒体数据上的应用。此外，深度网络在移动设备上的应用已经成为一种趋势。由于移动设备计算能力的限制，在不影响模型准确率的条件下简化深度网络模型已经成为迫切的需要。本节提出一种基于在线特征选取的模型简化算法，极大地减少了模型的参数。

4.6.1 深度卷积神经网络模型简化

当前主流的深度卷积神经网络倾向于使用更多的卷积层和较少的全连接层，因此本节主要研究卷积层参数的简化方法。假设第 l 个卷积层输入的通道数为 n_l ，高度为 h_l ，宽度为 w_l ，输入被表示为三维特征图 $X^l \in \mathbb{R}^{n_l \times h_l \times w_l}$ ，卷积操作将输入的三维特征图映射为新的三维特征图，表示为 $X^{l+1} \in \mathbb{R}^{n_{l+1} \times h_{l+1} \times w_{l+1}}$ ，作为下一个卷积层的输入。映射通过 n_{l+1} 个卷积核与输入特征图卷积得到，假设第 l 层卷积核参数表示为 $W^l \in \mathbb{R}^{n_{l+1} \times (n_l \times k \times k)}$ 。卷积操作表示为：

$$X_j^{l+1} = \sum_{c=1}^{n_l} W_{j,c}^l * X_c^l \quad (4.14)$$

卷积操作的运算量为 $n_{l+1}n_lk^2h_{l+1}w_{l+1}$ 。

卷积神经网络的简化可以从卷积核、通道以及卷积核内部等多个角度进行稀疏优化。本节主要研究卷积核的结构化稀疏方法，移除整个卷积核减少模型参数个数，加快模型运算速度。如图 4.7 所示，如果移除卷积核 W_j^l ，则输出特征图的一个通道 X_j^{l+1} 也被移除，可以减少 $n_lk^2h_{l+1}w_{l+1}$ 的运算量。由于输出通道数的减少，下一层卷积层也可以减少一个维度的卷积核，进一步减少 $n_{l+2}k^2h_{l+2}w_{l+2}$ 的运算量。已有算法通常采用组稀疏的方法实现结构化稀疏：

$$W^* = \arg \min_W \mathcal{L}(X; W) \quad \text{s.t.} \quad \sum_{j=1}^{n_l} \|W_j^l\|_2^p \leq \lambda_l, \forall l \in [1, L] \quad (4.15)$$

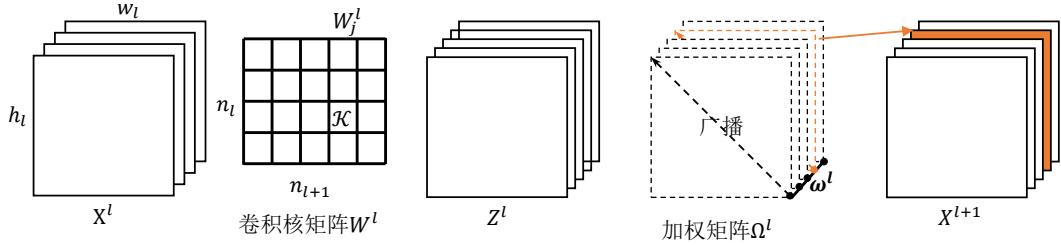


图 4.8 深度卷积神经网络辅助权重层模型简化

其中 p 取 0 时为 L_0 约束，取 1 时为 L_1 约束。组稀疏优化相对比较复杂， L_0 约束是非凸优化， L_1 约束不能显示地控制卷积核的个数。因此，本文提出了基于在线特征选取的深度卷积神经网络模型简化方法。

4.6.2 基于在线特征选取的模型简化

目前，深度网络的训练采取在线优化方法，本节提出将在线多维组稀疏优化问题转化为一维在线特征选取问题，实现模型简化。如图 4.8 所示，在卷积层后引入辅助权重层，权重层参数是对应特征图每个通道的一维卷积核权重向量。引入辅助权重层以后，模型简化问题的目标函数为：

$$W^* = \arg \min_W \mathcal{L}(X; W, \omega) \quad (4.16)$$

$$\text{s.t. } \sum_{j=1}^{n_l} (\omega_j^l)^0 \leq \lambda_l, -1 \leq \omega_j^l \leq 1, \forall l \in [1, L]. \quad (4.17)$$

辅助权重层的前向运算为：

$$Z_j^l = \sum_{c=1}^{n_l} W_{j,c}^l * X_c^l \quad (4.18)$$

$$X_j^{l+1} = Z_j^l \cdot \Omega_j^l \quad (4.19)$$

$$\Omega_{j,:}^l = \omega_j^l. \quad (4.20)$$

初始条件下所有卷积核的权重 ω_j^l 为 1，模型训练过程中更新权重，每次更新后利用在线特征选取算法保留部分权重并将剩余权重设为 0，设为 0 的权重对应的卷积核响应不能传递到更深的网络，因而在模型收敛以后可以将对应的卷积核移除，实现模型简化。在线特征选取算法可以在训练过程中动态调整需要保留的卷积核，减小模型简化对模型性能的影响。

模型简化问题中梯度通过反向传播得到。假设反向传播到第 l 层权重层的梯度为 $G^{l+1} \in \mathbb{R}^{n_{l+1} \times h_{l+1} \times w_{l+1}}$ ，传回第 l 层卷积层的梯度为 ∇Z^l ，第 l 层权重层的

表4.7 VGG-BN 网络结构

卷积层	输入大小	卷积核数目	参数个数	卷积层	输入大小	卷积核数目	参数个数
conv1_1	32x32	64	1,792	conv4_1	4x4	512	1,180,160
conv1_2	32x32	64	36,928	conv4_2	4x4	512	2,359,808
conv2_1	16x16	128	73,856	conv4_3	4x4	512	2,359,808
conv2_2	16x16	128	147,584	conv5_1	2x2	512	2,359,808
conv3_1	8x8	256	295,168	conv5_2	2x2	512	2,359,808
conv3_2	8x8	256	590,080	conv5_3	2x2	512	2,359,808
conv3_3	8x8	256	590,080	总计	-	4224	14,714,688

梯度为 $\nabla \omega^l$, 权重层参数的二阶对角协方差矩阵为 Σ^l , 梯度反向传播方式为:

$$\nabla Z^l = G^{l+1} \cdot \Omega^l \quad (4.21)$$

$$\nabla \Omega^l = G^{l+1} \cdot Z^l \quad (4.22)$$

$$\nabla \omega_j^l = \Sigma_j^l \sum_h \sum_w \nabla \Omega_{j,h,w}^l \quad (4.23)$$

$$(\Sigma_j^l)^{-1} = (\Sigma_j^l)^{-1} + \frac{\sum_h \sum_w Z_{j,h,w}^l}{\gamma}. \quad (4.24)$$

4.6.3 实验结果和评估

本文在 Cifar10 数据集上利用批量归一化 (Batch Normalization, BN) 的 VGG 模型 (VGG-BN) 研究卷积层简化对于模型准确率的影响^[124]。根据 Sergey 等人的结果, 该模型在 Cifar10 上达到了 92.45% 的准确率^[124]。VGG-BN 网络包含 13 层卷积层和 2 层全连接层, 网络结构如表 4.7 所示, 参数个数为 15M。

本实验首先在 Cifar10 数据集上预训练 VGG-BN 网络, 达到了 92.78% 的准确率, 然后在预训练的模型上进行模型简化。简化在 Cifar10 训练数据集上使用 0.01 的学习率继续训练 20 轮。

图 4.9 显示了简化不同卷积层对网络准确率的影响。可以发现, 深层次的网络存在更多的参数冗余。对于最深的 6 层卷积层 (conv4_1 至 conv5_3), 即使移除了 95% 的卷积核, 模型的准确率几乎不受到影。根据表格 4.7 中结果, 最深的卷积层占据了主要的参数个数。因此, 稀疏化这些层能够显著减少模型的大小和计算量。此外, 可以发现随着网络深度的减小, 卷积层对于简化的敏感性也逐渐增强。一方面可能是由于浅层卷积层的卷积核个数本身比较少, 另一方面可能是由于浅层网络提取更多的浅层语义, Cifar10 数据集的内容比较简单, 深层次语义对于图片分类提升作用相对较小, 因而对于模型简化的敏感性也较低。

为了更清晰地表示不同卷积层受模型简化的影响, 表 4.8 显示了在模型准确率下降至原来的 99%, 95% 和 90% 时不同层对应的稀疏度。当可容忍的准确率下降比例为 1% 时, 前 7 层卷积层能达到 60% 左右的稀疏度, 后 6 层能达到 95% 的稀疏度。随着可容忍下降比例的提高, 网络的稀疏度也迅速增加, 到 10% 的

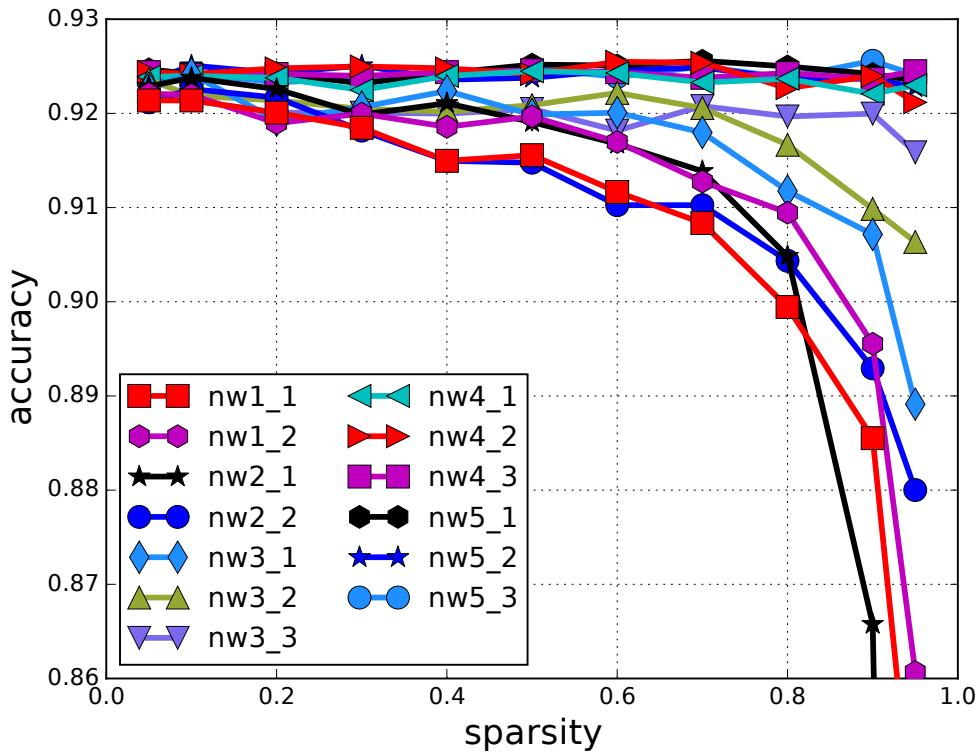


图 4.9 模型简化对模型准确率的影响

表 4.8 VGG-BN 网络不同卷积层在不同准确率下的稀疏度 (ρ 为下降百分比)

ρ	C1_1	C1_2	C2_1	C2_2	C3_1	C3_2	C3_3	C4_1	C4_2	C4_3	C5_1	C5_2	C5_3
0.01	60%	60%	40%	40%	60%	70%	90%	95%	95%	95%	95%	95%	95%
0.05	90%	90%	80%	90%	95%	95%	95%	95%	95%	95%	95%	95%	95%
0.10	95%	95%	90%	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%

下降比例时，几乎所有层都能达到 95% 的稀疏度或更高。

4.7 本章小结

本章主要解决大规模高维特征选取问题和深度卷积神经网络模型简化问题。特征选取是从所有特征中选取一小部分与具体问题相关的特征。本章提出了一个新的二阶在线特征选取算法 SOFS。区别于已有在线特征选取算法的复杂度与特征维度成正比，本章提出的算法的复杂度被显著减少到与每个训练样本的非零特征个数成正比。本章在中等规模和大规模的合成数据集和真实数据集上进行了充分实验，比较所提出算法相对于当前最好的批处理算法和在线特征选取算法的有效性和高效性。实验结果表明 SOFS 算法不仅能够显著减少训练的训练时间，还可以达到当前最好的批处理算法相近甚至更好的准确率，使得 SOFS 算法成为处理大规模高维数据的切实可行的特征选取算法。

此外，本章提出了深度卷积神经网络模型简化算法，将传统的多维卷积核组稀疏优化问题转化为一维特征选取问题，并基于在线特征选取算法给出了有效

的模型简化方案。实验部分分析了不同网络卷积层对于模型简化的敏感性，充分证明了模型简化的有效性。

本章的主要贡献包括：

- 提出了高效的二阶在线特征选取算法，用于解决大规模高维稀疏数据的特征选取问题。
- 提出了快速一阶在线特征选取算法和快速二阶在线特征选取算法，尤其是二阶算法的复杂度从与所有维度成正比降低到与非零特征个数成正比。
- 提出了高效的深度卷积神经网络模型简化算法，将多维卷积核组稀疏优化问题转化为一维特征选取问题。
- 进行了充分的实验，验证了所提出算法的有效性和高效性。

第5章 基于主题的照片集故事化表达

随着智能手机和智能移动设备的普及，人们几乎可以在任意时间任意地点拍摄照片。大量的照片被分享到社交网络以后，用户很少再次浏览或利用这部分数据，原因在于当前的社交多媒体缺少智能的服务去组织和挖掘用户数据。因此，本章提出一个基于主题的照片集故事化表达系统——Monet，总结整理用户照片集，并以故事化的表现形式重现照片集中的场景和事件。

本章首先介绍照片集故事化表达面临的主要问题和 Monet 系统的主要框架，然后依次介绍系统的两个主要组成部分：照片集分析与梳理和照片集故事合成，最后通过实验验证本章所提出系统的有效性。

5.1 主要问题与系统框架

从照片集中总结生成有吸引力和纪念意义的故事化表达面临一些挑战。首先，海量的用户照片集通常是无序的，浏览和分享照片的过程十分枯燥。然而大部分情况下，用户照片不是随机拍摄的，而是拍摄在不同事件的特殊时刻。对于故事化表达系统，挖掘照片集中的事件并将照片集按照事件进行组织十分必要。其次，用户会在事件中拍摄较多内容相似的照片，照片集存在数量多内容冗余的问题。此外，由于普通用户不具备专业的拍摄技术，很多用户照片存在质量较差的问题。因此，从照片集中选取一部分有代表性的子集是照片集故事化表达的关键步骤。再次，故事化表达系统需要将选取的照片用有吸引力的方式重新呈现给用户，提升用户体验。在本章提出的系统中，照片集通过音乐视频的方式表达。由于照片拍摄在不同的场景，系统需要自动选取合适的编辑风格达到不同的表达效果。最后，为了增加音乐视频的吸引力，应该在视频渲染时引入视频编辑元素，如视频特效、形状、颜色过滤器、转场等。从专业的视频编辑的角度，这些编辑元素依赖特定的内容和风格。因此，如何挖掘视频编辑语法，设计基于风格的编辑元素，并将它们运用到故事化表达系统中是一项具有很大挑战但又十分必要的部分。

根据以上总结的主要问题，本章提出的基于主题的照片集故事化系统如图 5.1 所示。系统首先检测照片集中的事件，然后在每个事件中选取一部分关键照片，完成对照片集的分析与梳理。在选取的关键照片上，系统自动为每张照片选取合适的编辑风格。根据预先定义的编辑语法，系统结合相机运动，视频特

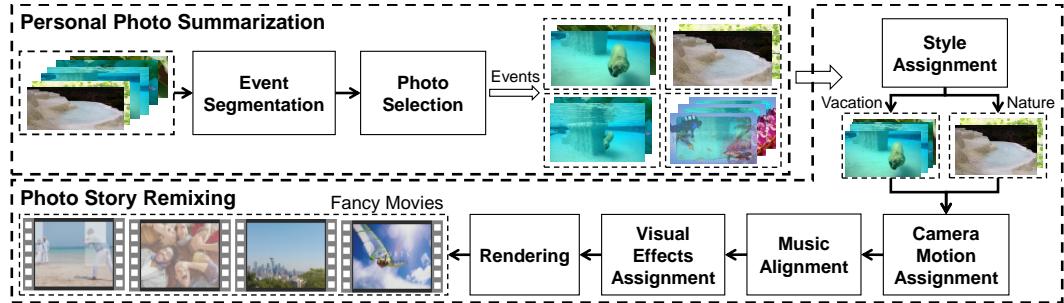


图 5.1 基于主题的照片集故事化表达系统框架

效和音乐产生视频片段。最后，通过在视频片段之间添加转场效果产生完整的音乐视频。系统的每个部分将在后续章节逐一介绍。

5.2 照片集分析与梳理

照片集分析与梳理包含两个步骤：事件检测和照片筛选。事件检测将照片按照事件进行组织。照片筛选首先移除低质量或者重复的照片，然后选取高质量的有代表性和均衡性的照片子集作为关键照片。

5.2.1 事件检测

从统计上看，用户倾向于在集中的时间点拍摄照片。为了清楚地解释事件检测的概念，本章定义“事件”如下：

定义 5.1 (事件) 事件是特定的情景下、相对较短的时间段内，用户记录值得留念的时刻的一组拍摄行为。

因此，同一事件中的照片在时间和地点上比较接近。用 N 和 K 分别代表照片集 \mathcal{X} 中照片和事件的数目，每个照片 $\mathbf{x}_i \in \mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ 属于事件 $e_j \in E = \{e_1, e_2, \dots, e_K\}$ 的概率可以表示为 $p(\mathbf{x}_i|e_j)$ ， $\mathbf{x}_i = (x_{i,1}, x_{i,2})$ ， $x_{i,1}$ 是时间 (\mathcal{T})， $x_{i,2}$ 是 $\text{GPS}(\mathcal{G})$ 。如果 $p(e_j|\mathbf{x}_i)$ 是最大后验概率，则 \mathbf{x}_i 被判定属于事件 e_j 。

给定事件 e_j ，假设时间和位置信息之间是独立的，似然函数 $p(\mathbf{x}_i|e_j)$ 表示为：

$$p(\mathbf{x}_i|e_j) = \prod_{l=1}^2 p(x_{i,l}|e_j) = p(\mathcal{T}_i|e_j)p(\mathcal{G}_i|e_j). \quad (5.1)$$

每个 $x_{i,l}$ 相对于事件 e_j 的概率服从高斯分布：

$$p(x_{i,l}|e_j) = \frac{1}{\sqrt{2\pi\delta_{j,l}^2}} e^{-\frac{(x_{i,l}-\mu_{j,l})^2}{2\delta_{j,l}^2}}. \quad (5.2)$$

事件检测的过程其实是学习方程 (5.1) 中混合高斯模型 (Gaussian Mixture Model, GMM) 参数的过程 $\Theta = \{\delta_{j,l}, \mu_{j,l}\}, j \in [0, K - 1], l \in \{1, 2\}$ 。优化的目标方程是

如下所示联合概率的对数似然函数：

$$l(\mathcal{X}; \Theta) = \log\left(\prod_{i=1}^N p(\mathbf{x}_i | \Theta)\right) = \sum_{i=1}^N \log\left(\sum_{j=1}^K p(e_j)p(\mathbf{x}_i | e_j, \Theta)\right), \quad (5.3)$$

$p(\mathbf{x}_i | e_j, \Theta)$ 通过方程 (5.1) 计算得到， $p(e_j)$ 是事件 e_j 的先验概率。本章提出的系统使用 EM 算法学习最优参数。系统用 K-means 算法的聚类中心初始化 GMM 模型的参数。为了决定事件的数目 K ，系统在不同数值的 K 上运用 EM 算法，产生一系列可能的事件分割结果。最优事件数目通过 Tao 等人提出的最小描述长度 (Minimum Description Length, MDL) 决定^[85]。

5.2.2 照片筛选

由于照片集中照片数量较多，并且多数照片是由没有专业拍摄技巧的普通用户拍摄的，系统需要滤除低质量和重复的照片，并进一步选取高质量、有代表性和均衡性的照片子集，得到照片集高质量的故事总结。

质量过滤：由于欠曝光、过曝光、模糊等问题，照片的质量可能十分低下。本章提出的 Monet 系统用 43 维人工设计的特征评价照片质量，包括：

- 暗度 (1D)，亮度 (1D)：欠曝光和过曝光的像素比例^[125]；
- 模糊度 (1D)^[126]，模糊差异 (1D)。模糊差异是指照片的模糊度与照片经过高斯模糊后的模糊度之间的差值。
- 锐度 (2D)^[126,127]，复杂性 (1D)^[128]，对比度 (1D)^[129]，动态范围 (1D)^[129]，景深 (1D)^[130]。这些特征是常见的照片质量评估的全局特征。
- HSV 分布 (12D)^[131]。首先将照片转换到 HSV 颜色空间，然后用非均匀量化将“hue”量化为 8 份，将“value”量化为 4 份。
- 最好块特征 (7D)，最差块特征 (7D)，主体块特征 (7D)。通过观察发现，某些情况下只有照片的一部分存在质量问题，导致整张照片被认为是低质量照片，此时全局特征并不能充分表示这种情况。因此系统将照片分成 5 个块（左上角、右上角、左下角、右下角和中间），从具有最大对比度、最小对比度和中间块提取常用的全局特征。

我们收集了一个包含 10,361 张高质量照片和 3,134 张存在质量问题的照片的数据集，所有照片都来自于用户拍摄的真实照片，并被人工标注为“good”或“bad”。利用上述 43 维特征，我们在这个数据集上训练 SVM 分类器。照片质量通过 SVM 模型预测的边界值评估，质量分数低于特定阈值的照片将会被移除。

重复照片过滤：为了更好地梳理照片集，需要检测内容重复的照片，并只保留每组内容重复照片中美学质量最高的照片。系统采用了 Winder 等人提出的局

部特征将每张照片表示为 64 维向量^[132]，向量的每个维度都是一个整数。照片的相似度定义为两个向量之间相同整数的个数。如果相似度大于某个阈值，则认为是重复照片。

关键照片选取：为了选取一部分照片子集作为关键照片代表整个照片集，系统考虑三个因素：美学质量、代表性和均衡性。

- **美学质量：**专业的视频制作需要选取美学质量很高的照片素材。本章采用了 Dong 等人提出的模型评价照片的美学质量^[133]。
- **代表性：**照片的代表性从两个方面进行评价：(1) 照片所在事件的重要性。当用户对某个事件更加感兴趣时，通常拍摄更多的照片，反之亦然。因此，对于事件 e_i ，如果事件中的照片数目为 n_i ，整个照片集中照片的数目为 N ，则事件 e_i 的重要性为 $\mathcal{E}I_i = \frac{n_i}{N}$ ；(2) 多样性。系统从事件中选取多样性最大的照片子集。通过提取照片的时间信息 ($\mathbf{t} \in R^1$)，位置信息 ($\mathbf{l} \in R^2$) 和颜色直方图 ($\mathbf{c} \in R^{64}$)，照片 \mathbf{x}_i 和 \mathbf{x}_j 之间的距离定义为：

$$d_{ij} = dist(\mathbf{t}_i, \mathbf{t}_j) + dist(\mathbf{l}_i, \mathbf{l}_j) + dist(\mathbf{c}_i, \mathbf{c}_j), \quad (5.4)$$

其中 $dist(\mathbf{a}, \mathbf{b}) = exp(-\frac{\|\mathbf{a}-\mathbf{b}\|^2}{\sigma^2})$ 。

对于照片 \mathbf{x}_i ，多样性定义为 $\mathcal{D}_i = \sum_j d_{ij} I_j$ 。 I_j 是一个指示函数，当 x_j 被选为关键照片时 $I_j = 1$ ，否则 $I_j = 0$ 。因此，照片 x_i 的代表性为：

$$\mathcal{R}_i = \mathcal{E}I_i + \mathcal{D}_i \quad (5.5)$$

- **均衡性：**美学质量和代表性是从照片的内容上进行关键照片选取，为了获得对照片集更加综合的总结概括，照片在时间上的均衡性也是一个重要的因素。系统利用照片之间的时间间隔的熵作为均衡性的表示。假设照片 \mathbf{x}_i 与 \mathbf{x}_i 前一个被选取的照片和下一个被选取的照片之间的时间间隔为 $t_{i,i-1}$ 和 $t_{i,i+1}$ ， \mathbf{x}_i 的熵定义为：

$$\mathcal{E}_i = t_{i,i-1} \log t_{i,i-1} + t_{i,i+1} \log t_{i,i+1} \quad (5.6)$$

照片被选取的适合度是上述三个因素的线性组合：

$$S_i = aQ_i + b\mathcal{R}_i + c\mathcal{E}_i. \quad (5.7)$$

其中， a, b, c 是满足 $a + b + c = 1$ 的非负加权参数。由于代表性和均衡性依赖于整体选取的照片，系统很难获取到关键照片选取的全局最优解。因此，系统用贪心算法获得关键照片子集。

5.3 照片集故事合成

视频制作通常包括 6 个步骤：(1) 选取合适的编辑风格和背景音乐；(2) 基于选取的风格添加照片；(3) 为每个静态照片设计运动效果将其转化为视频片段；(4) 对每个片段运用相机运动、形状、颜色过滤器、文本等视觉特效；(5) 为相邻片段选取转场效果；(6) 组合所有的片段和转场效果，添加开场和结束效果生成最终的音乐视频。一般来说，设计师需要决定编辑风格，并根据素材的语义内容设计视觉效果。

仿照上述专业视频制作的步骤，本节首先分析照片的语义内容。基于照片的内容，系统为不同照片分配不同的编辑风格，并选取合适的相机运动、视觉特效和转场等效果。

5.3.1 语义理解

在视频制作过程中，素材的语义特征和内容对于风格选取、运动设计，特效制作十分重要。Monet 系统采用了本论文提出的社交多媒体数据语义理解方法提取照片的语义特征。每张照片 \mathbf{x}_j 有 112 个可能的标签 (t_i) 以及相应的概率 $P(t_i|\mathbf{x}_j)$ ，这些标签和概率将会用来判断照片的编辑风格。

为了定义每个编辑风格可计算的编辑语法，这些标签被分成 20 个常见的用户照片类别，包括动物、建筑、暗黑、食物、室内、室外、物体、人物、多人、人群、植物、天空、文本、山和建筑。此外，人脸对于用户照片格外重要，系统检测照片中人脸的数目、性别、大小和位置。照片里的人脸特征分为“侧脸”、“一两个大比例人脸”、“一两个小比例人脸”、“三到五个小比例人脸”，“一组小比例人脸”和“一组大比例人脸”。性别信息包括“单个人”、“单个男性”、“单个女性”、“两个女性”、“两个男性”、“夫妻”和“人群”。所有这些特征信息将用于后续的生成视频片段和选取视觉特效等步骤。

5.3.2 风格选取

本节讨论如何将照片聚类到场景以及如何为这些场景选择合适的编辑风格，风格选取的流程如图 5.2 所示。

场景聚类：根据我们与设计师讨论的结果，用户通常在特定的场景中拍摄照片记录值得纪念的时刻。换句话说，照片不仅被时间和位置信息显式地组织，同时也被场景隐式地组织。本章提出的系统的主要目的是从用户照片中总结场景并生成音乐视频。为了衔接照片和场景之间的联系，每个照片用上节阐述的语义模型对应的概率向量表示 $\mathbf{p}_i = (P(t_1|x_i), P(t_2|x_i), \dots, P(t_N|x_i))$ 。我们用 Affinity

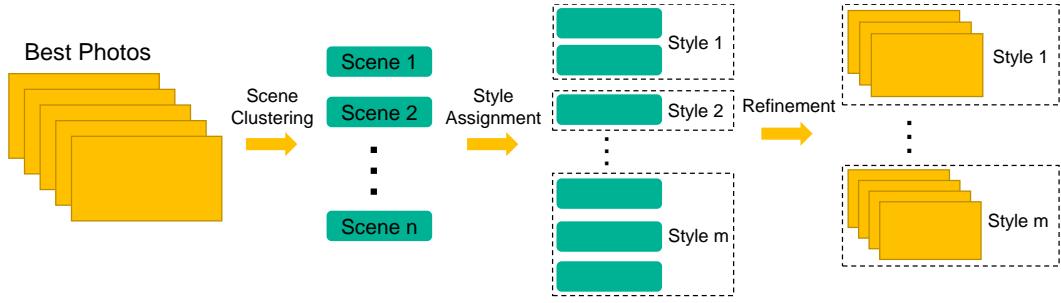


图 5.2 风格选取流程

Propagation^[134] 算法对照片聚类，每个聚类中心被当做是一个场景。

风格选取：在本章系统中，用不同的编辑风格表现不同的场景，使得故事化的表达更加智能更加具有吸引力。为了定义不同的编辑风格，我们邀请设计师设计了用户照片中最常见的主题风格以及相关的语义词汇，如“爱情”相关的词汇包括爱情、夫妻、甜蜜、婚礼、宝贝等。为了使得这些词汇表达的语义是可计算的，我们用这些词从 Flickr 上获取至少 5,000 张照片，并用本文提出的语义理解模型和特征选取算法提取照片的语义特征，训练多类 SVM 分类器区别不同的风格。相应的，照片 \mathbf{x}_k 属于风格 \mathcal{S}_j 的概率表示为 $P(\mathcal{S}_j|\mathbf{p}_k)$ 。

如果场景 $scene_i$ 包含 M 张照片， $scene_i = \{\mathbf{p}_1^i, \mathbf{p}_2^i, \dots, \mathbf{p}_M^i\}$ ，该场景属于风格 \mathcal{S}_j 的概率为：

$$P(scene_i, \mathcal{S}_j) = \sum_{k=0}^{M-1} P(\mathcal{S}_j|\mathbf{p}_k^i)/M. \quad (5.8)$$

最终，系统选取概率最大的风格编辑场景 $scene_i$ 。系统可能会给不同的场景选择相同的编辑风格，相同风格的照片将被合并到一起生成一个音乐视频。

5.3.3 生成视频片段

Monet 系统采用了 Hua 等人提出的从照片生成视频片段的方法^[88]，主要包括三个步骤：

- 关键帧选取。为了模拟相机运动，需要在照片中选取全景、中景和近景作为关键帧。
- 生成关键帧序列，决定关键帧的播放顺序。基于电影制作的规则，系统采用了 Hua 等人提出的 14 种播放策略产生关键帧序列^[88]。
- 生成相机运动。最终的视频片段通过在关键帧之间采用特定的相机运动实现。按照 Hua 等人提出的方法，系统构建了一个相机运动和关键帧之间的合适度矩阵，通过最大化相机运动合适度和相机运动分布均衡性为每个照片选取合适的相机运动。



图 5.3 视频特效、形状、颜色过滤器和转场样例

5.3.4 音乐分析

音乐视频需要将视频镜头切换和音乐节奏进行匹配。Monet 系统首先将音乐频率采样到 8kHz，并通过 Hua 等人提出的方法检测音乐的节奏^[88]。根据音乐的节奏和强度，系统检测最终音乐视频的切换点，决定视频片段的时长。切换点是指音乐视频从一个视频片段切换到另一个视频片段的时间点。切换点的检测算法将在 6.3.3 节中详细阐述。相比于已有的切换点检测算法，该方法提供了更好的音视频关联性，视频片段的切换频率和音乐节奏更加协调。

5.3.5 故事合成

专业的音乐视频需要对视频片段添加视频特效、形状、颜色过滤器、转场等视觉效果，使得音乐视频更加平滑更加具有吸引力^[135]。我们为每个主题风格总结了可计算的视频编辑语法并设计了风格模板，将视觉效果添加到视频片段中。最终的音乐视频根据视频片段、音乐和视觉特效生成。

5.3.5.1 风格模板设计

对于每个风格，我们设计了视频特效、形状、颜色过滤器和转场效果，每个视觉效果都依赖具体的照片内容。图 5.3 显示了一些视觉效果的样例。

不同的视觉效果有不同的表达效果，适用于不同的语义内容。比如，图 5.3(a) 中的条纹效果将照片分割成条纹依次展现，是“自然”主题风格的一种表现手法。该效果适用于户外拍摄的植物或天空，但不适用于包含人物的照片，因为没有用户愿意人物照片被分割成条纹。图 5.3(b) 的叶子形状中，黄色的树叶缓慢地漂浮在视频中，唤醒用户对于原始和旧时光的记忆，适用于包含草地或者树叶的照片。图 5.3(c) 的阳光颜色过滤器将照片和阳光颜色过滤器融合，能够帮助用户体验自然主题下的阳光照耀的效果。在图 5.3(d) 中，“聚会”风格的圆形转场逐步从中间到四周展现内容，能够有效地吸引观看者的注意力到照片中的主要物体或人物，特别适合照片中间仅有一个焦点人物或物体的照片，尤其是仅包含大比例人脸的照片。视觉效果的选取除了与照片内容相关，某些效果尤其是转场效果仅适用于特定的相机运动。总之，不同的视觉效果对于不同的语义内容和相机运动

有不同的合适度，我们为每个风格的视频特效、形状、颜色过滤器和转场定义了合适度语法。特效的语法规则定义为如下 xml 格式（形状和转场的语法类似）：

- **根节点 <Grammar>**：根节点包含风格信息以及视频特效、形状、颜色过滤器和转场语法的子节点。
- **特效节点 <Effect>**：特效节点描述视觉效果相对于不同语义特征和运动的合适度。每个特效节点包含一个或者多个“condition”子节点和一个可选的“percent”子节点。“condition”节点包含“feature”和“score”两个子节点。如果“feature”节点以“+”开头，则该特效适用于该特征，如果以“-”开头，则特效不适用。“score”子节点表示特效适用于该特征的程度。特效节点的“percent”子节点表明该特效在所有选择特效中适宜出现的比例。如果视频片段包含不适合的特征，特效的合适度为 0，否则特效的合适度为所有特征合适度之和。

根据特效语法，可以计算出每个视觉效果和每个视频片段的合适度以及部分效果期望出现的比例。

颜色过滤器的语法稍有不同。即使对于相同的视频片段，光照和颜色分布的变化也会导致不同的表达效果。因此，我们为不同的风格设计了不同的颜色过滤器，语法格式为：

- **视频颜色过滤器节点 <ColorFilter>**：该节点描述以视频形式出现的颜色过滤器的语法。“path”子节点包含视频的路径，“opacity”子节点控制过滤器视频和视频片段叠加时的不透明度，“overlay”子节点包含叠加的方式，“minPercent”和“maxPercent”控制该视频过滤器出现的最少和最大比例。
- **图片颜色过滤器节点 <ImageFilter>**：该节点描述图片颜色过滤器的语法。节点内容与视频颜色过滤器基本相同，除了“path”子节点被替换为“name”子节点，表明预先定义的图片颜色过滤器名称。此外，图片颜色过滤器节点包含**特效节点 <Effect>** 节点中的“condition”节点。

为了计算视频颜色过滤器和视频片段之间的合适度，系统首先提取颜色过滤器视频和视频片段的显著图 (Saliency Map)^[136]，合适度通过显著图之间的负相关系数得到。换句话说，视频颜色过滤器不应该将用户的注意力从照片的主题内容分散到颜色过滤器上。

5.3.5.2 视觉效果选取

视觉效果选取定义为一个优化问题进行求解。假设一共有 N_c 个视频片段，每个视频片段表示为 c_i ，一共有 N_e 个视觉效果，每个表示为 VE_j 。系统通过确定一

个选择矩阵 $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_c})$ 选取视觉效果， \mathbf{x}_j 是一个 N_e 维的仅包含一个非零元素的二值向量。由于不同的视觉效果有不同的期望出现比例 (percent)，首先需要确定它们的期望选择次数 n_j^* 。如果 VE_j 的期望比例为 p_j ，显然 $n_j^* = p_j N_c$ 。如果 p_j 没有指定，则根据它们与视频片段之间的合适度确定出现比例。假设视频片段 c_i 和视觉效果 VE_j 之间的合适度为 S_{ij} ， VE_j 的整体合适度为：

$$S_j = \sum_{i=0}^{N_c-1} S_{ij}(1 - I_j), \quad (5.9)$$

I_j 是一个指示向量，当 p_j 被指定时 I_j 为 1，否则为 0。 VE_j 期望出现的比例为：

$$p_j = \frac{S_j}{\sum_{k=0}^{N_e-1} S_k(1 - I_k)} p^*, \quad (5.10)$$

其中 $p^* = 1 - \sum_{k=0}^{N_e-1} p_k I_k$ 是没有被指定比例的视觉效果能够出现在视频片段中的剩余比例。

视觉效果选取定义为最大化整体合适度和出现比例的优化问题：

$$X^* = \arg \max_X \sum_{j=0}^{N_e-1} d_j \frac{\sum_{i=0}^{N_c-1} X_{ij} S_{ij}}{n_j}, \quad (5.11)$$

其中 $d_j = \exp(-\frac{(n_j - n_j^*)^2}{2})$ 是出现比例的合适度，用以衡量视觉效果的出现比例和预期出现比例之间的差异。

然而，仅仅最大化合适度和出现比例可能会导致相同的视觉效果被分配到连续的视频片段。为了保持视频片段原始的顺序，保证最终视频的故事性，同时防止这种连续效果的单一性，我们进一步约束视觉效果分布的均匀性。

假设 VE_j 在 N_c 个视频片段中出现 n_j 次， VE_j 应该均匀地出现在视频中。因此，选择 VE_j 的相邻视频片段之间的预期间隔为 $\frac{N_c - n_j}{n_j}$ 。假设选择 VE_j 的第 k 个和第 $k + 1$ 个视频片段之间的间隔为 δ_k ，则 VE_j 的均匀性分数 u_{jk} 定义为：

$$u_{jk} = \exp\left(-\frac{(\delta_k - \frac{N_c - n_j}{n_j})^2}{2}\right). \quad (5.12)$$

相应地，整体的均匀性分数为：

$$U = \sum_{j=0}^{N_e-1} \frac{\sum_{k=0}^{n_j-1} u_{jk}}{n_j^* - 1}. \quad (5.13)$$

考虑到合适度、出现比例和均匀性，视觉效果选取问题定义为如下优化目标方程：

$$X^* = \arg \max_{X_{ij}} \sum_{j=0}^{N_e-1} \left(d_j \frac{\sum_{i=0}^{N_c-1} X_{ij} S_{ij}}{n_j} + \lambda \frac{\sum_{k=0}^{n_j-1} u_{jk}}{n_j^* - 1} \right). \quad (5.14)$$

表 5.1 用户照片集详细信息

Dataset	User 1	User 2	User 3	User 4	User 5	User 6
#Photos	1080	481	496	564	702	866
#Events	95	32	40	107	28	58
#Best Photos	206	145	108	375	66	285

系统使用回溯法求解上述优化方程。

经过对用户照片集的事件检测、关键照片选取、视频片段生成、音乐匹配以及视觉效果选取，Monet 系统生成最终的音乐视频，完成对用户照片集的分析梳理，并按照主题对用户照片集进行故事化的表达。

5.4 实验结果和评估

据了解，目前还没有其他完整的系统能够自动对用户照片集进行分析与梳理并创建故事化的表达。为了评价 Monet 系统的有效性，本节从照片集分析与梳理和照片集故事合成两个角度评价 Monet 系统。

在实验中，所有的实验数据都从用户上传的照片集中选取。为了评价照片集分析与梳理的效果，我们让上传者对他们的照片进行标注，找出照片集中的事件和关键照片，并将用户照片上传到其他事件检测和关键照片选取系统，获取它们分析与梳理结果。对于故事合成，我们邀请用户从他们的照片集中针对每个主题风格推荐照片，然后针对每个风格将对应的照片和音乐提交到 Monet 系统和其他系统中，通过用户调查 (user study) 比较不同系统生成的视频。

5.4.1 事件检测和关键照片选取评估

我们邀请了 6 名用户分享他们在过去两年拍摄的照片，所有的照片都包含准确的拍摄时间，但只有一部分包含 GPS 信息。我们要求用户将他们的照片按照事件分组，并为每个事件选取最能代表这个事件的 1 到 6 张照片作为关键照片。这些用户标注的结果被当作评测事件检测和关键照片选取的真实数据。照片集的详细情况如表 5.1 所示。

实验采用了准确率 (Precision)、召回率 (Recall) 和 F-score 来评测事件检测的效果^[82]。准确率表示正确检测到的事件边界相对于检测到的事件边界的比例：

$$\text{Precision}_{seg} = \frac{\# \text{ 正确检测到的事件边界}}{\# \text{ 检测到的事件边界}}. \quad (5.15)$$

召回率表示正确检测到的事件边界相对真实事件边界的比例：

$$\text{Recall}_{seg} = \frac{\# \text{ 正确检测到的事件边界}}{\# \text{ 真实事件边界}}. \quad (5.16)$$

F-score 评价综合的性能：

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5.17)$$

关键照片选取通过准确率 (Accuracy) 评价：

$$\text{Accuracy}_{best} = \frac{\# \text{ 正确选取的关键照片}}{\# \text{ 真实关键照片}}. \quad (5.18)$$

事件检测的结果如表 5.2 所示。可以发现，Monet 在准确率和召回率上都比 PhotoToc 和 TEC 要好，因而也有更高的 F-score。高准确率和高召回率表明 Monet 不仅检测到相对容易的事件边界，也能找出较难检测的事件边界。对于关键照片选取，Monet 达到了 0.68 的准确率，高于目前现有的系统（如 Google+、OneDrive、Nokia StoryTeller 等），因而 Monet 能更准确地找出能够代表照片集的关键照片。

表 5.2 事件检测结果比较

Method	Precision	Recall	F-score
PhotoTOC ^[79]	0.50	0.71	0.59
TEC ^[82]	0.39	0.54	0.45
Monet	0.85	0.72	0.78

5.4.2 照片集故事合成评估

故事合成的客观评测比较困难，我们通过主观用户调研对系统进行评测。Monet 系统一共包含 10 个主题风格，我们邀请用户从他们的照片集中为每一个主题风格推荐照片。由于不是所有用户都有所有主题风格的照片，不同的主题风格对应的照片可能来自不同数目的用户。为了比较系统效果，我们为每个风格随机选取一个用户的照片作为测试数据。然后，我们将每个风格的照片以及相应的音乐上传到 Monet、Animoto、Magisto 和 Tilting Slide Show 系统中，最终生成 40 个音乐视频。在 Animoto 系统中，添加了视觉特效的照片在背景音乐下依次展现。Animoto 为照片之间添加了转场效果，但没有考虑相机运动。在 Magisto 系统中，照片伴随特效、转场、相机运动和背景音乐一起播放。在 Tilting Slide Show 系统中，照片伴随音乐的节奏以瓷贴的形式展现。

我们邀请了 20 名用户（12 名男性，8 名女性，用户的年龄从 22 岁到 28 岁不等）对一共 10 个主题 40 个视频进行打分。相同风格的视频在相同页面上按照随机顺序展现给用户。用户从以下方面对每个视频的满意度从 1 到 7（分数越高越好）打分：

表 5.3 故事生成主观评测结果

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Average
Magisto	5.83	5.67	5.33	5.78	5.67	5.22	5.05	5.54	5.51
Animoto	4.5	5	4.33	5.11	4.28	3.8	4.17	4.56	4.47
Monet	5.78	5.60	5.83	5.67	6	4.94	5.28	5.72	5.60
TS Show	-	3.62	3.22	3.05	2.9	2.5	2.56	2.83	2.59

- 问题 1：相机运动是否合理（对比专业视频）？
- 问题 2：视频片段之间的转场是否平滑？
- 问题 3：整个视频的视觉效果是否有吸引力？
- 问题 4：视频片段之间的转场是否与音乐节奏匹配？
- 问题 5：整个视频看上去是否像专业编辑的视频？
- 问题 6：视频是否讲述了有趣的故事？
- 问题 7：你有多少可能愿意分享该视频给你的朋友或社交网络？
- 问题 8：你对该视频的整体满意度是多少？

所有问题的用户打分以及平均分如表 5.3 所示。前 4 个问题是关于故事合成的具体方面。相比于 Tilting Slide Show 播放静态照片和 Animoto 简单的相机运动，Monet 和 Magisto 采用了更加专业的相机运动、更丰富的特效、形状和转场，因而它们都获得了更高的评分。Monet 在相机运动和转场上的评分略低于 Magisto，其中一个原因在于 Monet 根据当前照片的内容选取相机运动，转场根据当前的视频片段和它的上一个视频片段选取，视频片段切换点也基于马尔可夫假设确定（详见 6.3.3 节），这些方面都是基于局部信息的。我们猜测 Magisto 利用了一些全局信息或者优化方法来创建更加流畅和连续的视频，使得最终的视频编辑更具有一致性。尽管如此，Monet 仍然获得了和 Magisto 相差无几的评分。在视觉效果的吸引力上，Monet 获得了更高的分数，验证了本章提出的风格模板在故事合成中的有效性。

后续的四个问题是关于不同系统的整体评分。尽管 Magisto 在相机运动和转场上效果略好，Monet 在视频编辑的专业性上仍然获得了最高打分，再次表明 Monet 系统视觉效果设计和选取的优越性。由于 Magisto 实现了更加平滑的转场和相机运动，它在故事性上表现更优。尽管如此，用户在分享视频的评价上仍然更加倾向 Monet 系统。Monet 系统在整体满意度和平均满意度上也获得了最高的评分。这些结果表明：(1) 视觉效果对于生成专业的有吸引力的故事表达至关重要；(2) 在其他方面，Monet 与当前最好的 Magisto 系统表现相当。(3) 整体来说，Monet 在基于主题的照片集故事化表达上获得了最好的满意度。

5.5 本章小结

本文提出了一个从用户照片集中生成基于主题的故事化表达系统 Monet。该系统能够自动分析与梳理用户照片集并按照电影编辑语法和预先定义的编辑风格，以有吸引力的音乐视频的形式叙述照片集中的故事。该系统包含两个阶段：照片集分析与梳理和照片集故事合成。在照片集分析与梳理阶段，本章首先提出了一个新的多模态生成模型检测用户照片集中的事件。其次，系统根据照片的美学质量、代表性和均衡性选取一部分关键照片代表整个照片集。在故事合成阶段，为了叙述这些关键照片，系统检测照片的场景，并将场景分配到专业编辑人员预先定义的不同主题风格中，将属于同一个风格的照片重新组织起来。对于每一个照片，系统选取了特定的相机运动将它转化为动态的视频片段。为了提高最终表达的吸引力和平滑性，系统根据视频编辑语法和视觉内容将基于主题的视觉效果运用到视频片段中。最终，系统将视频和音乐混流生成音乐视频作为照片集的故事化表达。

本章的主要贡献包括：

- 提出了自动选取照片风格的模型。照片的风格选取是被当前故事化表达系统忽视的关键部分；
- 提出了关键照片选取算法，可以节省大量用户浏览和选取照片的时间。该部分也使得该系统成为首个全自动照片集故事化表达系统；
- 根据电影编辑语法定义了一系列主题风格模板，使得系统可以十分有效地运用视觉效果；
- 引入了多个维度的素材（音频，照片，视频）生成音乐视频表达。

第6章 移动多摄像头视频自动剪辑

随着移动设备的普及，用户几乎可以在任意时间任意地点拍摄视频记录他们经历的事件。虽然这些移动视频被大量分享到社交网络，由于视频角度的单一、内容冗余、声音嘈杂等问题，它们的观看体验十分有限。本章提出一个全自动移动多摄像头视频自动剪辑系统——MoVieUp，将同一事件中、多个摄像头从多个角度拍摄的时间上有重叠的一组视频剪辑成内容丰富、能体现专业编辑水平的单一音视频流。

本章首先分析移动多摄像头视频自动剪辑的主要问题，然后通过用户调研总结了一系列可计算的视频编辑语法。基于这些语法，本章提出了移动多摄像头视频自动剪辑系统的主要框架，并依次其中的三个主要步骤：音频剪辑、镜头切换点检测和视频镜头选取。最后，通过实验评估本章提出的 MoVieUp 系统各个部分的表现以及整体的用户体验。

6.1 主要问题

多摄像头视频是指在同一事件中，由多个用户从多个角度拍摄的时间上有重叠的一组视频^[91]。多摄像头视频的观看体验十分有限。首先，为了对整个事件有全面的了解，依次观看所有视频十分耗费时间。并且单个视频的角度单一，用户的观看过程十分枯燥乏味。其次，不同用户拍摄的内容可能十分类似，内容存在大量的冗余，使得多摄像头视频非常不利于收藏或分享。再次，移动多摄像头视频主要由业余用户在移动环境下拍摄的，视频的质量无法得到保证。为了解决这些困难，移动多摄像头视频自动剪辑将这些移动视频同步，并在每个时间段只选取一个视频，将一组视频剪辑成一个内容丰富、具有专业编辑水平的单一音视频流。图 6.1 是典型的移动多摄像头视频自动剪辑的示意图。给定某一事件的一组移动多摄像头视频，系统根据一系列剪辑规则剪辑生成单一的音视频流。每个时间段内选中的视频源在图中用深色表示。自动生成的剪辑视频提供了比任何单一视频更加丰富和更加专业的表达效果，显著提高了用户体验。

然而，移动多摄像头视频自动剪辑面临着一些困难。第一个困难来自于移动视频的质量。虽然拍摄设备已经获得了巨大的发展和提升，视频质量依然会受到抖动、模糊以及拍摄过程中其它因素的影响。此外，音频质量还受到周围环境以及麦克风本身的影响。据我们所知，目前很少有关于非侵入式 (non-intrusive) 音频质量评估的研究。另一个困难在于无法公式化地描述专业编辑人员剪辑视频

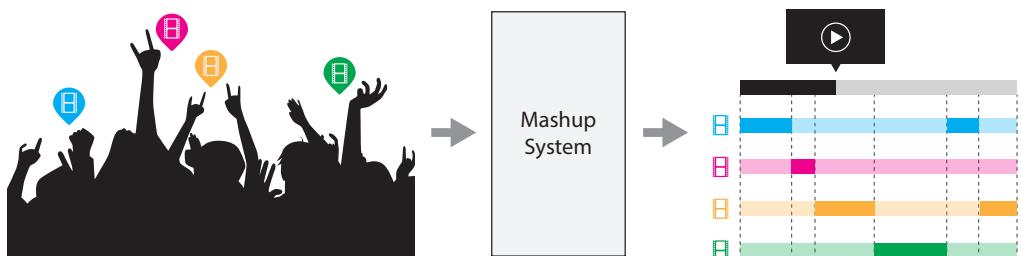


图 6.1 移动多摄像头视频自动剪辑示意图

的过程。不同的编辑人员有不同的编辑风格，很难学习通用的模型用于视频剪辑。即使对人工视频剪辑过程有了详细的了解，仍然需要将剪辑的过程转化成可计算的表达和算法，这个转化过程也是克服多摄像头视频剪辑两个基本问题的关键：何时应当从一个音/视频源切换到另一个音/视频源？切换时如何选择最佳的音/视频源？

6.2 可计算视频剪辑语法

移动视频自动剪辑和电影剪辑密切相关。电影剪辑人员运用专业的剪辑语法，通过影片的时间片段、阴影效果以及声音抓住用户的注意力，调动用户情绪，向用户讲述电影所要传达的故事。电影剪辑语法是指用来表述视觉形式、声音组合、它们存在和出现的功能以及播放时的相互关系的一套理论。因此，电影剪辑语法包含运动、声音、图像、颜色、电影符号、编辑、蒙太奇等元素^[137]。电影语法最基本的元素包括镜头、运动和距离（全景，中景，近景）^[98]。不同的镜头长度、拍摄角度以及距离远近可以传达不同的观感。例如，局部的物体运动可以通过中景镜头表达，而近景通常用来表达静态镜头或者缓慢运动的镜头。

移动视频自动剪辑模仿电影剪辑的过程来传达原始多摄像头视频记录的事件。移动多摄像头视频自动剪辑中，切换点是指镜头之间的边界时间点，镜头选取是指选取合适的视频流。为了创建内容丰富、剪辑专业的视频，首先需要了解电影剪辑语法以及剪辑人员如何运用这些语法到实际的剪辑中。

运用电影剪辑的一个主要难点在于电影剪辑并不是严格的准则。虽然已经有大量的工作讨论电影语法，本文的部分的结论也不是新的发现，总结已有的电影编辑规则，发掘新的与移动多摄像头视频剪辑相关的、并且方便转化成可计算规则的电影语法仍然十分必要，这也是本章系统的基础。我们查阅了已有的关于视频编辑的用户调研和论文，并进一步组织了关于视频剪辑两个基本问题的用户调研：

- 镜头/录音切换：何时该切换到另一个音频/视频源？
- 镜头/录音选取：切换时如何选择将要切换到的音频/视频源？

6.2.1 用户调研

我们邀请了一位艺术设计领域的教授和一名电影摄影专业的研究生参与了用户调研。该教授有 20 多年的视频相关领域的研究和从业经验，该研究生有着丰富的视频编辑经验尤其是和电视台的长期合作经验。

用户调研以讨论的形式进行。我们首先展示了典型的拍摄场景（如演唱会、比赛），移动多摄像头视频的主要问题（光照、抖动、遮挡等等）。我们向他们解释移动多摄像头视频剪辑的含义并咨询与之相关的问题。这些问题包括：

- **视频剪辑讨论：**该讨论主要调研切换频率和镜头选取。咨询的问题包括：视频镜头的时长是否有要求？哪些因素会影响到镜头时长？它们是如何影响的？这些因素与时长之间是否有确定的联系？镜头切换有哪些要求？哪些因素跟视频质量密切相关？如何避免移动多摄像头视频的单一性问题？如何能够平滑地切换视频镜头？是否有提高观看体验的建议？
- **音频剪辑讨论：**该讨论主要是关于音频的选取，问题包括：何时需要从一个音频源切换到另一个音频源？什么样的音频是比较好的？视频剪辑和音频剪辑之间的差异是什么？不同音频源的片段之间如何拼接？

6.2.2 视频剪辑调研结果

镜头切换。两个编辑人员认为视频镜头的时长应该有个范围，太短的镜头内容表达不完整，太长的镜头则显得枯燥。镜头时长不是一个固定的常数值。Shrestha 等人选取的演唱会视频镜头的长度为 3 秒到 7 秒^[138]。然而，镜头时长也不是严格固定的，更长的镜头可以用于运动镜头或者定场镜头，因为运动镜头不断拍摄新的内容，防止了内容的枯燥性，而定场镜头展示了完整的场景和丰富的内容，也不会导致内容单一引起的枯燥性。

镜头切换的频率跟音频和视频都比较相关。较高的切换频率适用于较快的音频节奏，剧烈的物体运动、以及快速的光照变化，平缓的镜头适宜配以较少的镜头切换。镜头切换频率和这些影响因素之间没有明确的关系，采用线性关系或者非线性关系取决于编辑者的编辑风格。

合适的切换点应该选择在说话或者歌唱的间歇。开始说话或者唱歌能够抓住观看者的注意力，并在结束时释放。为了避免打断观看者的注意力，在说话或者唱歌的间歇切换通常是比较好的选择。

镜头选取。与镜头选取相关的可计算的因素包括：视频质量、多样性、相机运动和语义完整性。

- **视频质量。**视频质量保证了视频的清晰度和观看体验的愉悦性。有关视频

质量的发现包括：太暗或者太亮的镜头应该被排除；应避免模糊的镜头；遮挡的镜头会破坏用户兴趣；倾斜镜头多数情况下会让观看者感到不适，虽然特殊情况下能够达到特殊的表达效果。两位参与者还特别提到由不规律或者不专业的相机运动引起负面效果（如手的抖动、快速运动等）的镜头必须排除。

- 多样性。多样性表示对事件丰富的内容表达。根据调研，视频剪辑中没有明确的准则用来选择相机的角度和距离。不同的剪辑人员有不同的剪辑风格，因而在有多个镜头可选时可能也会做出不同的选择。然而，镜头选取中也有一些不能违反的规则。比如，一个关键的准则是避免跳切：拍摄相同主体的两个相邻镜头的拍摄位置和角度不能过于接近。 30 度准则表明相邻镜头至少应该有 30 度拍摄角度的差异，从而避免镜头内容大量的重合。当相机位置不确定时，剪辑人员推荐应当选择镜头使得帧与帧之间的差别足够大，给观看者呈现新的内容。
- 相机运动。为了良好的观看体验，相邻镜头的相机运动应该尽量平滑，意外的相机运动会导致令人厌烦的视觉影响。一些通用的准则包括：(1) 静态镜头应该与静态镜头连接；(2) 运动镜头不适宜放置一起；(3) 可以通过减慢的相机运动消除运动镜头和静态镜头连接导致的部分视觉影响。
- 语义完整性。两位参与者提到视频剪辑语义上的一些考虑。每个用户拍摄的视频都是由一系列语义完整的部分（后文中我们称之为子镜头）构成。视频剪辑不应该破坏每个部分的语义完整性，切换时应该选取语义完整的镜头，否则镜头切换会显得十分突兀。

6.2.3 音频剪辑调研结果

不同与视频剪辑，音频切换的次数应该越少越好，我们称之为最小切换准则。在音频剪辑中不存在一直播放同一音频源导致的单一性问题。相反，即使多个视频之间在时间上精确同步，由于音频音量音色的差异，连接不同的音频源会导致不连贯的问题。这种不连贯性既有可能是麦克风本身也有可能是录音时的周围环境引起的，使得剪辑音频质量下降。因此，音频剪辑更倾向于从一个或少数几个音频源创建单一音频流。考虑到移动拍摄环境，应该避免嘈杂的音频，选择声音清晰干净的音频片段，音频剪辑应该具有鉴别高质量音频片段的能力。

6.2.4 可计算视频剪辑语法

根据用户调研，我们对可计算的视频剪辑语法做了总结。这些语法构成了本章节提出的基础。

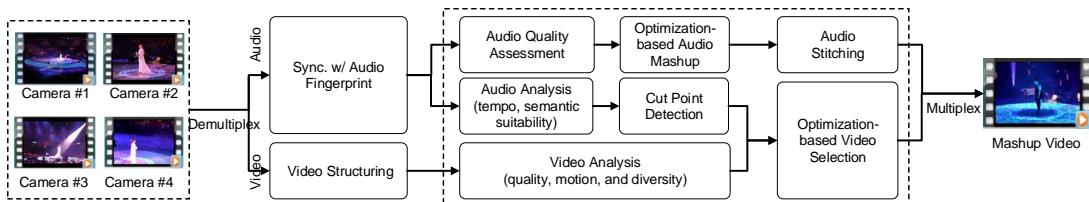


图 6.2 移动多摄像头视频自动剪辑系统框架

- 对于视频切换点检测，需要满足以下条件：(1) 镜头时长应该在一个范围内；(2) 镜头切换的频率应该与音频的节奏相匹配；(3) 切换点应该选择在说话或者唱歌的间歇。
- 视频镜头选取应该满足以下准则：(1) 镜头应该清晰稳定（没有模糊、遮挡、抖动等问题）；(2) 相邻镜头之间的帧差应该较大，避免跳切；(3) 切换点附近的相机运动应该平滑自然；(4) 每个被选择的镜头语义上应该完整。
- 音频剪辑应该满足：(1) 音频片段清晰干净；(2) 满足最少切换准则：音频切换的次数应该越少越好。

6.3 移动多摄像头视频自动剪辑系统

基于以上总结的可计算视频剪辑语法，本节提出 MoVieUp 系统用于解决移动多摄像头视频自动剪辑问题。后续章节中，视频和音频分别表示视觉信号和音频信号，用录像同时表示两个信号。

6.3.1 系统框架

MoVieUp 系统包含音频剪辑和视频剪辑两个部分。为了更清晰地阐述系统原理，定义以下术语：

- 子镜头：**包含连续相机运动和独立语义内容的基本视频单元。
- 镜头：**剪辑后的视频中来自于同一个视频源的连续子镜头。
- 切换点：**从一个信号源切换到另一个信号源的时间点。需要注意的是，在每个切换点，可能有多个信号源作为选择。

图 6.2 显示了 MoVieUp 系统的框架。给定一组录像，系统将它们分流成音频和视频、在时间上对录像同步、根据可计算的视频剪辑语法生成剪辑后的音频和剪辑后的视频，并混流成最终的单一音视频流。对于音频剪辑，系统在最少切换准则下根据音频质量选择音频片段。视频剪辑包含两个步骤：切换点检测和视频镜头选取。系统通过匹配切换频率和音频节奏并在说话或唱歌的间歇检测切换点。给定检测到的切换点，镜头选取在保证相机运动一致性的条件下最大化视频质量和多样性。质量和内容的分析在子镜头的粒度上完成。系统对视频剪辑的结

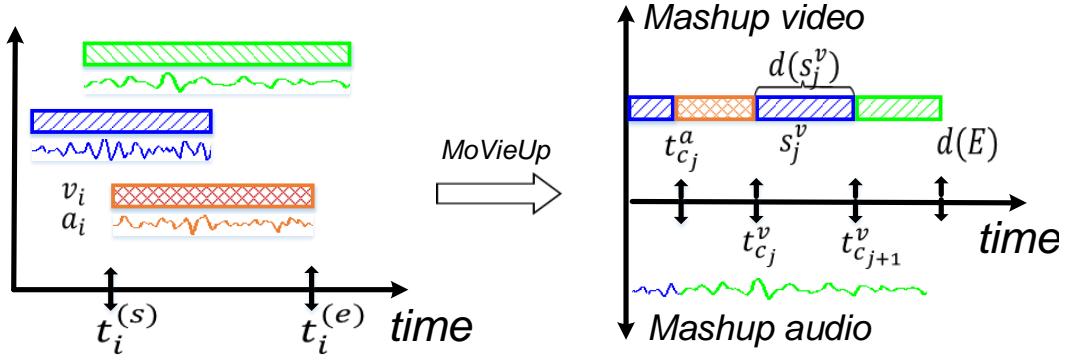


图 6.3 移动多摄像头视频自动剪辑系统符号表示

果进一步微调，从而满足语义完整性。视频抖动矫正作为可选的步骤可以进一步提高剪辑视频的观看体验。系统完成音频剪辑和视频剪辑以后，对两个剪辑结果混流生成最终的结果。

符号表示：假设一共有 N 个录像 $\mathcal{R} = \{r_1, r_2, \dots, r_N\}$ 。每个录像 r_i 被分流成音频 a_i 和视频 v_i 。 r_i 开始的时间记为 $t_i^{(s)}$ ，结束的时间记为 $t_i^{(e)}$ ， a_i 和 v_i 开始和结束的时间与 r_i 相同。第 j 个被选取的音频和视频片段表示为 s_j^a 和 s_j^v （第 j 个镜头），相应的时长表示为 $d(s_j^a)$ 和 $d(s_j^v)$ 。上标 a 和 v 用来区分音频和视频。剪辑后的音频和视频表示为：

$$\begin{aligned}\mathcal{M}^a &= (s_1^a, s_2^a, \dots, s_{M^a}^a) \\ \mathcal{M}^v &= (s_1^v, s_2^v, \dots, s_{M^v}^v).\end{aligned}$$

对于视频剪辑，每个镜头最大时长为 d_{max} ，最小时长为 d_{min} : $d_{min} \leq d(s_j^v) \leq d_{max}$ 。切换点 c_j 是视频从 s_{j-1}^v 切换到 s_j^v 的时间点，对应的时间记为 $t_{c_j}^v$ 。为了表示方便，事件开始的时间也认为是一个切换点，即 $t_{c_1} = 0$ 。图 6.3 表示了以上符号的含义。

预处理：系统首先对音视频做预处理，方便后续的分析和优化。音频被采样到 8kHz，视频帧率被采样到 25 帧/秒，分辨率被采样到 640×360 。

输入的一组录像记录了同一事件的不同时间段。为了进一步的处理，需要将录像在时间上同步，确定每个录像的起止时间 $t_i^{(s)}$ 和 $t_i^{(e)}$, $\forall r_i \in \mathcal{R}$ 。同步的基本假设是在事件的任意时间点至少存在一个录像。MoViewUp 采用了 Shrestha 等人提出的基于音频指纹的方法对录像同步^[138]。系统首先提取每个录像的音频指纹，通过比较计算出每一对录像之间的时间差，再通过投票的方法决定所有录像之间的时间差。

视频和音频剪辑都需要满足同步约束：被选取的信号源的开始时间必须早于当前切换点，结束时间必须晚于当前切换点。

$$t_{s_j}^{(s)} \leq t_{c_j} \leq t_{c_{j+1}} \leq t_{s_j}^{(e)} \quad (6.1)$$

当不满足上述约束时，对应的信号源将不作为当前切换点的备选信号源。

对于视频的处理有三个粒度：帧，子镜头和镜头。如同 Mei 等人的分析^[139]，帧并不是视频最具有信息量的语义单元，帧级别的操作不仅十分耗时，也不利于进一步的内容分析。镜头是由于用户开始和结束操作引起的物理结构，持续的时间相对较长，包含的内容不一定具有一致性。本章剪辑系统选取了包含连续相机运动和完备语义内容的子镜头作为基本视频单元，并采用 Kim 等人提出的基于颜色和运动阈值的方法对输入视频做结构分析，得到视频的子镜头^[140]。

6.3.2 音频剪辑

在多摄像头视频剪辑中，每个录音只记录了整个事件一段时间，音频剪辑是把所有录音综合起来生成事件的单一的完整的音频的过程。本章系统在最少切换准则下最大化音频质量剪辑音频。

根据用户调研的结果，音频剪辑的主要目的是最大化选取的音频片段的整体质量 $Q(\mathcal{M}^a)$ 。由于音频切换会降低最终生成音频的质量， $Q(\mathcal{M}^a)$ 并不是简单的各个音频片段的质量之和。通过仔细研究，我们发现每个移动设备拍摄的音频质量并不会频繁剧烈的抖动。基于这个发现，我们将最小切换准则具体化为一个音频切换的硬性约束。用 $q_{s_j^a}(t)$ 表示音频片段 s_j^a 在时间 t 的质量，仅当其它音频的质量明显好于当前音频时才发生切换：

$$q_{s_{j+1}^a}(t) > \gamma \cdot q_{s_j^a}(t), \quad (6.2)$$

γ 是音频切换的惩罚系数，在实验中设为 1.2。

系统采用贪心算法每隔一秒检查所有备选音频的质量，生成最终的剪辑音频。音频切换发生在当前音频结束时或另一个音频的质量明显好于当前音频时。

音频质量评估：上述方案需要评估音频的质量。据了解，除了一些在语音信号上的工作^[141]，很少有无参考的通用音频质量评估的解决方案。Li 等人将无参考音频质量评估建模为排序学习问题^[142]，该工作也是通用音乐质量评估的第一个工作。然而该方法并不适用于本章的音频剪辑场景，如方程 (6.2) 所示，音频剪辑需要有意义的音频质量评分。

通常，移动多摄像头视频拍摄的事件中音频信号的主体是说话或者歌唱。因此本章系统采用了 P.563^[143]——一种非侵入式语音质量评估算法评价输入音频的质量。系统用 5 秒滑动时间窗评价每秒的音频质量。

为了验证 P.563 算法是否适用于移动多摄像头视频的音频质量评估，我们随机选取了四个移动设备录制的音乐会录音，并下载了对应的音乐，称之为参考音频。我们用 P.563 算法分别评价这两种类型的音频质量，质量分数如图 6.4 所示。图中横轴是时间，纵轴是质量评分 (Mean Opinion Score, MOS)。左图中是四个移

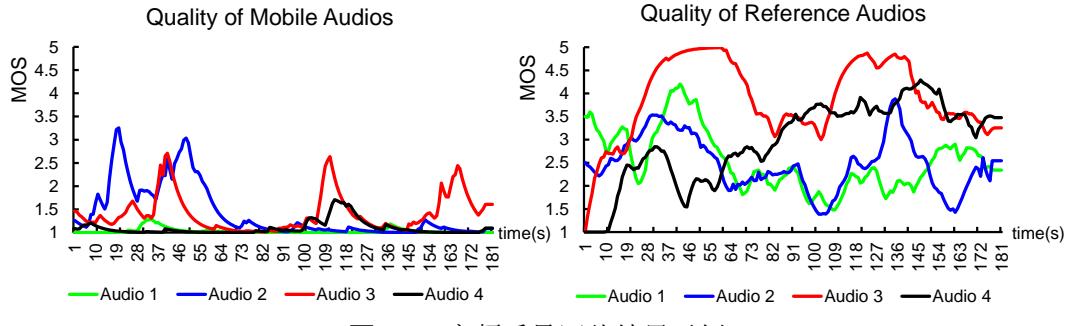


图 6.4 音频质量评估结果示例

动音频的质量评估结果，右图中是对应的参考音频质量评估结果。可以发现移动音频的质量评分明显低于参考音频的评分，验证了 P.563 算法在移动多摄像头视频场景下的有效性。音频剪辑的目标是选取图 6.4 所示的 Audio 2 开始阶段持续时间长的高质量音频片段。

音频拼接：选取音频片段以后，系统需要将音频片段拼接成单一音频。不同于视频镜头可以直接拼接，用户对声音的突然变化比较敏感。由于输入的录音是由不同设备在不同位置不同周围环境下录制的，直接拼接音频片段会导致音量和音色不一致引起的声音突变问题。为了克服这个问题，系统首先用直流矫正使得音频的音量增益相同，并使用类似与图片融合的基于拉普拉斯金字塔的算法将音频片段拼接成单一音频。

6.3.3 镜头切换点检测

视频剪辑是从输入视频中选取镜头并将它们拼接成内容丰富、具有专业观赏品质的单一视频流。视频剪辑分为两个步骤：切换点检测和视频镜头选取。本节介绍切换点检测，研究视频剪辑应该何时从一个视频源切换到另一个视频源。

根据可计算的视频剪辑语法，视频切换点检测需要综合考虑音频（音频节奏、说话/歌唱间隔）和视频（相机运动、子镜头完整性等）。MoVieUp 系统首先根据剪辑后的音频检测备选的切换点。对于视频，系统要求在备选的切换点处选取镜头时需要满足运动一致性和语义完整性。我们提出了两个合适度用于检测视频切换点：节奏合适度 $S^T(t)$ 和语义合适度 $S^S(t)$ 。节奏合适度衡量切换频率和音频节奏之间的关系，语义合适度避免打断说话或歌唱等行为。切换点检测假设当前检测点以前的切换点已经被确定，用数学形式表达为：

$$t_{c_j} = \arg \min_t \{ S^T(t|t_{c_{j-1}}) + S^S(t|t_{c_{j-1}}) \}, \quad s.t. \quad d_{min} \leq t - t_{c_{j-1}} \leq d_{max} \quad (6.3)$$

节奏合适度 $S^T(t|t_{c_{j-1}})$ ：视频切换频率应该与音频节奏匹配。快节奏的音频应该配以频繁的镜头切换，低频镜头切换更适合用在节奏缓慢的事件中。我们用

音频起始点 (onset) 之间的间隔近似音频节奏^[101,144]。如同在 6.2.2 节讨论的，音频节奏和切换频率之间没有明确的关系。我们将时刻 t 的节奏 $b(t)$ 线性地映射到预期时长 $d(b(t))$:

$$d(b(t)) = d_{max} - \frac{d_{max} - d_{min}}{b_{max} - b_{min}}(b(t) - b_{min}), \quad (6.4)$$

其中 b_{max} 和 b_{min} 分别是最大和最小音频节奏。

t 时刻的节奏合适度定义为:

$$S^T(t|t_{c_{j-1}}) = \left| \int_{t_{c_{j-1}}}^t \frac{1}{d(b(t))} dt - 1 \right|. \quad (6.5)$$

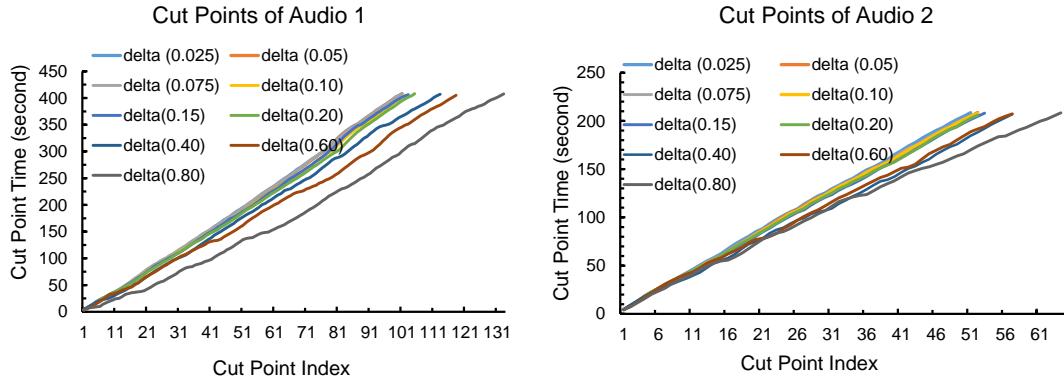
语义合适度 $S^S(t|t_{c_{j-1}})$: 语义合适度 $S^S(t)$ 表示应该避免让镜头切换分散用户的注意力。系统通过选择音频能量较低的时间点（如说话/歌唱间隔）作为切换点达到这个目的。语义合适度通过归一化到 $[0, 1]$ 范围内的音频能量 $e(t)$ 衡量:

$$S^S(t) = e(t). \quad (6.6)$$

切换点检测: 目标方程 (6.3) 是一个连续函数，由于节奏合适度和语义合适度都不是平滑函数，很难求出该目标方程的闭合解。因此，我们将时间离散化，离散化步长为 δ 秒，系统枚举从上一个切换点开始的 $[d_{min}, d_{max}]$ 范围内的所有可能时间点，目标方程的求解转化为:

$$\begin{aligned} t_{c_j} &= t_{c_{j-1}} + K\delta, \\ \text{where } K &= \arg \min_K \{ S^T(K) + S^S(K) \}, \\ S^T(K) &= \left| \sum_{k=1}^K \frac{\delta}{d(b(k))} - 1 \right|, \\ S^S(K) &= e(K) = e(t_{c_{j-1}^v} + K\delta), \\ b(k) &= b(t_{c_{j-1}^v} + k\delta). \end{aligned} \quad (6.7)$$

在理想情况下， δ 的值应该尽可能小，从而逼近连续的目标方程。为了选取合适的 δ 值，我们随机选取了一些音频并用不同的 δ 值检测切换点。图 6.5 显示了在两个随机选取的音频上检测切换点的结果，横轴为切换点的序号，纵轴为对应的时间点。我们发现当 δ 小于 0.2 秒时，检测到的切换点十分类似。因此，我们在实验中 δ 取值为 0.1。用 $d(e)$ 表示事件的时长，当 $d(e)$ 在几分钟的量级时，目标方程 (6.3) 需要枚举的时间点个数为 $d(e)/\delta$ ，一般不超过几千个检测点的量级，相应的计算复杂度也在合理的范围内。

图 6.5 不同 δ 取值对视频切换点检测的影响

6.3.4 视频镜头选取

视频镜头选取是在每个切换点确定如何选取应该切换到的视频源。本节将视频镜头选取建模为相机运动一致约束下的优化问题。

根据用户调研的结果，视频镜头选取需要综合考虑视频质量、多样性、相机运动和语义完整性。由于语义完整性依赖被选取的镜头，我们将会在后处理中考虑语义完整性。因此，视频镜头选取建模为相机运动一致性约束下最大化视频质量和多样性问题。具体来说，为了简化相机运动一致性的约束，我们禁止选取在切换点有相机运动的镜头。该简化方案能达到两个效果：(1) 减小连接运动镜头带来的视觉冲击；(2) 平滑相机运动的变化，避免打断相机运动。

用 $\mathbf{m}^-(s_j^v)$ 和 $\mathbf{m}^+(s_j^v)$ 表示第 j 个被选取的镜头 s_j^v 左右两边的相机运动，相机运动一致性表示为：

$$\mathbf{m}^-(s_j^v) = \mathbf{m}^+(s_j^v) = 0, \quad (6.8)$$

用 $Q(\mathcal{M}^v)$ 表示剪辑视频的质量， $D(\mathcal{M}^v)$ 表示多样性，视频镜头选取定义为：

$$\begin{aligned} \mathcal{M}^v &= \arg \max_{(s_1^v, \dots, s_{M^v}^v)} \{Q(\mathcal{M}^v) + D(\mathcal{M}^v)\}, \quad s.t. \\ \mathbf{m}^-(s_j^v) &= \mathbf{m}^+(s_{j-1}^v) = 0, \forall j \in [2, M^v]. \end{aligned} \quad (6.9)$$

视频质量评估：不同与音频剪辑，视频质量不会因为镜头切换而降低。视频剪辑的整体质量表示为：

$$Q(\mathcal{M}^v) = \sum_{j=1}^{M^v} Q(s_j^v), \quad (6.10)$$

其中 $Q(s_j^v)$ 是时间段 $[t_{c_j}, t_{c_{j+1}}]$ 内镜头 s_j^v 的质量分数。

MoViUp 采用了无参考的视频质量评估方法衡量移动视频的质量。视频质量从两大类别（时间和空间）六个方面进行衡量^[139]。时间因素，包括不稳定性 (unstability) 和急动 (jerkiness)，是由不规律的相机运动引起的。空间因素，包括

失真(inidelity)、亮度(brightness)、模糊(blurring)和倾斜(tilting)，是由于恶劣的拍摄环境引起的。系统在子镜头的粒度上评估六个方面的分数 $u_i^v \in [0, 1], i \in [1, 6]$ 。子镜头的不合适分数 U 和质量分数 Q 通过两个步骤计算得到：

- 预筛选低质量的子镜头。如果任一子镜头的某个质量分数低于对应的阈值，则认为该子镜头不适合被选取。系统将它的不合适度设为一个较大的值(在实验中设为 1,000)，从而保证在该子镜头持续时间内，仅有该子镜头可选时该子镜头仍会被选中。
- 基于准则的方法计算子镜头的整体不合适分数 U 和质量分数 $Q^{[139]}$ ：

$$\begin{aligned} U &= E(u^v) + \frac{1}{10 + 6\gamma} \sum_{i=1}^6 (u_i^v - E(u^v)), \\ Q &= 1 - U. \end{aligned} \quad (6.11)$$

其中 $E(u^v)$ 是子镜头六个方面质量分数的平均值， γ 是预先定义的常数值，用以调节 u_i^v 和 $E(u^v)$ 之间的差异对整体分数的影响。按照 Tao 等人在论文中的设定^[139]， γ 取值为 0.20。镜头质量 $Q(s_j^v)$ 取所有子镜头质量的最小值。

多样性：整体多样性是所有选取镜头的多样性之和：

$$D(\mathcal{M}^v) = \sum_{j=1}^{M^v} D(s_j^v), \quad (6.12)$$

其中 $D(s_j^v)$ 是选取的第 j 个镜头的多样性。多样性主要与用户观看过的仍停留在记忆中内容有关。系统根据多大程度上镜头能唤起用户记忆中的内容来衡量多样性。 $D(s_j^v)$ 的计算方式为：

$$D(s_j^v) = D(s_j^v, s_{j-1}^v). \quad (6.13)$$

类似于 Sundaram 等人提出的记忆模型^[97]，两个镜头之间的记忆和多样性通过对称的相邻两个子镜头 a 和 b 计算得到：

$$\begin{aligned} R(a, b) &= s(a, b) \cdot f_a \cdot f_b \cdot \left(1 - \frac{\Delta t}{T_m}\right), \\ D(a, b) &= 1 - R(a, b), \end{aligned} \quad (6.14)$$

其中 T_m 是记忆大小。 $s(a, b)$ 是相邻两个子镜头之间的相似度，在 MoVieUp 系统的实现中采用了传统低层次的特征表达和相似度度量方法 SSIM^[145]。为了进一步加强相似性度量，还可以收集常见的物体和场景类别，并从社交多媒体数据中收集相关的弱标注数据，利用本文提出的弱监督深度学习方法，学习这些类别和场景的深度识别网络，并利用本文提出的特征选取算法选取最能体现视频帧语义的特征，获取更好的镜头相似度的计算。 f_a 和 f_b 是子镜头长度相对于记忆大小 T_m 的比例。 δt 是两个子镜头的时长差。

镜头选取: 为了求解目标方程(6.9), 我们将约束条件表示成指示函数 $I(s_j^v)$, 当约束条件满足时 $I(s_j^v)$ 等于 0, 否则取一个很大的值作为惩罚项 (系统实现中取值为 1,000), 从而保证当且仅当该镜头为唯一选择时才会被选中。目标方程被重新表示为:

$$\mathcal{M}^v = \arg \min_{(s_1^v, \dots, s_{M^v}^v)} \left\{ \sum_{j=1}^{M^v} \{U(s_j^v) + I(s_j^v)\} + \sum_{j=2}^{M^v} R(s_j^v, s_{j-1}^v) \right\}. \quad (6.15)$$

上述目标方程可以定义成递归的形式。用 $f(s_m^v : s_n^v)$ 表示切换点 m 到 n (不包括 m) 的上述目标方程的最优值:

$$\begin{aligned} f(s_m^v : s_n^v) &= \min_{(s_{m+1}^v, \dots, s_n^v)} \left\{ R(s_m^v, s_{m+1}^v) \right. \\ &\quad \left. + \sum_{j=m+1}^n (U(s_j^v) + I(s_j^v)) + \sum_{j=m+2}^n R(s_{j-1}^v, s_j^v) \right\} \\ &= \min_{s_{m+1}^v} \left\{ U(s_{m+1}^v) + I(s_{m+1}^v) + R(s_m^v, s_{m+1}^v) + f(s_{m+1}^v : s_n^v) \right\}, \\ \text{where } &1 \leq m \leq n \leq M^v. \end{aligned} \quad (6.16)$$

递归方程(6.16)具有最优子结构性质, 为了优化 $f(s_m^v : s_n^v)$, 首先需要针对每一个可能的 s_{m+1}^v 优化 $f(s_{m+1}^v : s_n^v)$ 。利用最优子结构性质以及递归方程, 系统通过动态规划获得最优解。

假设存在虚拟的唯一初始镜头 s_0^v 并满足所有约束条件。初始目标方程(6.15)可以利用动态规划和方程(6.16)求解得到。

$$\begin{aligned} \mathcal{M}^v &= \arg \min_{(s_1^v, \dots, s_{M^v}^v)} f(s_0^v : s_{M^v}^v), \\ U(s_0^v) &= I(s_0^v) = 0, R(s_0^v, s_1^v) = 0. \end{aligned} \quad (6.17)$$

后处理: 根据用户调研的结果, 每个录像都是由许多语义完整的子镜头构成的。上述剪辑步骤没有考虑语义完整性。切换点可能出现在子镜头中间, 使得被选取的子镜头太短, 破坏了语义完整性。此外, 应该进一步减少不规则相机运动引起的抖动, 提高用户体验。

因此, 系统对剪辑结果进行两步后处理操作: 语义完整性和视频抖动矫正。对于语义完整性, 系统对切换点附近的子镜头增加时长约束: 如果某个子镜头持续的时间少于一秒, 则对切换点的位置进行微调使得子镜头的时长满足约束。对于视频稳定性, 系统采用了 Liu 等人提出的视频抖动矫正方法减小移动视频抖动的影响^[146]。

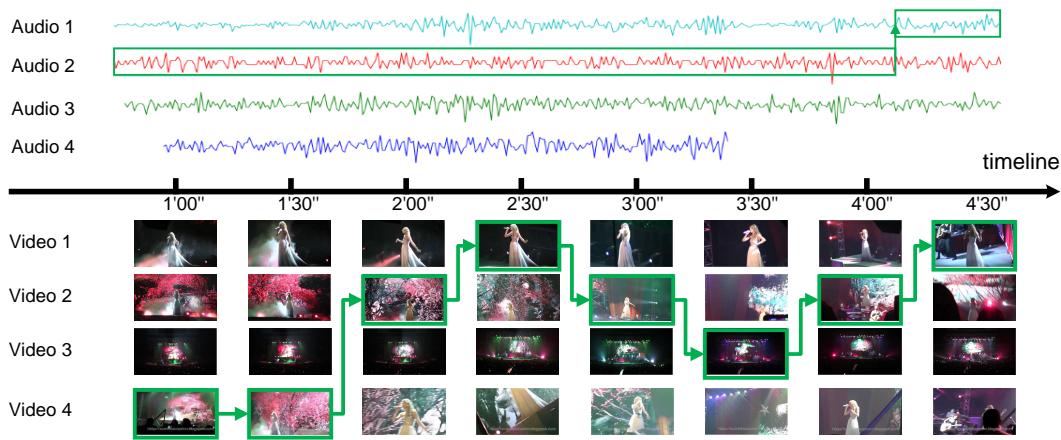


图 6.6 移动多摄像头视频自动剪辑示例

6.4 实验结果和评估

本节评价本章提出的移动多摄像头视频自动剪辑系统的效果。图 6.6 显示了移动多摄像头视频剪辑的一个样例。时间轴上方是移动设备录制的备选音频，波形表示音频的能量曲线。下方是移动设备拍摄的备选视频，视频通过每隔 30 秒采样的帧来表示。系统从这些备选音频/视频中选取片段/镜头生成最终的剪辑音频/视频，如图中绿色方框所示。

本实验从三个方面对系统进行评估：(1) 音频剪辑是否提高了最终的音频质量。实验与 Virtual Director 系统^[91]进行对比，该系统在选取视频镜头的同时选取对应的音频组成最终的音频输出；(2) 相比于人工选取的切换点以及 Mukesh 等人提出的基于学习的方法^[94]，本章提出的切换点检测算法效果如何？(3) 本章提出的 MoVieUp 自动剪辑系统是否提高了移动多摄像头的观看体验？比较的基准包括两个现有系统：Virtual Director^[91] 和 Jiku Director^[94]。评价主要从视频的质量、多样性、稳定性和整体的观看体验四个方面进行。

6.4.1 数据集

我们从 Youtube 网站¹上收集了 6 个事件一共 46 个录像，也是目前关于移动多摄像头视频自动剪辑最大的数据集。每个录像包含音频流和视频流。所有录像都是由不专业的用户通过移动设备拍摄的。上述章节中提到的视频质量问题在数据集中普遍存在。前三个事件的 14 个录像与 Virtual Director 使用的数据相同，我们用 Virtual Director 的作者提供的剪辑视频²作为比较对象。剩下的 32 个录像被提交到 Jiku Director 中得到相应的剪辑视频。表格 6.1 显示了数据集的详细情况。我们在一台 8CPU 16GB 内存的 Windows 服务器上运行 MoVieUp 系统，平

¹<http://www.youtube.com>

²<http://www.youtube.com/AutomaticMashup>

均需要 45 秒左右的时间评估一分钟音频的质量。对于一分钟的视频，需要 240 秒左右的时间做视频结构化分析，65 秒的时间做运动分析，13 秒的时间做质量评估。系统优化占用的时间如表 6.1 所示。

表 6.1 移动多摄像头视频数据集及算法优化时间

事件	录像数目	时长	同步	音频剪辑	切换点检测	视频剪辑
E1	5	4'37"	1'44"	0.11"	1.58"	26.19"
E2	5	7'01"	2'38"	0.14"	2.26"	36.26"
E3	4	5'15"	27"	0.11"	1.90"	10.26"
E4	8	6'25"	2'55"	0.10"	2.11"	2'49"
E5	11	3'32"	6'07"	0.11"	1.36"	6'26"
E6	27	5'22"	33'27"	0.19"	1.90"	19'24"

6.4.2 实验设置

音频剪辑评估采用用户调研的主观评估方法，评估遵循 MOS 评分策略：1 代表很差，2 代表差，3 代表一般，4 代表好，5 代表很好。在用户调研中，我们给用户播放本章算法和 Virtual Director 产生的一对音频，并要求用户给出评分。生成每个音频的系统对于用户是透明的，每组音频的内部顺序对不同用户是随机的。

由于普通用户对于视频镜头切换点的敏感性比较低，很难对切换点的质量给出客观公正的评价。为了更好地评价视频切换点，我们邀请了两位专业的视频编辑人员（与用户调研中的参与者不同）评价视频切换点的合适度。同时我们也邀请两位编辑对切换点给出他们的意见。该评价围绕两个方面开展：

- 切换频率是否合适？
- 切换点是否出现在合适的时机？

我们随机打乱了 Virtual Director、Jiku Director 和 MoVieUp 系统生成的一共 12 个剪辑视频的顺序。两位专业人员需要依次观看这些视频并针对上面两个问题给出他们的评分，分数范围和含义与音频剪辑评估相同。

对于视频镜头选取，我们组织了在线用户调研比较不同系统产生的视频观看体验。在调研开始前，我们提供了介绍页面帮助用户了解移动多摄像头视频自动剪辑以及用户即将要完成的评价内容，包括多样性、视频质量、稳定性和总体评价四个方面，每个方面的评价围绕以下问题展开：

- **多样性：**视频是否给出了事件丰富的概览？该问题调查剪辑的视频内容是否有单一枯燥的问题。



图 6.7 视频剪辑评价页面

- **视觉质量：**视频的视觉质量是否良好？此处的视觉质量主要是影响质量的空间因素。
- **稳定性：**视频是否稳定？由于抖动是移动视频的主要质量问题之一，我们单独评价视频的稳定性。
- **总体评价：**该视频剪辑是否专业？

我们要求用户针对每个视频回答上述四个问题并给出从 1（完全不同意）到 7（完全同意）的评分。

我们邀请了一位用户体验设计师设计了评价页面。不同于已有系统在实验中依次播放每个视频，设计师建议我们同时播放两个对比视频，两个视频按图 6.7 所示分左右两边分别摆放。两个视频的顺序是随机的。用户在不知道视频顺序以及对应剪辑系统的情况下给两个视频分别打分。这种评价方式的优点包括：

- 用户可以直观地感受到两个视频之间的差异。在已有系统的实验中，用户需要记住对比视频的所有内容和细节。
- 同时观看两个视频不会让用户感到迷惑，如果同时播放更多的视频会分散用户的注意力。

根据以上设定，本章提出的 MoVieUp 系统在前三个事件上与 Virtual Director 比较，在后三个事件上与 Jiku Director 比较。

6.4.3 音频剪辑评价

在音频剪辑评价中，我们进行了两组不同设置的实验。第一组实验评价本章提出的音频拼接方式是否有效。由于 MoVieUp 系统遵循最少切换准则，相应的音频剪辑的切换次数较少（如表格 6.2 所示）。因此，实验从三个剪辑音频中选取了 30 秒切换最频繁的片段作为评价对象，并与 Virtual Director 产生的对应时间段的音频作比较。

实验邀请了 18 名用户（包括 14 名男性和 4 名女性）参与用户调研。用户年龄在 22 岁到 26 岁之间。其中，观看视频录像（演唱会、比赛等）的频率分布

表 6.2 音频剪辑切换次数比较

系统	设定 1			设定 2		
	音频 1	音频 2	音频 3	音频 4	音频 5	音频 6
MoVieUp	1	2	1	2	1	4
VD	5	5	4	86	23	76

为：几乎不看-2人，每月-3人，每周-6人，每天-7人。评价结果如表格 6.3 所示。本章提出的音频拼接方法可以有效减少音频切换带来的听觉影响，尤其在音频的音量和音色差异较大时（音频 2）时改善更加明显。

表 6.3 音频剪辑主观评价 (MOS) 比较

系统	设定 1			设定 2		
	音频 1	音频 2	音频 3	音频 4	音频 5	音频 6
MoVieUp	2.56	2.28	3.56	3.00	3.81	3.0
VD	2.33	1.28	3.5	1.75	2.81	2.38

第二组实验评价另外三个完整的剪辑音频。该设定的目的是检验本章提出的音频剪辑方法是否能够提高音频剪辑的质量。我们邀请了另外 16 名用户参与用户调研，包括 5 名女性和 11 名男性，年龄分布从 20 岁到 33 岁。观看视频录像的频率分布为：几乎不看-2人，每月-1人，每周-10人，每天-3人。根据表格 6.3 中设定 2 的结果可以发现，MoVieUp 系统的音频质量明显高于 Virtual Director 系统。

根据以上两个设定的实验结果，本章提出的系统从三个方面生成了更好的剪辑音频：1) 提出了更好的音频选取算法；2) 显著减少了音频切换的次数；3) 系统更平滑地拼接各个音频片段，减少了音频切换引起的突兀的听觉效果。

6.4.4 切换点检测评估

本节比较本章提出的切换点检测算法相比于 Virtual Director 和 Jiku Director 的效果，其中 Virtual Director 根据所有录像中质量最好的音频人工选取切换点，Jiku Director 和 MoVieUp 系统根据算法自动选取切换点。

图 6.8 显示了关于切换频率的实验比较结果。MoVieUp 系统的切换点频率明显优于 Virtual Director 系统。考虑到 Virtual Director 系统通过人工选取的方法获得切换点，我们与编辑人员讨论发现 Virtual Director 选取了很多无意义的切换点。这些无意义的切换点是由于视频中出现了负面效果（抖动、遮挡等）引起的。虽然我们的系统同样根据音频选取视频切换点，但它在视频质量方面的表现更

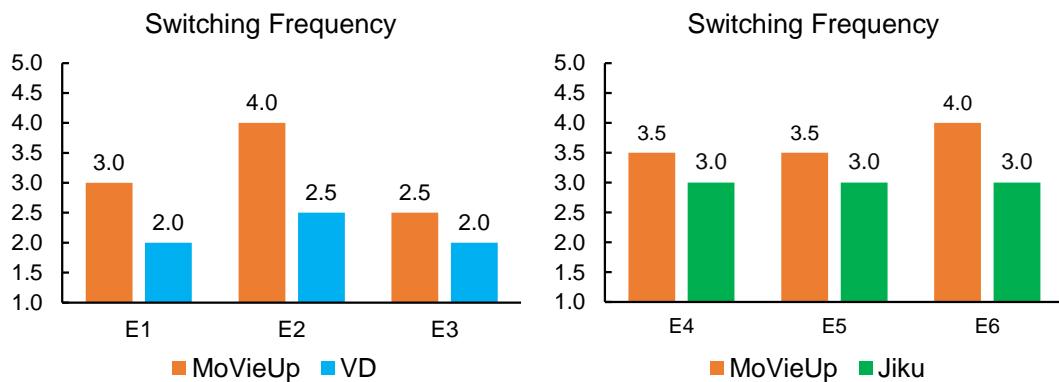


图 6.8 视频切换频率比较

好，避免了无意义的切换。编辑人员提醒我们当没有合适的视频可供选取时，不切换通常是更好的选择。

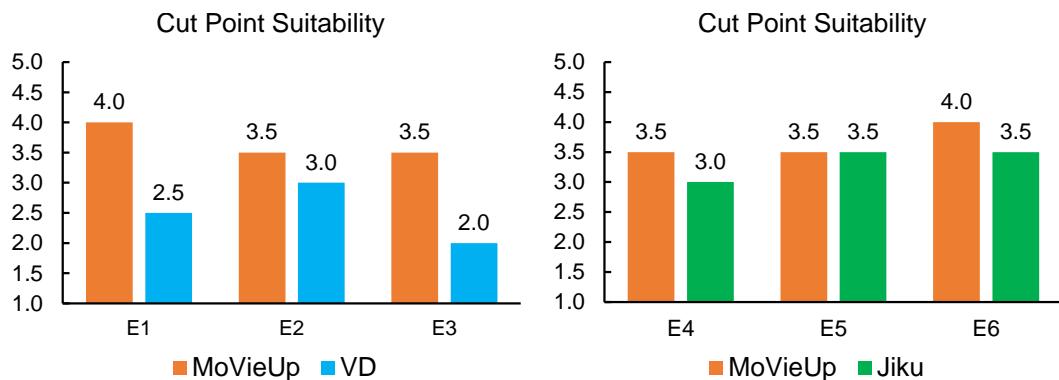


图 6.9 视频切换点位置比较

视频切换点位置合适度的比较结果如图 6.9 所示。同样由于无意义的切换点，Virtual Director 获得了很差的评分。该结果验证了切换点检测不仅与音频有关，也依赖于视频信息。根据评分，MoVieUp 获得了比 Jiku Director 略好的结果。在用户调研中，编辑人员推荐在说话或歌唱等语义行为的间隔切换，通过调研发现，部分文献建议在音乐的节奏点切换，编辑人员也很难给出明确的切换点选取规则。即使如此，试验结果仍然表明 MoVieUp 提供了一个可行的切换点检测方案，检测结果达到了一般甚至好的评价级别。

6.4.5 视频剪辑评估

本节从视频剪辑的角度评价 MoVieUp 系统相对于 Virtual Director 和 Jiku Director 系统的观看体验。为了公平比较三个系统的表现，实验中 MoVieUp 系统没有采用视频稳定等后处理措施。

与 Virtual Director 的比较：我们邀请参与音频剪辑第一组实验的 18 名用户

参加 MoViUp 系统与 Virtual Director 系统的比较实验，比较结果如图 6.10 所示。

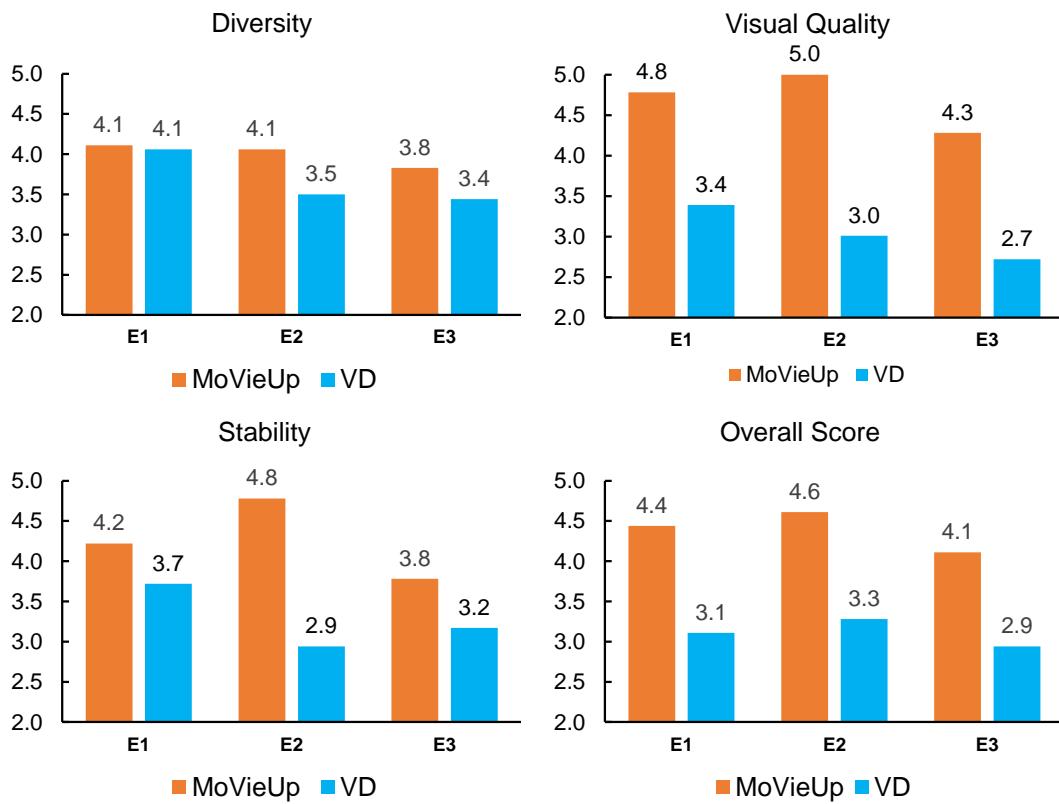


图 6.10 MoViUp 系统与 Virtual Director 系统视频剪辑结果比较

通过比较发现，MoViUp 系统在多样性方面比 Virtual Director 系统表现更好，但相对于其他三个评价因素，多样性的优势并不明显。该结果也在合理的范围之内，因为两个系统都在切换点选取不同的视频源避免单调性问题，区别在于 MoViUp 系统考虑了时间因素，而 Virtual Director 系统仅依靠相邻两帧衡量多样性。此外，我们分析了三个事件对应的视频，发现每个事件仅包含 4 到 5 个从不同角度和距离拍摄的视频，视频源本身也能够避免一部分的单调性问题。

在视觉质量上，MoViUp 系统的表现更好。Virtual Director 系统考虑了四个方面的质量因素：块效应（blockiness）、模糊（blurriness）、亮度（brightness）和抖动（shakiness）。MoViUp 系统考虑了更多的空间因素和时间因素，如倾斜（tilting）、失真（infidelity）和急动（jerkiness）。质量预处理操作也滤除了质量很差的镜头。为了进一步研究系统表现好的原因，我们邀请了三位专业的编辑人员从抖动、过暗和失真（包括由于遮挡和强光等原因引起的失真）三个角度将剪辑视频的镜头标注成“好”或“差”。倾斜在本组视频中出现较少，模糊往往伴随抖动出现，因而这两个因素没有单独标注。标注的结果如表格 6.4 所示。在第一个事件中，Virtual Director 系统选取的 46 个镜头中有 24 个镜头存在过暗的问题（超过一半的画面是黑的）。与之相应，MoViUp 系统选取的 27 个镜头中只有 7 个镜头存在过暗问题。在第二个事件中，Virtual Director 系统选取的 55 个镜头中有

表 6.4 MoVieUp 和 Virtual Director 系统存在质量问题的镜头数量比较

因素	MoVieUp			Virtual Director		
	事件 1	事件 2	事件 3	事件 1	事件 2	事件 3
镜头	27	27	29	46	55	43
抖动	0	2	1	2	14	8
过暗	7	0	0	24	1	0
失真	0	2	5	0	5	8

14 个镜头受不规则相机运动的影响，导致了抖动问题，同样的问题在 MoVieUp 系统中只存在于两个镜头中。我们对该事件的视频做了进一步的分析发现，5 个视频中有 4 个视频存在抖动的问题。MoVieUp 系统选取的 2 个抖动镜头很可能是出于多样性的考虑。此外，失真也是影响视觉质量的关键因素。Virtual Director 系统选取了 5 个被强光污染的镜头，而 MoVieUp 系统中只有 2 个镜头有同样的问题。在第三个事件中，Virtual Director 系统选取的 43 个镜头中有 8 个存在抖动问题，MoVieUp 系统只选取了 1 个抖动镜头。根据实验结果和用户反馈，可以发现本章提出的系统可以从空间和时间两个方面选取出高质量的视频镜头。

本章提出的 MoVieUp 系统在三个事件上从三个具体的方面都取得了比 Virtual Director 系统更好的效果，验证了系统采用的多样性和视觉质量评估方法的有效性，同时也表明了系统采用的算法可以获得在多样性和视觉质量上更优的结果和观看体验，也因此 MoVieUp 系统在整体评价上获得了更高的评分。

与 Jiku Director 的比较：我们邀请了参与音频剪辑评估第二组实验的 15 名用户参加与 Jiku Director 系统的比较实验。实验结果如图 6.11 所示。

表 6.5 MoVieUp 和 Jiku Director 系统存在质量问题的镜头数量比较

因素	MoVieUp			Jiku Director		
	事件 4	事件 5	事件 6	事件 4	事件 5	事件 6
镜头	49	21	37	49	26	49
抖动	3	0	0	11	3	3
过暗	1	0	0	0	0	1
失真	5	1	1	7	3	4

多样性方面，MoVieUp 系统获得了更好的效果，类似于与 Virtual Director 系统的比较结果，多样性方面的提升并不明显。两个系统都采用了基于关键帧的相似度方法，都采用了用户对内容的兴趣随着镜头时长递减的多样性模型。多样性方面的区别在于 Jiku Director 首先确定镜头的视角（拍摄距离和角度），再从对

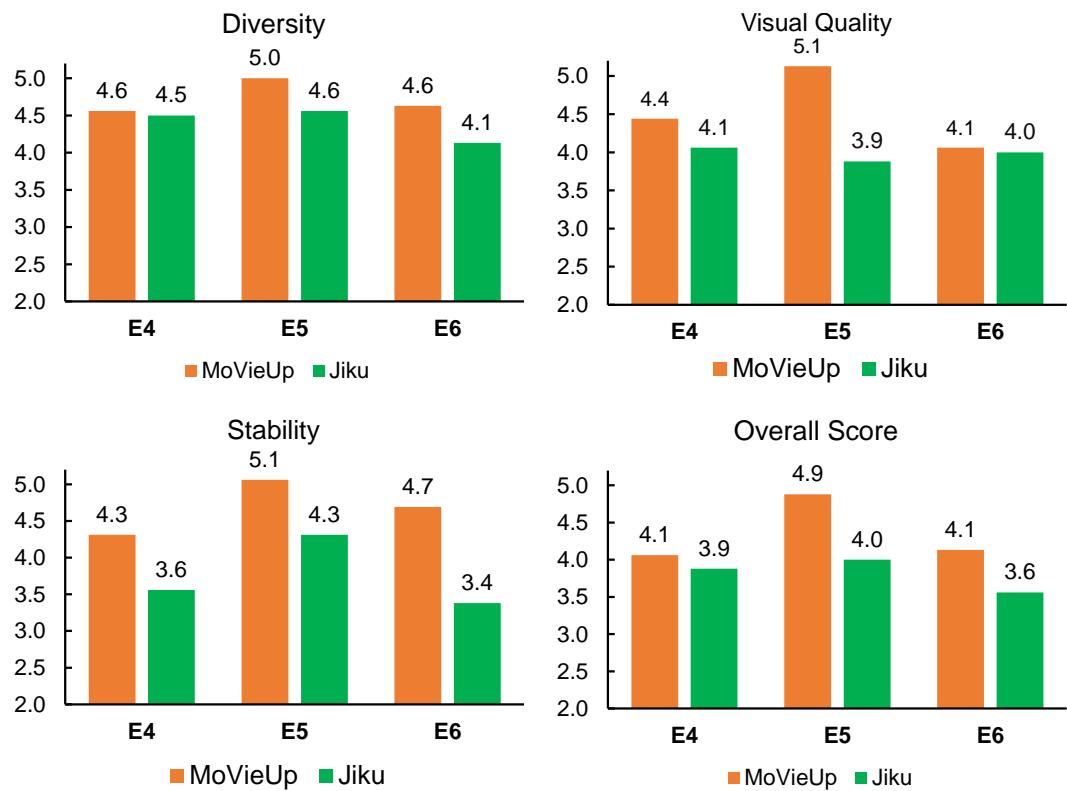


图 6.11 MoVieUp 系统与 Jiku Director 系统视频剪辑结果比较

应视角选取视频镜头。该策略可能在连续选取的视角比较接近时的多样性较差，与之对比，MoVieUp 系统基于记忆模型选取镜头。

视觉质量方面，两个系统在事件 4 和事件 6 上的表现比较接近。系统均考虑了空间和时间上的质量因素。我们采用了与 Virtual Director 系统相同的客观比较方法对有问题的镜头进行标注，结果如表格 6.5 所示。MoVieUp 系统在事件 4 上选取了更少的抖动镜头，但选取了相对较多的失真镜头，这可能是导致视觉质量评分相差不大的主要因素。在事件 5 中，MoVieUp 系统表现更好。我们发现 Jiku Director 系统产生的视频受遮挡、强光和不规律的相机运动影响。对于事件 6，编辑人员反应视频存在模糊问题。根据图 6.5 和表格 6.5 的结果，我们可以总结出 MoVieUp 系统在视觉质量上比 Jiku Director 系统表现更好。

总结起来，本章提出的系统相对于已有的剪辑系统能达到更好的多样性、视觉质量和稳定性。整体的评分表明该系统提供了更好的移动多摄像头视频观看体验。

6.5 本章小结

本章内容介绍了一个全自动移动多摄像头视频自动剪辑系统。该系统从多摄像头视频中剪辑音频和视频，相比于已有的剪辑系统达到了更好的观看体验。系统基于从用户调研中得到的可计算的视频编辑语法。为了生成高质量的剪辑

音频，系统对输入音频进行质量评估，并在最少切换准则下剪辑音频。在剪辑后的音频上，系统衡量节奏合适度和语义合适度，检测视频镜头的切换点。对于视频镜头选取，系统考虑了相机运动一致性使得镜头平滑切换。为了保证视频质量，系统考虑了空间和时间上的质量因素。为了增强内容的多样性，系统基于记忆模型解决视频内容的单调性问题。视频剪辑最终通过优化问题求解。后处理操作进一步完善了剪辑视频的语义完整性，增加了视频的稳定性。系统最后对剪辑音频和剪辑视频混流得到最终的剪辑结果。

本章的主要贡献包括：

- 提出了全自动移动多摄像头视频自动剪辑系统，与已有系统需要一部分人工干预有显著区别。
- 系统进行了视频剪辑调研，引入了一系列可计算的视频剪辑语法，这些可计算的语法提供了音视频剪辑的指导规则。
- 首次考虑了音频剪辑，音频剪辑的作用被现有系统忽视，然而它对于提升用户体验十分重要。

第7章 总结与展望

本论文的主要目标是研究并实现社交多媒体数据内容分析和处理系统，完成对社交多媒体数据的语义理解和关联表达。语义理解的内容包括：(1)解决语义理解标注难的问题：从大规模标注不准确的社交多媒体数据中学习语义理解模型；(2)解决大规模社交多媒体数据处理慢的问题：从大量数据特征中选取与目标任务相关的特征子集，加快语义理解相关问题的处理速度；简化语义理解模型，减少深度卷积神经网络的模型参数，加快语义理解的速度。社交多媒体数据关联表达是指根据用户个性化的需求，从社交多媒体数据中选择有关联的数据，并以一定的表达形式将这些关联的数据呈现给用户。本论文分别从照片和视频的角度研究关联表达的具体应用：(1)解决社交多媒体数据中基于主题的照片集故事化表达问题：选取语义上有关联的代表性图片，通过可计算的视频编辑语法对照片集进行故事化表达；(2)解决社交多媒体数据中移动多摄像头视频自动剪辑问题：将移动多摄像头视频在时间上同步，通过可计算的视频剪辑语法，选取镜头和录音，将移动多摄像头视频剪辑成单一的音视频流。

本章首先对本论文的工作进行总结，再给出今后研究方向的展望。

7.1 本文总结

本论文的主要研究内容是社交多媒体数据的语义理解和关联表达，主要解决社交多媒体数据标注难、处理慢以及基于主题的照片集故事化表达和移动多摄像头视频自动剪辑的关联表达问题。本文的主要创新点包括：

(1) 弱监督社交多媒体数据语义理解

社交多媒体数据语义理解是处理社交多媒体数据的基础。传统的语义理解方法通常依赖准确的数据标注信息，而社交多媒体数据的标注包含大量的噪音，影响了已有方法的准确性。现有的噪音鲁棒算法依赖特定的噪音假设模型或相应的训练方法过于复杂。针对这一现状，本文提出了一种新的噪音鲁棒的弱监督相关反馈深度学习算法，解决有噪音情况下任意语义类别的学习问题。该方法基于感知连续性的假设，即语义上接近的数据在特征空间上也比较接近，利用特征之间的相关性使得不同数据在训练过程中有不同的梯度贡献。算法将数据特征转化为相似性表示，并利用相似矩阵的低秩近似实现感知连续性。为了减小模型训练的复杂度，本文进一步对目标函数进行了简化和近似，提出了高效的弱监督相关反馈深度学习算法。与传统的语义理解方法以及现有的弱监督语义理解方

法相比，本文提出的相关反馈网络具有更好的噪音鲁棒性。

(2) 大规模社交多媒体数据快速处理

社交多媒体数据的快速处理是将语义理解应用到实际问题的关键。社交多媒体数据具有数据规模大、特征种类多、应用场景复杂多样的特点。不同应用场景需要不同的特征表示，冗余特征的存在会影响模型的效率和效果。此外，当前移动端的处理逐渐成为趋势，移动设备的计算能力、存储空间以及电池容量都十分有限，提高社交多媒体数据的处理速度十分必要。本文首先从特征选取的角度实现社交多媒体数据的快速处理。现有的批处理特征选取算法存在计算慢、内存占用高、不能处理流数据的缺点，已有的在线特征选取算法的效果与批处理算法有明显的差距。本文利用二阶在线学习算法，基于特征的置信度选取特征，并利用最大/最小堆结构提出快速在线特征选取算法，将二阶在线特征选取算法的复杂度降低成与非零特征数目成正比。在不同规模的合成数据集和公开数据集上的比较结果验证了本文提出算法的有效性和高效性。

此外，深度卷积神经网络广泛应用于社交多媒体数据的语义理解。深度网络的参数个数较多，需要大量的计算资源和时间。现有的深度卷积神经网络简化算法需要依赖特定的硬件或软件库的支持，或存在优化困难的问题。本文利用二阶在线特征选取算法提出了模型简化算法，在不影响网络表达能力的情况下减少了模型参数，提高了模型速度。

(3) 社交多媒体数据的关联表达

社交多媒体数据的关联表达是语义理解的最终目的。当前，社交多媒体数据的产生、分享和获取都十分便利，然而这些数据都以碎片化的形式存在于社交多媒体上。本文从照片和视频两个角度分别对社交多媒体数据关联表达的具体应用进行了研究。

对于照片集，本文提出了基于主题的照片集故事化表达系统——Monet。系统首先对照片集进行分析与梳理，根据时间和位置信息将照片划分到不同的事件，再根据照片的视觉质量、代表性和均衡性选取一部分子集作为关键性照片。其次，系统根据照片的内容选取合适的基于主题的编辑风格。根据主题风格的设计模板，系统选取合适的相机运动将关键照片转换为视频片段，并将一系列的视频特效、颜色过滤器、形状和转场等视觉效果应用到视频片段中。最终的视频和音乐经过混流生成故事化的音乐视频表达。实验结果表明，本文提出的 Monet 系统达到了最好的照片集分析与梳理和故事合成效果。

对于视频，本文针对移动多摄像头视频提出了自动剪辑系统——MoVieUp，将同一事件中，多个摄像头从多个角度拍摄的时间上有重叠的一组视频剪辑成内容丰富、能体现专业编辑水平的单一音视频流。我们通过用户调研总结了一系列可计算的视频剪辑规则。基于这些规则，系统通过音频指纹同步多摄像头视

频、评价音视频质量、在最小切换准则下最大化音频质量生成音视剪辑结果。对于视频剪辑，系统根据音频的节奏和语义信息检测视频切换点。在切换点上基础上，系统在镜头运动一致性的约束下最大化视频质量和多样性，选取视频镜头。实验结果显示 MoViUp 系统达到了最好的移动多摄像头视频自动剪辑的效果，提供了更好的用户体验。

7.2 研究工作展望

社交多媒体数据的语义理解和关联表达是一个充满挑战的研究课题。随着移动设备进一步的普及和性能的提升，社交多媒体数据的规模和内容都将继续快速增长，相应的研究内容和应用场景也将更加复杂多样，并受到学术界和工业界越来越多的共同关注。随着计算机视觉、机器学习和多媒体计算相关研究领域的发展，可以从以下几个方面对本文的工作进行改进和扩展：

(1) 语义理解方面。社交多媒体数据的语义理解既包含图片数据上的目标识别、目标检测、物体分割、内容描述等研究内容，也包括视频数据上的场景识别、目标检测与跟踪、视频内容描述等。本文主要解决照片数据的弱监督目标识别问题，利用大规模社交多媒体数据开展更多的弱监督语义理解研究具有很大的研究意义和应用价值。例如，Bilen 等人通过修改在标准数据集上预训练的深度卷积神经网络同时学习区域选取和目标识别模型，实现了弱监督目标检测^[147]。Pinheiro 等人仅利用物体是否在照片中出现的弱监督信息在深度卷积神经网络上学习目标分割模型^[148]。Rochan 等人研究了在弱标注视频中的目标定位和分割问题^[149]，训练数据仅需要提供视频中出现的主要物体，算法就可以自动在每一帧中定位出物体的位置并将物体与背景分开。近年来，从视频生成文字描述获得了广泛的关注和研究，Shen 等人提出了不需要准确语句标注的弱监督视频内容描述算法^[150]。这些弱监督语义理解方法为社交多媒体数据提供了更全面的分析和有益的借鉴。

(2) 快速处理方面。虽然本文提出了快速有效的二阶在线特征选取算法用于社交多媒体数据的快速处理，但当选取的特征数目较少时，算法的效果与批处理算法仍然有一定差距。如何从数据中挖掘更多的信息，提高当前在线特征选取算法在特征数目较少时的效果仍然需要进一步的研究。当前的特征选取算法需要手动指定特征的数目。在实际应用中，算法应该能够不断接收新的数据并自动输出最紧凑最准确的模型，如何自适应地选取特征数目也是需要解决的主要问题之一。本文提出的特征选取算法主要解决单个特征的特征选取问题。当前，研究人员探索了基于结构化信息的特征在文本问题上的应用，如文本分类^[151] 和文本聚类^[152]。如何快速有效地选取结构化特征并将它们应用到社交多媒体数据上是极

富挑战性的问题。对于深度模型简化，本文提出的简化算法同样存在不能自适应地决定卷积核数目的问题。此外，对于网络深度的简化仍需要进一步地探索。

(3) 关联表达方面。本文提出的基于主题的照片集故事化表达系统主要侧重对照片集的内容进行故事化表达，对于照片集的社交性没有做针对性的深入处理。从社交网络中挖掘不同用户具有社交联系的照片集并针对不同用户生成个性化的故事化表达，进一步加强社交照片数据的关联表达，是今后的一个主要研究问题。

对于视频的关联表达，当前提出的移动多摄像头视频自动剪辑系统基于内容上的相似性进行多样性选取，为了利用更多可计算的视频编辑语法，需要确定移动摄像头拍摄的角度和距离^[153]，对视频做更深入的语义分析，提高剪辑结果的多样性和专业性。此外，当前系统在选取视频切换点时主要考虑音频的节奏和简单的语义信息，引入视频的语义内容和更丰富的音频语义从而提高切换点检测的效果仍然有待进一步的研究。

参考文献

- [1] Earth photographed over eight hundred billion mobile phones to take nearly 80%, Online; accessed 2-January-2017[M]. [S.l.]: [s.n.].
- [2] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision. 2015, 115 (3): 211–252.
- [3] Lowe D G. Object recognition from local scale-invariant features[C]//Ieee, Computer vision, 1999. The proceedings of the seventh IEEE international conference on: volume 2. [S.l.]: Ieee, 1999: 1150–1157.
- [4] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]//Springer, European conference on computer vision. [S.l.]: Springer, 2006: 404–417.
- [5] Danyluk A, Provost F. Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network[C]//Proc. of Tenth International Conference on Machine Learning. [S.l.], 2014: 81–88.
- [6] Brodley C E, Friedl M A. Identifying mislabeled training data[J]. Journal of Artificial Intelligence Research. 1999, 11: 131–167.
- [7] Zhou B, Jagadeesh V, Piramuthu R. ConceptLearner: Discovering visual concepts from weakly labeled image collections[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 1492–1500.
- [8] Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data[C]//Advances in neural information processing systems. [S.l.], 2006: 601–608.
- [9] Vo P D, Ginsca A, Borgne H L, et al. On deep representation learning from noisy web images[J]. arXiv preprint arXiv:1512.04785. 2015.
- [10] Schroff F, Criminisi A, Zisserman A. Harvesting image databases from the web[J]. IEEE transactions on pattern analysis and machine intelligence. 2011, 33 (4): 754–766.
- [11] Chatzilari E, Nikolopoulos S, Kompatsiaris Y, et al. Salic: Social active learning for image classification[M]//[S.l.]: IEEE.
- [12] Manwani N, Sastry P. Noise tolerance under risk minimization[J]. IEEE transactions on cybernetics. 2013, 43 (3): 1146–1151.
- [13] Thathachar M A, Sastry P S. Networks of learning automata: Techniques for online stochastic optimization[M]. [S.l.]: Springer Science & Business Media, 2011.
- [14] Sastry P, Nagendra G, Manwani N. A team of continuous-action learning automata for noise-

- tolerant learning of half-spaces[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2010, 40 (1): 19–28.
- [15] Beigman E, Klebanov B B. Learning with annotation noise[C]//Association for Computational Linguistics, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. [S.l.]: Association for Computational Linguistics, 2009: 280–287.
- [16] Bunescu R C, Mooney R J. Multiple instance learning for sparse positive bags[C]//ACM, Proceedings of the 24th international conference on Machine learning. [S.l.]: ACM, 2007: 105–112.
- [17] Vijayanarasimhan S, Grauman K. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization[C]//IEEE, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. [S.l.]: IEEE, 2008: 1–8.
- [18] Sindhwani V, Keerthi S S. Large scale semi-supervised linear svms[C]//ACM, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. [S.l.]: ACM, 2006: 477–484.
- [19] Feng J, Xu H, Mannor S, et al. Robust logistic regression and classification[C]//Advances in Neural Information Processing Systems. [S.l.], 2014: 253–261.
- [20] Izadinia H, Farhadi A, Hertzmann A, et al. Image classification and retrieval from user-supplied tags[J]. arXiv preprint arXiv:1411.6909. 2014.
- [21] Izadinia H, Russell B C, Farhadi A, et al. Deep classifiers from image tags in the wild[C]//ACM, Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions. [S.l.]: ACM, 2015: 13–18.
- [22] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE. 1998, 86 (11): 2278–2324.
- [23] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. [S.l.], 2012: 1097–1105.
- [24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556. 2014.
- [25] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 1–9.
- [26] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2016: 2818–2826.
- [27] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of

- residual connections on learning[J]. arXiv preprint arXiv:1602.07261. 2016.
- [28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2016: 770–778.
- [29] Reed S, Lee H, Anguelov D, et al. Training deep neural networks on noisy labels with bootstrapping[J]. arXiv preprint arXiv:1412.6596. 2014.
- [30] Sukhbaatar S, Bruna J, Paluri M, et al. Training convolutional networks with noisy labels[J]. arXiv preprint arXiv:1406.2080. 2014.
- [31] Xiao T, Xia T, Yang Y, et al. Learning from massive noisy labeled data for image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 2691–2699.
- [32] Azadi S, Feng J, Jegelka S, et al. Auxiliary image regularization for deep cnns with noisy labels[J]. arXiv preprint arXiv:1511.07069. 2015.
- [33] Jain A K, Vailaya A. Image retrieval using color and shape[J]. Pattern recognition. 1996, 29 (8): 1233–1244.
- [34] Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data[J]. IEEE Transactions on pattern analysis and machine intelligence. 1996, 18 (8): 837–842.
- [35] Yang J, Jiang Y G, Hauptmann A G, et al. Evaluating bag-of-visual-words representations in scene classification[C]//ACM, Proceedings of the international workshop on Workshop on multimedia information retrieval. [S.l.]: ACM, 2007: 197–206.
- [36] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]//ICML: volume 3. [S.l.], 2003: 856–863.
- [37] Jiang F, Sui Y, Zhou L. A relative decision entropy-based feature selection approach[J]. Pattern Recognition. 2015, 48 (7): 2151–2163.
- [38] Li F, Zhang Z, Jin C. Feature selection with partition differentiation entropy for large-scale data sets[J]. Information Sciences. 2016, 329: 690–700.
- [39] Yang H, Lyu M R, King I. Efficient online learning for multitask feature selection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD). 2013, 7 (2): 6.
- [40] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial intelligence. 1997, 97 (1-2): 273–324.
- [41] Pappu V, Panagopoulos O P, Xanthopoulos P, et al. Sparse proximal support vector machines for feature selection in high dimensional datasets[J]. Expert Systems with Applications. 2015, 42 (23): 9183–9191.
- [42] Le Thi H A, Vo X T, Dinh T P. Feature selection for linear svms under uncertain data: Robust optimization based on difference of convex functions algorithms[J]. Neural Networks. 2014, 59: 36–50.

- [43] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain.[J]. Psychological review. 1958, 65 (6): 386.
- [44] Crammer K, Dekel O, Keshet J, et al. Online passive-aggressive algorithms[J]. Journal of Machine Learning Research. 2006, 7 (Mar): 551–585.
- [45] Crammer K, Dredze M, Kulesza A. Multi-class confidence weighted algorithms[C]//Association for Computational Linguistics, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. [S.l.]: Association for Computational Linguistics, 2009: 496–504.
- [46] Crammer K, Kulesza A, Dredze M. Adaptive regularization of weight vectors[C]//Advances in neural information processing systems. [S.l.], 2009: 414–422.
- [47] Langford J, Li L, Zhang T. Sparse online learning via truncated gradient[J]. Journal of Machine Learning Research. 2009, 10 (Mar): 777–801.
- [48] Duchi J, Singer Y. Efficient online and batch learning using forward backward splitting[J]. Journal of Machine Learning Research. 2009, 10 (Dec): 2899–2934.
- [49] Xiao L. Dual averaging methods for regularized stochastic learning and online optimization[J]. Journal of Machine Learning Research. 2010, 11 (Oct): 2543–2596.
- [50] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research. 2011, 12: 2121–2159.
- [51] Wu X, Yu K, Wang H, et al. Online streaming feature selection[C]//Proceedings of the 27th international conference on machine learning (ICML-10). [S.l.], 2010: 1159–1166.
- [52] Huang H, Yoo S, Kasiviswanathan S P. Unsupervised feature selection on data streams[C]//ACM, Proceedings of the 24th ACM International Conference on Information and Knowledge Management. [S.l.]: ACM, 2015: 1031–1040.
- [53] Wang J, Zhao P, Hoi S C, et al. Online feature selection and its applications[J]. IEEE Transactions on Knowledge and Data Engineering. 2014, 26 (3): 698–710.
- [54] Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146. 2016.
- [55] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387. 2015.
- [56] Denil M, Shakibi B, Dinh L, et al. Predicting parameters in deep learning[C]//Advances in Neural Information Processing Systems. [S.l.], 2013: 2148–2156.
- [57] Denton E L, Zaremba W, Bruna J, et al. Exploiting linear structure within convolutional networks for efficient evaluation[C]//Advances in Neural Information Processing Systems. [S.l.], 2014: 1269–1277.
- [58] Rigamonti R, Sironi A, Lepetit V, et al. Learning separable filters[C]//Proceedings of the

- IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2013: 2754–2761.
- [59] Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions[J]. arXiv preprint arXiv:1405.3866. 2014.
- [60] Ioannou Y, Robertson D, Shotton J, et al. Training cnns with low-rank filters for efficient image classification[J]. arXiv preprint arXiv:1511.06744. 2015.
- [61] Tai C, Xiao T, Zhang Y, et al. Convolutional neural networks with low-rank regularization[J]. arXiv preprint arXiv:1511.06067. 2015.
- [62] Mamalet F, Garcia C. Simplifying convnets for fast learning[C]//Springer, International Conference on Artificial Neural Networks. [S.l.]: Springer, 2012: 58–65.
- [63] Hwang K, Sung W. Fixed-point feedforward deep neural network design using weights+ 1, 0, and- 1[C]//IEEE, Signal Processing Systems (SiPS), 2014 IEEE Workshop on. [S.l.]: IEEE, 2014: 1–6.
- [64] Arora S, Bhaskara A, Ge R, et al. Provable bounds for learning some deep representations.[C]//ICML. [S.l.], 2014: 584–592.
- [65] Courbariaux M, Bengio Y, David J P. Binaryconnect: Training deep neural networks with binary weights during propagations[C]//Advances in Neural Information Processing Systems. [S.l.], 2015: 3123–3131.
- [66] Rastegari M, Ordonez V, Redmon J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks[C]//Springer, European Conference on Computer Vision. [S.l.]: Springer, 2016: 525–542.
- [67] Gong Y, Liu L, Yang M, et al. Compressing deep convolutional networks using vector quantization[J]. arXiv preprint arXiv:1412.6115. 2014.
- [68] Wu J, Leng C, Wang Y, et al. Quantized convolutional neural networks for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2016: 4820–4828.
- [69] Anwar S, Hwang K, Sung W. Fixed point optimization of deep convolutional neural networks for object recognition[C]//IEEE, Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. [S.l.]: IEEE, 2015: 1131–1135.
- [70] Chen W, Wilson J T, Tyree S, et al. Compressing neural networks with the hashing trick.[C]//ICML. [S.l.], 2015: 2285–2294.
- [71] Liu B, Wang M, Foroosh H, et al. Sparse convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2015: 806–814.
- [72] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Advances in Neural Information Processing Systems. [S.l.], 2015: 1135–1143.
- [73] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with

- pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149. 2015.
- [74] Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convnets[J]. arXiv preprint arXiv:1608.08710. 2016.
- [75] Murray K, Chiang D. Auto-sizing neural networks: With applications to n-gram language models[J]. arXiv preprint arXiv:1508.05051. 2015.
- [76] Anwar S, Hwang K, Sung W. Structured pruning of deep convolutional neural networks[J]. arXiv preprint arXiv:1512.08571. 2015.
- [77] Hu H, Peng R, Tai Y W, et al. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures[J]. arXiv preprint arXiv:1607.03250. 2016.
- [78] Changpinyo S, Sandler M, Zhmoginov A. The power of sparsity in convolutional neural networks[J]. arXiv preprint arXiv:1702.06257. 2017.
- [79] Platt J C, Czerwinski M, Field B A. Phototoc: Automatic clustering for browsing personal photographs[C]//IEEE, Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on: volume 1. [S.l.]: IEEE, 2003: 6–10.
- [80] Graham A, Garcia-Molina H, Paepcke A, et al. Time as essence for photo browsing through personal digital libraries[C]//ACM, Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. [S.l.]: ACM, 2002: 326–335.
- [81] Gargi U. Modeling and clustering of photo capture streams[C]//ACM, Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval. [S.l.]: ACM, 2003: 47–54.
- [82] Cooper M, Foote J, Girgensohn A, et al. Temporal event clustering for digital photo collections[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2005, 1 (3): 269–288.
- [83] Loui A C, Savakis A E. Automatic image event segmentation and quality screening for albuming applications[C]//IEEE, Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on: volume 2. [S.l.]: IEEE, 2000: 1125–1128.
- [84] Gong B, Jain R. Segmenting photo streams in events based on optical metadata[C]//IEEE, Semantic Computing, 2007. ICSC 2007. International Conference on. [S.l.]: IEEE, 2007: 71–78.
- [85] Mei T, Wang B, Hua X S, et al. Probabilistic multimodality fusion for event based home photo clustering[C]//IEEE, Multimedia and Expo, 2006 IEEE International Conference on. [S.l.]: IEEE, 2006: 1757–1760.
- [86] Shen X, Tian X. Multi-modal and multi-scale photo collection summarization[J]. Multimedia Tools and Applications. 2016, 75 (5): 2527–2541.

- [87] Chu W T, Lin C H. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection[C]//ACM, Proceedings of the 16th ACM international conference on Multimedia. [S.I.]: ACM, 2008: 829–832.
- [88] Hua X S, Lu L, Zhang H J. Photo2video—a system for automatically converting photographic series into video[J]. IEEE Transactions on circuits and systems for video technology. 2006, 16 (7): 803–819.
- [89] Chu W T, Chen J C, Wu J L. Tiling slideshow: an audiovisual presentation method for consumer photos[J]. IEEE MultiMedia. 2007, 14 (3).
- [90] Kuo T H, Tsai C Y, Cheng K Y, et al. Sewing photos: Smooth transition between photos[C]//Springer, International Conference on Multimedia Modeling. [S.I.]: Springer, 2011: 73–83.
- [91] Shrestha P, de With P H N, Weda H, et al. Automatic mashup generation from multiple-camera concert recordings[C]//ACM Multimedia. [S.I.], 2010: 541-550.
- [92] Russell S J, Norvig P. Artificial Intelligence — A Modern Approach[M]. [S.I.]: Pearson Education, 2010: I-XVIII, 1-1132.
- [93] Nguyen D T D, Saini M, Nguyen V T, et al. Jiku director: A mobile video mashup system[C]//ACM Multimedia. [S.I.], 2013: 477–478.
- [94] Saini M K, Gadde R, Yan S, et al. MoViMash: online mobile video mashup[C]//ACM Multimedia. [S.I.], 2012: 139-148.
- [95] Hua X S, Lu L, Zhang H. Automatic music video generation based on temporal pattern analysis[C]//ACM Multimedia. [S.I.], 2004: 472-475.
- [96] Arev I, Park H S, Sheikh Y, et al. Automatic editing of footage from multiple social cameras[J/OL]. ACM Trans. Graph. July 2014, 33 (4): 81:1–81:11. <http://doi.acm.org/10.1145/2601097.2601198>. DOI: 10.1145/2601097.2601198.
- [97] Sundaram H, Chang S F. Computable scenes and structures in films[J]. IEEE Transactions on Multimedia. 2002, 4 (4): 482–491.
- [98] Sharff S. The Elements of Cinema: Toward a Theory of Cinesthetic Impact[M]. [S.I.]: Columbia University Press, 1982.
- [99] Lampi F, Kopf S, Benz M, et al. A virtual camera team for lecture recording[J]. IEEE MultiMedia. 2008, 15 (3): 58-61.
- [100] Sumec S. Multi camera automatic video editing[M]. [S.I.]: [s.n.], 2006: 935–945.
- [101] Hua X S, Lu L, Zhang H J. Optimization-based automated home video editing system[J]. IEEE Transactions on Circuits and Systems for Video Technology. 2004, 14 (5): 572–583.
- [102] Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE transactions on neural networks and learning systems. 2014, 25 (5): 845–869.

-
- [103] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]//NIPS: volume 14. [S.l.], 2001: 585–591.
 - [104] Candès E J, Li X, Ma Y, et al. Robust principal component analysis?[J]. Journal of the ACM (JACM). 2011, 58 (3): 11.
 - [105] Eckart C, Young G. The approximation of one matrix by another of lower rank[J]. Psychometrika. 1936, 1 (3): 211–218.
 - [106] Golub G H, Van Loan C F. Matrix computations: volume 3[M]. [S.l.]: JHU Press, 2012.
 - [107] Yang Y, Shen H T, Ma Z, et al. L2, 1-norm regularized discriminative feature selection for unsupervised learning[C]//IJCAI proceedings-international joint conference on artificial intelligence: volume 22. [S.l.], 2011: 1589.
 - [108] Ye J, Zhao Z, Wu M. Discriminative k-means for clustering[C]//Advances in neural information processing systems. [S.l.], 2008: 1649–1656.
 - [109] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[M]//[S.l.]: Citeseer, 2009.
 - [110] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results[M]. [S.l.]: [s.n.], 2007.
 - [111] Luo P, Wang X, Tang X. Hierarchical face parsing via deep learning[C]//IEEE, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. [S.l.]: IEEE, 2012: 2480–2487.
 - [112] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.], 2014: 1717–1724.
 - [113] Divvala S K, Farhadi A, Guestrin C. Learning everything about anything: Webly-supervised visual concept learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2014: 3270–3277.
 - [114] Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model[C]//Advances in neural information processing systems. [S.l.], 2013: 2121–2129.
 - [115] Bychkovsky V, Paris S, Chan E, et al. Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs[C]//The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2011.
 - [116] Bolón-Canedo V, Sánchez-Marcano N, Alonso-Betanzos A. Recent Advances and Emerging Challenges of Feature Selection in the Context of Big Data[J]. Knowledge Based System. September 2015, 86 (C): 33–45.
 - [117] Zhai Y, Ong Y S, Tsang I. The Emerging "Big Dimensionality"[J]. Computational Intelligence Magazine, IEEE. Aug 2014, 9 (3): 14-26.

- [118] Dredze M, Crammer K, Pereira F. Confidence-weighted linear classification[C]//Proceedings of the 25th International Conference on Machine Learning. New York, NY, USA: ACM, 2008: 264–271.
- [119] Ma J, Kulesza A, Dredze M, et al. Exploiting Feature Covariance in High-Dimensional Online Learning[C]//Proceedings of the Artificial Intelligence and Statistics. [S.I.], 2010: 493–500.
- [120] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence. 2005, 27 (8): 1226–1238.
- [121] Ramírez-Gallego S, Lastra I, Martínez-Rego D, et al. Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data[J]. International Journal of Intelligent Systems. 2017, 32 (2): 134–152.
- [122] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research. 2008, 9: 1871–1874.
- [123] Tan M, Tsang I W, Wang L. Towards ultrahigh dimensional feature selection for big data[J]. Journal of Machine Learning Research. 2014, 15 (1): 1371-1429.
- [124] Zagoruyko S. 92.45% on CIFAR-10 in Torch, Online; accessed 2-January-2017[M]. [S.I.]: [s.n.], 2015.
- [125] Girgensohn A, Boreczky J, Chiu P, et al. A Semi-automatic Approach to Home Video Editing[C]//the 13th Annual ACM Symposium on User Interface Software and Technology. [S.I.], 2000: 81–89.
- [126] Tong H. Blur detection for digital images using wavelet transform[C]//IEEE International Conference on Multimedia and Expo (ICME). [S.I.], 2004: 17–20.
- [127] Ke Y, Tang X, Jing F. The Design of High-Level Features for Photo Quality Assessment[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). [S.I.]: IEEE Computer Society, 2006: 419-426.
- [128] Luo Y, Tang X. Photo and Video Quality Evaluation: Focusing on the Subject[C]//Proceedings of the 10th European Conference on Computer Vision: Part III. [S.I.], 2008: 386–399.
- [129] Xia T, Mei T, Hua G, et al. Visual quality assessment for web videos[J]. Journal of Visual Communication and Image Representation. 2010, 21: 826–837.
- [130] Dong Z, Tian X. Effective and efficient photo quality assessment[C]//2014 IEEE International Conference on Systems, Man, and Cybernetics, SMC. [S.I.], 2014: 2859–2864.
- [131] Ch.Kavitha, Rao D, Dr.A.Govardhan. Image Retrieval Based On Color and Texture Features of the Image Sub-blocks[J]. International Journal of Computer Applications. 2011, 15: 33–37.

- [132] Winder S A, Brown M. Learning local image descriptors[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.], 2007: 1–8.
- [133] Dong Z, Tian X. Effective and efficient photo quality assessment[C]//2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). [S.l.], Oct 2014: 2859-2864.
- [134] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science. 2007, 315 (5814): 972–976.
- [135] Mei T, Hua X S, Yang L, et al. Videosense: towards effective online video advertising[C]//ACM, Proceedings of the 15th ACM international conference on Multimedia. [S.l.]: ACM, 2007: 1075–1084.
- [136] Ma Y F, Zhang H J. Contrast-based Image Attention Analysis by Using Fuzzy Grouping[C]//ACM Multimedia. [S.l.], 2003: 374–381.
- [137] Manchel F. Film study: an analytical bibliography: volume 1[M]. [S.l.]: Fairleigh Dickinson Univ Press, 1990.
- [138] Shrestha P, Barbieri M, Weda H, et al. Synchronization of multiple camera videos using audio-visual features[J]. IEEE Transactions on Multimedia. 2010, 12 (1): 79-92.
- [139] Mei T, Hua X S, Zhu C Z, et al. Home video visual quality assessment with spatiotemporal factors[J]. IEEE Transactions on Circuits and Systems for Video Technology. 2007, 17 (6): 699-706.
- [140] Kim J G, Chang H S, Kim J, et al. Efficient camera motion characterization for mpeg video indexing[C]//IEEE International Conference on Multimedia and Expo. [S.l.], 2000: 1171-1174.
- [141] Campbell D, Jones E, Glavin M. Audio quality assessment techniques — a review, and recent developments[J]. Signal Processing. 2009, 89 (8): 1489-1500.
- [142] Li Z, Wang J C, Cai J, et al. Non-reference audio quality assessment for online live music recordings[C]//ACM Multimedia. [S.l.], 2013: 63-72.
- [143] an J. G. Beerends A W R, Kim D S, Kroon P, et al. Objective assessment of speech and audio quality — technology and applications[J]. IEEE Transactions on Audio, Speech & Language Processing. 2006, 14 (6): 1890-1901.
- [144] Stowell D, Plumley M, Mary Q. Adaptive whitening for improved real-time audio onset detection[C]//Proceedings of the International Computer Music Conference. [S.l.], 2007.
- [145] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing. 2004, 13 (4): 600-612.
- [146] Liu S, Yuan L, Tan P, et al. Bundled camera paths for video stabilization[J]. ACM Transactions on Graphics (TOG). 2013, 32 (4): 78.
- [147] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]//The IEEE Conference

- on Computer Vision and Pattern Recognition (CVPR). [S.l.], June 2016.
- [148] Pinheiro P H, Collobert R, Pinheiro P O. Weakly supervised object segmentation with convolutional neural networks[M]//[S.l.]: Citeseer, 2014.
- [149] Rochan M, Rahman S, Bruce N D, et al. Weakly supervised object localization and segmentation in videos[J]. Image and Vision Computing. 2016, 56: 1–12.
- [150] Shen Z, Li J, Su Z, et al. Weakly Supervised Dense Video Captioning[J]. arXiv preprint arXiv:1704.01502. 2017.
- [151] Wang C, Song Y, Li H, et al. Text Classification with Heterogeneous Information Network Kernels[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press, 2016: 2130–2136.
- [152] Wang C, Song Y, El-Kishky A, et al. Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW, Australia: ACM, 2015: 1215–1224.
- [153] Liu H, Mei T, Luo J, et al. Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing[C]//Proceedings of the 20th ACM International Conference on Multimedia. [S.l.], 2012: 9–18.
- [154] Wen W, Wu C, Wang Y, et al. Learning structured sparsity in deep neural networks[C]//Advances in Neural Information Processing Systems. [S.l.], 2016: 2074–2082.
- [155] Yang X, Mei T, Xu Y Q, et al. Automatic generation of visual-textual presentation layout[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2016, 12 (2): 33.
- [156] Ranjan A, Henrikson R, Birnholtz J P, et al. Automatic camera control using unobtrusive vision and audio tracking[C]//Graphics Interface. [S.l.], 2010: 47-54.
- [157] Zhang C, Rui Y, Crawford J, et al. An automated end-to-end lecture capture and broadcasting system[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP). 2008, 4 (1): 6.
- [158] Wang Y, Mei T, Wang J, et al. JIGSAW: interactive mobile visual search with multimodal queries[C]//Proceedings of the 19th International Conference on Multimedia 2011. [S.l.], 2011: 73–82.

致 谢

回首二十年求学生涯，我在不同阶段都得到了许多老师、同学和亲友们无私的鼓励和帮助，值此论文完成之际，我想向他们表示衷心的感谢。

从 2008 年 8 月到 2017 年 7 月，我在中国科学技术大学度过了充实而难忘的九年时间，包括四年的本科和五年的研究生生涯。我要特别感谢我的导师俞能海教授。我从本科阶段进入俞老师的实验室参与图像处理方面的学习，在大四年进入微软亚洲研究院参加多媒体计算方面的学习和研究并加入多媒体计算与通信教育部—微软重点实验室联合培养项目，到研究生的部分阶段参与实验室移动视觉搜索以及智能交通项目的开发和研究，俞老师为我提供了优良的学习环境和广阔的研究平台。俞老师始终给予我无微不至的关怀和孜孜不倦的指导，始终鼓励我要有勇气有信心做出更好的研究工作，始终用他渊博的学识和超群的远见引导着我的前进。俞老师是我学识渊博、治学严谨的导师，也是热情真诚、和蔼可亲的长辈，我所取得的每一点进步都与俞老师的关爱息息相关。俞老师教导我要做“敢/赶牛的人”，不仅鼓舞着我不断拼搏努力，也勉励着我常怀感恩之心，将他“俯首甘为孺子牛”的精神作为今后做学问做人应当追求的目标。

感谢我的导师李世鹏教授。我从本科阶段开始进入李老师在微软亚洲研究院的团队实习。作为一名懵懂的本科生，李老师为我打开了广阔的学术视野，提供了丰富的科研资源，为我在后来的研究工作打下了良好的基础。在读研期间我有幸成为了李老师的博士生，李老师对我的学习和科研给予了精心的指导，勉励我要成为一名既有深度又有广度的“T”字型科研人员。他渊博的学识、开放的思路将对我今后的研究和职业生涯产生深远的影响。

感谢我在微软亚洲研究院实习期间的导师——梅涛教授。从 2011 年 7 月份我进入微软亚洲研究院实习开始，梅老师对我进行了精心的培养和帮助，逐步带领我参见科研实践，选择合适的研究课题，指导我研究的思路和方法，引导我解决科研上遇到的问题，逐字逐句地帮助我修改学术论文。梅老师是我在科研上的启蒙导师，他用严谨的科研态度，开阔的科研眼界，高涨的科研热情，以及积极阳光的生活态度，不断鞭策着我的成长。在我实习结束以后，梅老师依然十分关心我的研究进展和职业规划，给予了悉心的指导和无私的帮助，在此表示真挚的感谢。此外，还要感谢卢适旸博士、刘衡博士、姚霆博士、傅建龙博士和吴国斌博士，你们在我实习期间给予了热情的指导和帮助，并用你们良好的科研和工作态度为我树立了榜样。

致谢

感谢我在新加坡管理大学交流学习期间的导师——Steven C.H. Hoi 教授，感谢他给予我在新加坡交流学习的宝贵机会。Steven 教授带我进入了机器学习和深度学习最前沿的研究课题，他扎实的数学基础、严谨的科研态度以及耐心的指导为我后续的研究工作打下了坚实的基础。同时，Steven 教授在生活上也给予了无微不至的关心和帮助，让我快速适应了新的环境，开始了自己的研究工作。感谢交流期间帮助我的吴鹏程博士、赵沛霖博士、王大勇博士，你们对待科研的态度和研究问题的思路方法让我受益良多。

感谢在微软亚洲研究院实习期间一起学习和进步的许多同学和朋友，包括中国科学技术大学的杨绪勇、夏睿、张婷、刘海峰、何栋梁、李明磊、练建勋、唐傲、沈旭等，中国科学院大学的刘武、杨晓鹏、吴波，清华大学的于俊杰、孙世华、钱韵子、杨也，以及在新加坡交流期间互相帮助共同进步的万吉、高兴宇、刘成昊、吕静、刘汉唐、贺智超等。感谢实验室给我帮助和关心、陪我一起进步的同学们，包括蒋锴、胡校成、王晶晶、孟垂实、朱烽、王雨农、汤闻易、费驰、殷国君、朱子平、陈胜、王国坤、李小丹、陈竹、赵坤、朱辉辉、Fangbemi Abassin Sourou 等。感谢那些属于奋斗的时光！

最后，由衷感谢我的父母，谢谢你们给了我生命，抚养我长大，教会我做人，谢谢你们给了我一个温暖的家。从我蹒跚学步，到走出学堂，你们始终是我坚实的依靠，也是我奋斗不息的动力。

在读期间发表的学术论文与取得的研究成果

已发表论文

1. **Yue Wu**, Steven C.H. Hoi, Tao Mei, Nenghai Yu. Large-scale Online Feature Selection for Ultra-high Dimensional Sparse Data. ACM Transactions on Knowledge Discovery from Data(TKDD). 2017. (Accepted, SCI, IF:1.000)
2. **Yue Wu**, Steven C.H. Hoi, Chenghao Liu, Jing Lu, Doyen Sahoo, Nenghai Yu. SOL: A library for scalable online learning Algorithms. Neurocomputing. 2017. (Accepted, SCI, IF:2.392)
3. **Yue Wu**, Xu Shen, Tao Mei, Xinmei Tian, Nenghai Yu, Yong Rui. Monet: A System for Reliving Your Memories by Theme-Based Photo Storytelling. IEEE Transactions on Multimedia(TMM). 2016, 18(11): 2206-2216. (The first two authors contributed equally, SCI, IF:2.536, EI: 20164502985964)
4. **Yue Wu**, Tao Mei, Ying-Qing Xu, Nenghai Yu, Shipeng Li. MoVieUp: Automatic Mobile Video Mashup. IEEE Transactions on Circuits and Systems for Video Technology(TCSVT). 2015, 25(12): 1941-1954. (SCI, IF:2.254, EI:20161202112799)
5. Jianlong Fu, **Yue Wu**, Tao Mei, Jinqiao Wang, Hanqing Lu, Yong Rui. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In Proceedings of the IEEE International Conference on Computer Vision(ICCV). 2015: 1985-1993. (The first two authors contributed equally, EI:20162502506587)
6. **Yue Wu**, Shiyang Lu, Tao Mei, Jian Zhang, and Shipeng Li. Local visual words coding for low bit rate mobile visual search. In Proceedings of the 20th ACM International Conference on Multimedia(MM). 2012: 989-992. (EI:20125215836578)
7. Dayong Wang, Pengcheng Wu, Peilin Zhao, **Yue Wu**, Chunyan Miao, Steven CH Hoi. High-dimensional data stream classification via sparse online learning. In Proceedings of the IEEE International Conference on Data Mining(ICDM). 2014: 1007-1012. (EI:20152801030181)

在读期间参与的科研项目

1. 国家自然科学基金面上项目，用于交通管理的复杂拥挤环境下协同视频监控理论和方法研究，(编号：61371192，2014.1-2017.12)

主要工作：协助进行项目任务分解与讨论；协调项目各项进展；负责大规模数据快速处理和多摄像头视频分析处理；发表 TKDD, TMM, TCSVT 国际期刊论文各一篇。

2. 国家重大科技专项，新型移动多媒体音视频编解码关键技术研究，(编号：2010ZX03004-003，2010.1-2012.12).

主要工作：协调项目各项进展；负责移动图片搜索核心功能的设计与实现；发表 MM 国际会议论文一篇。