# 基于图像和视频信息的社交关系理解研究综述

王 正[1),2)] 吴 斌[1),2)] 王文哲[1),2)] 滕一阳[1),2)] 帅 杰[3)]
肖云鹏[4)] 白 婷[1),2)]

[1)](北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)
[2)](北京邮电大学计算机学院(国家示范性软件学院) 北京 100876)
[3)](合肥工业大学媒体计算实验室 合肥 230601)
[4)](重庆邮电大学网络与信息安全技术重庆市工程实验室 重庆 400065)

**摘 要** 随着多媒体技术的快速发展,互联网上涌现了大量的文本、图像、视频、音频等多媒体数据.多媒体数据的特点表现为形式上多源异构、语义上互相联系.基于多媒体信息的社交关系理解是利用各种手段和方法从海量异构的多媒体数据中挖掘出有价值的信息,帮助人们快速地理解多媒体信息中的社交关系,促进多媒体内容理解、人物追踪、知识图谱的构建等多媒体数据检索和智能商业服务的发展.图像和视频是多媒体信息的重要组成部分,基于图像和视频信息的社交关系理解研究逐渐引起了学术界和工业界的广泛关注.本文主要对近年来基于图像和视频信息的社交关系理解的分类和研究现状进行总结.首先,给出问题定义并对基于图像和视频信息的社交关系理解过程进行介绍.其次,从图像和视频两个角度概括总结社交关系理解的主要研究现状.然后,在介绍已有的图像和视频数据集的基础上,对现有的主要算法进行比较分析.最后,对基于图像和视频信息的社交关系理解中的主要问题和挑战作进一步阐述.本文旨在为感兴趣的研究人员提供有益的参考,帮助其更全面地了解基于图像和视频信息的社交关系理解的研究现状,推动该领域的进一步发展.

**关键词** 多媒体特征抽取;图像内容理解;视频内容理解;社交关系理解;多元关系判定;社交理解应用
**中图法分类号** TP391 **DOI号** 10.11897/SP.J.1016.2021.01168

# A Survey of Social Relation Understanding Based on Image and Video Information

WANG Zheng[1),2)] WU Bin[1),2)] WANG Wen-Zhe[1),2)] TENG Yi-Yang[1),2)] SHUAI Jie[3)]
XIAO Yun-Peng[4)] BAI Ting[1),2)]

[1)](*Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing* 100876)
[2)](*School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing* 100876)
[3)](*Lab for Media Computing, Hefei University of Technology, Hefei* 230601)
[4)](*Chongqing Engineering Laboratory of Internet and Information Security, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

**Abstract** With the rapid development of multimedia technology, a large amount of multimedia data such as text, image, video, and audio have emerged. These multi-sources data are heterogeneous in the form while interrelated in semantics. By using the rich information from massive heterogeneous multimedia data, the aim of multimedia social relation understanding is to learn the social relation in multimedia, so as to promote the intelligent business services, such as multimedia content

收稿日期:2020-01-04;在线发布日期:2020-05-15.本课题得到国家重点研发计划项目(2018YFC0831500)、国家自然科学基金(61972047)、国家自然科学基金(U1936220)、中央高校基本科研业务费专项资金(500420824)资助.**王 正**,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为多媒体内容理解、机器学习、计算机视觉.E-mail:wangzheng123@bupt.edu.cn.**吴 斌**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为数据挖掘、复杂网络、云计算.**王文哲**,硕士研究生,主要研究方向为多媒体内容理解、机器学习.**滕一阳**,博士研究生,主要研究方向为多媒体内容理解、机器学习、计算机视觉.**帅 杰**,博士研究生,主要研究方向为数据挖掘、推荐系统.**肖云鹏**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为社交网络、机器学习.**白 婷**(通信作者),博士,讲师,中国计算机学会(CCF)会员,主要研究方向为推荐系统、多维数据挖掘、网络表示学习.E-mail:baiting@bupt.edu.cn.

understanding, character tracking, knowledge graph construction and so on. Image and video are important parts of multimedia information. The research of social relation understanding based on image and video information have gradually attracted increasing attention from both academic and industry areas. In this paper, we summarize the existing studies of social relation understanding based on image and video information in recent years. We first briefly introduce the research background and the general organization of our paper: relevant definitions, formal description of the problem, research methods and the process of social relation understanding. In the definitions of relevant concepts, we mainly introduce nine definitions from the aspects of node, edge, feature, network and so on. Problem formalization is mainly described from two aspects: relation existence judgment and relation type judgment. Then we look into the studies of social relation understanding based on image and video information. The process of social relation understanding includes four parts, namely data preprocessing, feature extraction, social relation extraction and research application. We also summarize the similarities and differences between the studies of social relation understanding in image and video areas. Afterwards, we give detailed introductions of the existing methods in social relation understanding based on both image and video information. And analyze the experiment from three parts: evaluation method, data set and comparison method based on image and video data. Finally, we make a conclusion of the problems and challenges on the social relation understanding based on image and video information. In particular, based on the technology development in social relation understanding, we divide the existing methods of social relation understanding into seven categories: co-occurrence-based methods, traditional graph-based methods, supervision-based methods, machine learning-based methods, deep learning-based methods, multimodal information-based methods and GNN-based methods. As for social relation understanding in both relation existence judgment and relation type judgment, we further classify the methods into two categories based on the number of relations: single and multiple relations. In the part of experiments, we summarize five evaluation methods, namely accuracy, precision, recall, F1 and mAP. Then we introduce the image and video datasets related to social relation understanding in recent years. In the experiments based on image information, we chose PISC and PIPA datasets for method comparisons. As for the experiments based on video information, we chose SRIV and ViSR datasets for method comparisons. Moreover, we analyze the advantages and disadvantages of the existing methods based on the experimental results. And finally summarize the problems and challenges from seven aspects, namely small sample learning, multi-source data fusion, unsupervised social relation understanding, multi-role different relation recognition, efficient relation understanding algorithm, real-time data feedback and multimedia knowledge graphs. The aim of this paper is to provide a research scope of social relation understanding based on image and video information, which may be helpful for the researchers to have a quick understanding of the field, and promote the further development in this area.

**Keywords**　multimedia feature extraction; image content understanding; video content understanding; social relation understanding; multiple relations predication; social understanding application

## 1　引　言

随着科技的日新月异，多媒体技术得到了长足的发展，并涌现出海量的多媒体数据，如文本、图像、视频、音频等. 国际数据公司(International Data Corporation, IDC)发布的最新版白皮书《Data Age 2025》[①]指出，预计到 2025 年，全球数据量总和将达到 175 ZB，这说明在当前所处的科技时代，信息数据的增长是爆炸性的. 其中，图像和视频数据已经占据总数

① Data Age 2025. https://www.seagate.com/cn/zh/our-story/data-age-2025

据量的 90%以上[1]. 因为这些具有多源异构性和多样性的非结构化数据正在快速增长并且具有很高的研究前景,所以如何从海量数据中挖掘出有价值的信息,帮助人们快速地理解社交关系并取得实际应用成为研究的热点问题[2].

社交关系是人与人之间的关系,主要是在社会生产和生活的直接交往中形成的[3-4]. 图像和视频是多媒体信息的重要组成部分,近年来基于图像和视频信息的社交关系理解有了新的进展,引起了学术界和工业界的广泛关注. 本文主要针对多媒体中基于图像和视频信息的社交关系理解进行分析和总结. 基于多媒体(特指图像和视频)信息的社交关系理解是利用各种手段和方法从海量异构的多媒体数据中挖掘出有价值的信息,帮助人们快速地理解多媒体信息中的社交关系,促进多媒体内容理解[5]、人物追踪[6]、角色发现[7]、知识图谱的构建[8]等多媒体数据检索和智能商业服务[9-10]的发展. 同时,基于多媒体信息的社交关系理解也逐渐受到谷歌、京东、爱奇艺等国内外知名企业的关注. 可见,基于多媒体信息的社交关系研究与国家和社会的发展紧密相连,顺应社会的发展趋势.

多媒体数据具有数据量大、非结构化、语义抽象、种类多样等特点,对基于多媒体信息的社交关系理解的研究不仅需要分析图像、视频等单一媒体数据,也需要综合分析多种媒体数据进而实现语义协同[1]. 其中,这涉及多个领域的知识,如多媒体[11]、计算机视觉[12]、传统机器学习[13]、深度学习[14]、模式识别[15]等,吸引了来自社会学、数学、计算机科学、语言学、复杂性科学等众多领域的研究者. 每年在 CVPR[16]、ICCV[17]、ECCV[18]、IJCAI[19]、ACM MM[20]、IJCV[21]、TMM[22]、PR[23]等国际顶级会议和期刊上都刊出了相关的工作,推动着相关研究的发展.

本文主要为了给对基于图像和视频信息的社交关系理解方向感兴趣的研究人员提供有益的参考,帮助其更全面地了解基于图像和视频信息的社交关系理解的研究现状,推动该领域的进一步发展. 对此,在广泛阅读相关文献的基础上,本文重点对近年来基于图像和视频信息的社交关系理解的分类和研究现状进行全面的分析总结,并整理出相关的图像和视频数据集以及相应的方法对比. 贡献是系统地整理出基于图像和视频信息的社交关系理解研究综述. 主要对问题定义、问题形式化以及具体处理方法进行描述,总结了基于图像和视频信息的社交关系理解过程(包括数据预处理、特征提取、社交关系抽取、研究应用). 此外,对方法分类说明、比较分析与选择进行描述,从图像和视频两个角度介绍和分析当前的研究现状. 同时,对实验中涉及的评价指标、数据集、基于图像和视频数据的实验对比方法以及当前存在的问题与挑战进行了相应的分析与总结.

本文第 1 节为引言,讲述基于图像和视频信息的社交关系理解研究背景;第 2 节对问题定义、形式化、具体处理方法以及基于图像和视频信息的社交关系理解过程进行描述;第 3 节对方法分类说明、比较分析与选择进行描述,并从图像和视频两个角度概括总结社交关系理解的主要研究现状;第 4 节在介绍已有图像和视频数据集的基础上,对现有主要算法进行比较分析;第 5 节对基于图像和视频信息的社交关系理解中的主要问题和挑战作进一步阐述;第 6 节对全文进行总结.

## 2 问题描述

本文主要对基于图像和视频信息的社交关系理解进行研究,通过从图像和视频数据中识别出人物实体,并借助提取的各种特征属性,利用学习模型去分析和推断人物之间是否存在社交关系、存在何种类型的社交关系. 其中,涉及到节点、边、特征、网络等相关定义、形式化描述、研究方法以及具体过程. 首先,我们对相关概念进行定义.

### 2.1 相关定义

**定义 1.** 角色节点集. 角色是网络中的节点,通过人脸或人体检测和识别算法,对图像和视频中的人物进行识别和标识,形成角色节点集合 $U=\{u_1,u_2,\cdots,u_n\}$.

**定义 2.** 角色边集. 角色边即网络中的角色对 $\langle u_i,u_j\rangle$ 之间的连边,$E$ 表示角色边集合. 当图像或视频中的角色 $u_i$ 和 $u_j$ 之间存在关系 $e_{ij}$ 时,将 $e_{ij}$ 添加到角色边集合 $E$ 中. 其中,$u_i\in U$,$u_j\in U$.

**定义 3.** 关系权重矩阵. 关系权重矩阵 $\boldsymbol{W}$ 是一个 $n\times n$ 的矩阵 $[w_{ij}]_{n\times n}$. 其中,$w_{ij}$ 是角色 $i$ 和 $j$ 之间边的权重,通过分析图像和视频中出现的相应角色之间的交互确定权重,它用来衡量角色之间的关系强度.

**定义 4.** 图像集. 图像集用 $D=\{d_1,d_2,\cdots,d_k\}$ 来表示. 其中,$d_o$ 表示图像集中的任意一张图片.

**定义 5.** 视频帧集. 视频是由帧组成的,我们用 $V=\{F_t\}$ 来表示整个视频,也表示一个具有相应

时间戳 $t$ 的帧 $F_t$ 的流媒体集合.

**定义 6.** 社交关系集. 社交关系集合 $R$ 是由生活中各种类型的关系组成的，通常是预先定义好的社交关系集合，也可以用来表示是否存在社交关系，每一个角色对 $\langle u_i,u_j\rangle$ 从 $R$ 中识别出关系 $r_{ij}$ 或者判别是否存在关系.

**定义 7.** 社交特征集. 所谓的社交特征集合 $P=\{p_1,p_2,\cdots,p_m\}$ 是由各种学习算法提取的时空特征(语音、光流、场景等)、语义对象(人脸、人物、物体等)和语义属性(年龄、性别、穿着等)等特征组成的. 其中，$p_1,p_2,\cdots,p_m$ 表示不同的社交特征，用来辅助社交关系的判定.

**定义 8.** 角色社交关系网络. 角色社交关系网络用 $G^r=\langle U,E,R\rangle$ 来表示. 其中，$U$ 表示角色节点集合，$E$ 表示角色之间边的集合，$R$ 表示角色社交关系集合，社交关系根据设定的阈值来判定.

**定义 9.** 角色权重网络. 角色权重网络用 $G^w=\langle U,E,\boldsymbol{W}\rangle$ 来表示. 其中，$U$ 表示角色节点集合，$E$ 表示角色之间边的集合，$\boldsymbol{W}$ 表示关系权重矩阵.

## 2.2 问题形式化描述

本节主要根据问题的不同从关系存在判定和关系类型判定两种问题出发进行形式化描述. 关系存在判定主要是判断图像和视频中的人物之间是否存在关系，从而构建网络. 关系类型判定是判断图像和视频中的人物之间具体是哪种类型的关系，从而构建角色关系网络. 下面分别从关系存在判定和关系类型判定两方面进行形式化描述.

### 2.2.1 关系存在判定问题的形式化描述

首先，利用相应的算法识别出图像集 $D=\{d_1,d_2,\cdots,d_k\}$ 或视频帧集 $V=\{F_t\}$ 中的角色节点集 $U=\{u_1,u_2,\cdots,u_n\}$ 以及抽取相应的社交特征 $P=\{p_1,p_2,\cdots,p_m\}$. 然后，设计相应的框架判断每个角色对 $\langle u_i,u_j\rangle$ 之间是否存在关系，如果存在关系则计算关系之间的权重矩阵 $\boldsymbol{W}$，从而构建角色权重网络 $G^w=\langle U,E,\boldsymbol{W}\rangle$. 具体的形式化定义如下：

$$\left.\begin{array}{l}D=\{d_1,d_2,\cdots,d_k\},V=\{F_t\}\\U=\{u_1,u_2,\cdots,u_n\}\\P=\{p_1,p_2,\cdots,p_m\}\end{array}\right\}\Rightarrow f:(D,V,U,P)$$

$$\rightarrow\begin{cases}R=1, & u_i\text{和}u_j\text{存在连边}\\R=0, & \text{其他}\end{cases}$$

$$\rightarrow G^w=\langle U,E,\boldsymbol{W}\rangle.$$

### 2.2.2 关系类型判定问题的形式化描述

首先，利用相应的算法识别出图像集 $D=\{d_1,d_2,\cdots,d_k\}$ 或视频帧集 $V=\{F_t\}$ 中的角色节点集 $U=\{u_1,u_2,\cdots,u_n\}$ 以及抽取相应的社交特征 $P=\{p_1,p_2,\cdots,p_m\}$. 然后，设计相应的框架判断每个角色对 $\langle u_i,u_j\rangle$ 之间的关系类型，从而构建角色社交关系网络 $G^r=\langle U,E,R\rangle$. 具体的形式化定义如下：

$$\left.\begin{array}{l}D=\{d_1,d_2,\cdots,d_k\},V=\{F_t\}\\U=\{u_1,u_2,\cdots,u_n\}\\P=\{p_1,p_2,\cdots,p_m\}\end{array}\right\}\Rightarrow f:(D,V,U,P)$$

$$\rightarrow R\rightarrow G^r=\langle U,E,R\rangle.$$

## 2.3 研究方法

基于图像和视频信息的社交关系理解的具体处理方法为从图像和视频数据中识别出人物实体，利用各种学习方法提取时空特征、语义对象和语义属性，通过学习模型去分析和推断人物之间是否存在社交关系，存在何种类型的社交关系，并进行更深层次的应用研究. 基于图像和视频信息的社交关系理解的整体研究框架如图 1 所示. 其中，问题的输入根据多媒体中不同数据的模态类型，分为图像和视频两种. 采用的方法是在社交关系理解任务中设计不同的学习模型. 问题的输出分为关系存在判定和关系类型判定. 当然，社交关系不仅包括关系存在判定和关系类型判定，还包括关系强度和方向的判定等，只是本文综述的内容限定在关系存在判定和关系类型判定. 研究应用包括多角色关系识别、人和关系同时识别、网络构建、知识图谱构建等. 针对图像和视频信息的社交关系理解的具体研究过程框架将在后续章节分别展开描述.

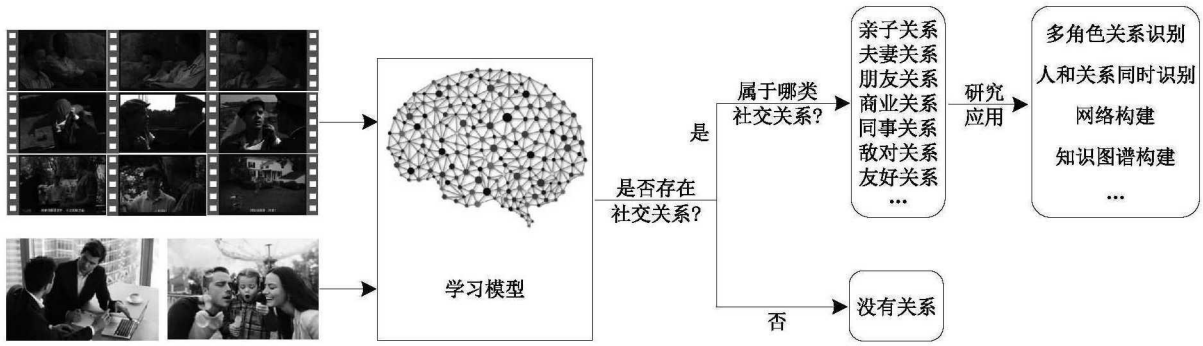


图 1　基于图像和视频信息的社交关系理解整体研究框架

### 2.4 基于图像和视频信息的社交关系理解过程

基于图像和视频信息的社交关系理解过程主要包括数据预处理、特征提取、社交关系抽取、研究应用,研究的过程框架如图 2 所示.此外,本节对图像和视频相关研究异同点进行总结.下面分别对框架的各个方面进行详细介绍.

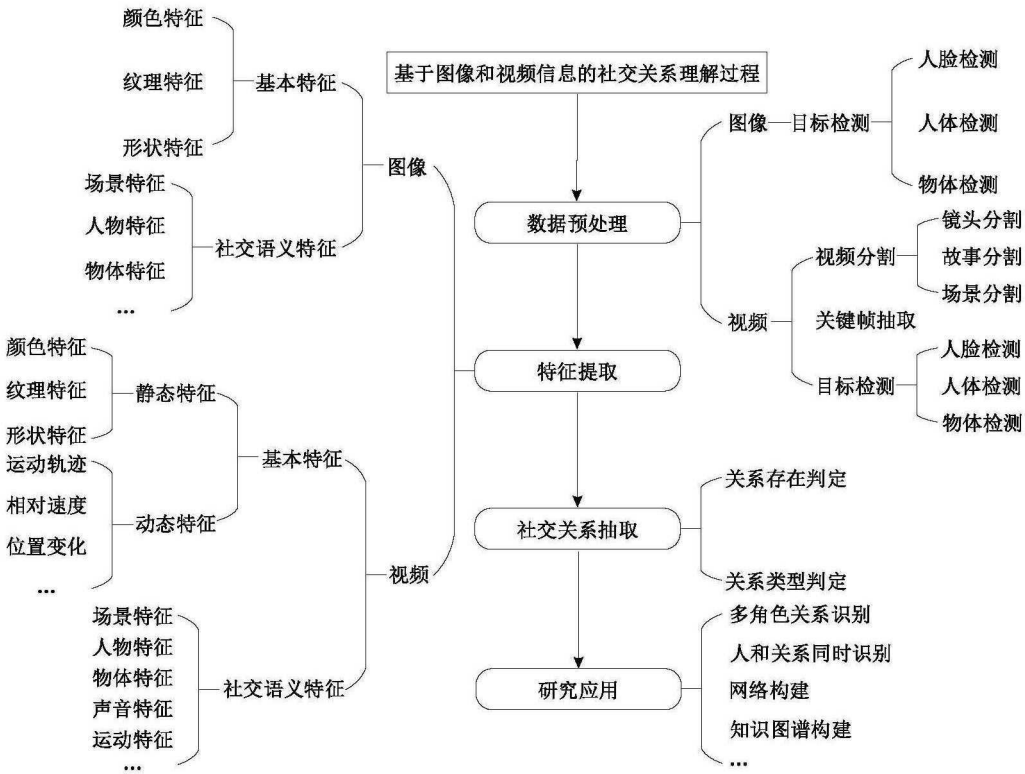


图 2 基于图像和视频信息的社交关系理解过程框架

#### 2.4.1 数据预处理

在本文中,数据预处理是针对图像和视频进行的.其中,图像和视频预处理都包括目标检测[24],目标检测一般包括人脸检测[25]、人体检测[26]和物体检测[27].而视频预处理还包括视频分割[28]、关键帧抽取[29],这些预处理工作属于社交关系理解中非常重要的部分.首先,我们对视频分割和关键帧抽取进行相应的介绍.然后,对目标检测进行相应的介绍.

(1)视频分割

视频分割是将一段视频按照不同的分割方法划分为不同的视频片段,目的是从中划分出相关的实体对象,通常将这种实体对象称为视频对象[30].简单来说就是通过某种手段或方法,把视频按照需求进行截断分割,选取需要的部分.在视频分割中,最常用的是基于视频镜头进行分割.之所以进行视频镜头分割,主要是因为镜头分割作为视频处理的第一步,可以为分类、分析、索引和查询高层内容打下坚实的基础[31].镜头分割的准确性将直接影响社交关系理解的准确性,并且为关键帧的抽取奠定基础.视频镜头分割的典型方法包括基于像素[32]、基于直方图[33]、基于轮廓[34]的方法等,以及一些改进的视频镜头分割方法如基于双重检测[35]、基于 GIST 特征和条件判定[36]的方法等.

除了基于视频镜头分割外,还有基于视频故事分割和基于视频场景分割的方法,不同的分割方法对于社交关系的理解存在差异.例如,因为新闻视频的视频片段和语音内容差异比较大,所以对于新闻视频大多使用基于故事分割的方法.而影视视频中的场景信息比较丰富,能够为视频内容理解提供重要信息,所以对于影视视频使用场景分割的方法也非常普遍.然而,对于故事分割点不明确的影视视频来说,使用先前的故事分割方法相对困难[2].对此,Lv 等人[37]提出一种基于视频多层次特征的故事分割方法,利用分水岭算法并结合分层提取的视频内容特征进行故事分割,构建人物关系网络.

(2)关键帧抽取

关键帧指用来描述一个镜头内部主要内容的某帧或某几帧图像,关键帧抽取是在视频分割的基础上进行的.之所以进行关键帧抽取,主要是因为它可以减少视频帧之间存在的大量冗余信息,并且更凝练地表达一段视频中包含的信息,便于对视频内容建立索引并进行管理.在基于视频信息的社交关系

理解中，进行关键帧抽取可以在不影响实验结果的情况下极大缩短处理时间. 关键帧抽取的算法包括基于聚类的方法[38]、基于运动分析的方法[39]、基于改进的谱聚类算法[40]、基于卷积神经网络(Convolutional Neural Network，CNN)和图形处理单元算法[41]等.

(3) 目标检测

目标检测作为基于图像和视频信息的社交关系理解中非常重要的步骤，能够对社交关系理解的准确性产生直接的影响. 目标检测的任务是对任意一幅图像中存在的对象进行定位和分类，并用矩形框对其进行标记. 一般将信息区域选择、特征提取、分类作为传统目标检测模型的三个阶段[24]. 在计算机视觉的基本问题中，目标检测具有重要的研究意义，它能够为理解图像和视频中的语义提供有价值的信息，涉及到许多应用领域，包括图像分类[42-43]、人类行为分析[44]、社交关系理解[45]和自动驾驶[46]等. 在基于图像和视频信息的社交关系理解过程中，目标检测通常分为不同的子任务，如人脸检测、人体检测和物体检测. 其中，人脸检测和人体检测主要对人物的面部和整体进行检测，而物体检测主要对除人物以外的其它物体进行检测. 通过利用目标检测提供的有价值信息能够提高社交关系理解的准确度. 通用的目标检测方法框架主要分为两类，第一类为基于区域生成的方法，第二类为基于回归/分类的方法，每一类都包含多种方法，表 1 列出了其中的部分方法.

**表 1　目标检测方法分类**

| 目标检测分类 | 方法 | 发表会议 |
|---|---|---|
| 基于区域生成 | R-CNN[47] | CVPR 2014 |
| | Fast R-CNN[48] | ICCV 2015 |
| | Faster R-CNN[49] | NeurIPS 2015 |
| | R-FCN[50] | NeurIPS 2016 |
| | FPN[51] | CVPR 2017 |
| | Mask R-CNN[52] | ICCV 2017 |
| | Cascade R-CNN[53] | CVPR 2018 |
| | R-DAD[54] | AAAI 2019 |
| | RRPN[55] | AAAI 2020 |
| 基于回归/分类 | MultiBox[56] | CVPR 2014 |
| | AttentionNet[57] | ICCV 2015 |
| | YOLO[58] | CVPR 2016 |
| | G-CNN[59] | CVPR 2016 |
| | SSD[60] | ECCV 2016 |
| | YOLO9000[61] | CVPR 2017 |
| | YOLOv3[62] | arXiv 2018 |
| | M2Det[63] | AAAI 2019 |
| | Spiking-YOLO[64] | AAAI 2020 |

### 2.4.2　特征提取

随着多媒体和网络技术的快速发展，随时都有大量的图像和视频数据产生. 视频作为一种重要的媒体形式，包含了丰富的时空特征信息，如何全面地提取特征信息对基于图像和视频信息的社交关系理解具有重要的意义. 特征提取可以理解为计算机视觉任务中的方法和处理过程，它是社交关系理解过程中的关键步骤，是否能够提取尽可能完整地反映多媒体内容的理想特征将直接影响社交关系理解的准确率，因为特征是不同的并且不同特征的重要性也不相同.

多媒体特征提取是针对图像和视频进行的，两者之间的特征既存在相同点也存在不同点. 我们将图像特征分为基本特征和社交语义特征，基本特征包括颜色、纹理和形状特征，社交语义特征包括场景、人物和物体特征等. 视频特征也分为基本特征和社交语义特征，基本特征又分为静态特征和动态特征. 静态特征包括颜色、纹理和形状特征，动态特征包括运动轨迹、相对速度和位置变化等信息，社交语义特征包括场景、人物、物体、声音和运动特征等.

(1) 图像的基本特征

颜色作为图像最重要的特征之一，主要由颜色空间或模型定义. 颜色空间包括 RGB、LUV、HSV 和 HMMD 等[65]. 纹理是另外一种重要的特征，具有很强的识别能力. 通常，颜色是像素属性，而纹理只能从一组像素中测量[66]. 形状也是一种重要的特征，人们认识世间万物主要是以形状为线索，将简单的几何形状进行编码[67].

(2) 图像的社交语义特征

图像的社交语义特征根据其描述对象的不同可以分为场景特征[68]、人物特征[69]和物体特征[5]. 社交语义特征相比于基本特征来说包含更多能够反映多媒体内容的丰富信息，导致在基于图像信息的社交关系理解中社交语义特征所起到的作用更大，而基本特征所起到的作用相比于社交语义特征来说微乎其微. 因此，当前的研究中绝大多数使用的是社交语义特征. 其中，所谓的场景特征是指图像中人物所处的环境以及周围的事物等信息，物体特征也属于场景中的一部分，它们能够为基于图像信息的社交关系理解提供非常重要的线索. 人物特征相对来说比较丰富，例如人脸表情、穿着、年龄、动作等，再加上人和物体之间的交互信息，同样能够很好地促进社交关系理解. 而社交语义特征的提取广泛使用 CNN[43]等深度学习方法.

(3) 视频的基本特征

视频的每一帧都代表一个图像，每一帧的图像特征都是静态特征. 因此，使用上述的静态图像特征

提取方法就可以处理帧特征.动态特征是图像所不具备的,例如摄像机在拍摄过程中会出现摇镜头、推拉、跟踪等操作以及镜头中对象的运动轨迹、相对速度、位置变化等[70].

(4)视频的社交语义特征

视频的社交语义特征根据其描述对象的不同可以分为:场景特征[71]、人物特征[72]、物体特征[73]、声音特征[74]和运动特征[75].其中,场景特征指视频中人物所处的环境以及周围的事物等信息,但相比于图像的场景特征,视频中的场景是动态变化的,情况更加复杂.人物特征中的人脸表情和人物姿态随着时间的推移是不断变化的,而图像中的表情和姿态是固定不变的.物体特征是指视频中出现的物体,例如汽车、电脑、椅子、书籍等,因为物体通常只占视频帧的一部分,所以我们又称之为局部社交语义特征.声音特征和运动特征是视频所特有的,声音特征主要是指视频中的音频信息,随着场景和情绪的变化而变化.运动特征的存在主要是因为视频是由连续的图像帧组成的.在基于视频信息的社交关系理解中,社交语义特征是最核心的特征,将对最终结果产生直接的影响.然而从底层物理特征映射到社交语义特征存在"语义鸿沟"的问题[76],尽管已经得到相应的解决,但是由于数据的复杂性,仍然存在大量的问题.图像和视频的特征提取内容总结如表2所示.

**表2 特征提取的内容分类**

| 类别 | 基本特征 | | | | | | 社交语义特征 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 颜色特征 | 纹理特征 | 形状特征 | 运动轨迹 | 相对速度 | 位置变化 | 场景特征 | 人物特征 | 物体特征 | 声音特征 | 运动特征 |
| 图像 | √ | √ | √ | × | × | × | √ | √ | √ | × | × |
| 视频 | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |

2.4.3 社交关系抽取

社交关系抽取首先要对多媒体中的不同实体进行识别;然后利用相应的方法对数据进行分析,判断不同的两个实体之间是否存在关系;最后,如果实体间有关系存在,那么需要进一步确定存在的关系类型[2],其准确率的高低取决于数据预处理、特征提取的准确性以及学习模型的选择.解决此任务需要运用各种学习算法将与人物行为相关的语义线索进行整合与分析[77].近年来,基于静态图像数据的人物社交关系抽取已经取得一定的研究成果,其中关系存在判定的方法主要是基于共现的方法[78],而关系类型判定主要有基于传统机器学习的方法[79-80]、基于深度学习的方法[12,81]等.由于视频中的人物关系具有抽象性、动态性以及人物交互具有复杂的时序性,导致基于图像信息的社交关系抽取方法并不适用于复杂的视频数据.因此学者们提出适用于视频数据中人物社交关系抽取和推断的研究方法,如CNN+LSTM(Long Short-Term Memory)[82]、CNN+TSN(Temporal Segment Networks)[83]、GGNN(Gated Graph Neural Network)+LSTM[84]、PGCN(Pyramid Graph Convolution Network)+TSN[5]等.通过上述方法可以推断图像和视频中人物之间的社交关系,从而帮助我们更好地理解多媒体内容.

2.4.4 研究应用

研究应用是在社交关系抽取的基础上进行的,当前的研究主要针对多媒体中两个角色之间的社交关系进行抽取,因此可以考虑多角色之间的关系识别.同时,还可以考虑角色和关系同时识别,从而促进多媒体内容理解[5]、人物追踪[6]、角色发现[7]、知识图谱的构建[8]等多媒体数据检索和智能商业服务[9-10]的发展.

2.4.5 图像和视频研究的异同点总结

在本节中,我们对图像和视频这两种多媒体数据研究的各方面异同点进行分析总结,详细信息如表3所示.

**表3 图像和视频相关研究异同点总结**

| 类别 | 图像 | 视频 |
|---|---|---|
| 不同点 | (1)静态的,不具有时序性.<br>(2)单模态信息,包括人脸、人体和场景等空间信息.<br>(3)图像是静态的,并且人物都在同一张图像中.<br>(4)只需要进行图像的预处理.<br>(5)研究比较成熟. | (1)动态的,具有时序性.<br>(2)多模态信息,包括时间、空间和语音信息等.<br>(3)视频是动态的,具有时序性,多个人物甚至不在同一帧中出现,导致人物之间的交互更加复杂.<br>(4)在图像预处理基础上需要进行视频分割、关键帧抽取等操作.<br>(5)研究相对较少. |
| 相同点 | (1)研究的问题包括人物关系识别和网络构建,评价指标相同.<br>(2)处理流程都包括目标检测,特征提取以及人物关系抽取.<br>(3)都使用了颜色、纹理和形状等底层特征以及表情、性别和年龄等高层特征.<br>(4)都可以使用基于图数据的图神经网络(Graph Neural Network, GNN)方法处理人和物体之间的交互问题. | |

由表 3 可知，图像和视频之间的不同点主要体现在数据特点、属性、场景、处理流程以及研究现状等方面；相同点主要体现在研究内容、评价指标、处理流程、特征属性以及处理方法等方面.

## 3　基于图像和视频信息的社交关系理解

在本节中，根据数据类型的不同我们将社交关系理解分为图像和视频两种；包括两类任务，分别为关系存在判定和关系类型判定；涉及 7 种方法，包括共现方法、传统图方法、监督信息、机器学习、深度学习、多模态以及 GNN. 下面分别进行介绍.

### 3.1　基于图像和视频信息的社交关系理解方法

以社交关系理解方法的发展和技术演化的时间脉络为主线，基于图像和视频信息的社交关系理解方法可分为共现方法、传统图方法、监督信息、机器学习、深度学习、多模态和 GNN. 接下来对方法说明、方法比较分析与选择进行阐述.

#### 3.1.1　方法说明

(1) 共现方法

所谓的共现方法是指如果在图像或视频中同时出现两个实体，那么他们之间可能存在某种关系. 当两者共现的次数越多，说明他们之间的联系越密切，通过挖掘出的共现信息来进行社交关系理解. 共现方法是基于图像和视频信息的社交关系理解中最基本的方法.

(2) 传统图方法

传统图方法是指以点和边组成的图结构. 其中，点表示图像或视频中的角色，边表示角色之间的关系. 此外，边既可以是有向的，也可以是无向的. 传统图方法可以非常直观地表示出图像和视频中的社交关系，让人快速地理解其中的内容. 传统图方法也是基于图像和视频信息的社交关系理解中最基本的方法.

(3) 监督信息

所谓的监督信息就是在模型训练之前需要对数据进行人工标注，然后借助于标注数据去训练得到最优模型. 例如，在进行社交关系理解过程中，需要对人物关系、年龄、性别等特征进行标注，借助于标注信息更好地进行社交关系理解. 监督信息经常与机器学习、深度学习相结合，从而进行有监督学习.

(4) 机器学习

机器学习是一门多领域交叉学科，涉及概率论、统计学、凸优化、算法复杂度理论等多门学科. 传统机器学习中包含多种算法，例如最大期望(Expectation-Maximization，EM)、支持向量机(Support Vector Machine，SVM)、条件随机场(Conditional Random Field，CRF)、逻辑斯蒂回归(Logistic Regression，LR)等，不同算法的作用也各不相同. 机器学习主要包括两类任务：回归和分类. 其中，分类就是将新数据划分到合适的类别中，一般用于类别型的目标特征. 如果目标特征为连续型，则往往采用回归方法. 机器学习主要分为监督学习和无监督学习，监督学习是基于输入数据和指定目标值进行学习，而无监督学习是不指定目标值或预先无法知道目标值，将相似或相近的数据划分到一起，常用的解决方法为聚类. 除此之外机器学习还包括半监督学习和强化学习. 当前已有的社交关系理解方法大多是有监督学习.

(5) 深度学习

深度学习是用于建立、模拟人脑进行分析学习的神经网络，用来学习样本数据的内在规律和表示层次，在学习过程中获得的信息对基于图像和视频信息的社交关系理解帮助很大. 传统的深度学习相关算法包括 CNN、循环神经网络(Recurrent Neural Network，RNN)、LSTM 等多种算法，广泛应用于计算机视觉、语音识别、自然语言处理等多个领域并表现出优异的性能. 深度学习模型有两大优势，第一，深度学习模型可以挖掘数据中更多的隐含信息；第二，深度学习模型可以进行表示学习，能够从原始数据中自动学习到有用的特征，从而减少人为设计特征造成的不完备性，并能够产生更为有效的特征表示，更好地促进社交关系理解. 在当前社交关系理解问题中最常用的方法为 CNN.

(6) 多模态

本文中的多模态主要是指融合多种模态的数据进行社交关系理解，例如文本、图像、视频、音频等. 相比于单模态，多种模态的数据能够弥补单模态数据彼此的不足，提供更加丰富的特征信息，有助于更好地促进社交关系理解.

(7) GNN

除了传统的深度学习方法外，图神经网络(GNN)[85]可以说是近几年 AI 领域的新贵，它是一种连接模型，通过在图节点之间传递消息来捕获图的依赖性. 相比于传统的深度学习方法，GNN 保留了一种状态，这种状态可以用任意深度表示来自其邻域的信息. GNN 包括 GCN、图注意力网络(Graph

Attention Network,GAT)、GGNN 等,可以用于各种图数据上的监督、半监督及强化学习,是一种广泛应用的图分析方法.

3.1.2 方法比较分析与选择

在所有方法中,共现方法和传统图方法是基于图像和视频信息的社交关系理解中最基本的方法,也是最常用的方法.当然,如果在视频帧中两个人没有同时出现,那么就需要借助于其它方法,例如使用多模态中的文本信息辅助角色关系的挖掘.如果需要保留邻域的信息,传统图方法就显得力不从心,这时候可以借助于 GNN 去保留一种用任意深度表示其邻域信息的状态,弥补传统图方法的不足.

监督信息经常与机器学习和深度学习方法联合使用,从而进行有监督学习.使用监督信息的优点是可以提高社交关系理解的准确度,缺点是需要对数据进行标注.如果数据量过大,数据标注既费时又费力.

传统机器学习方法相比于深度学习方法来说,首先对小数据量表现更好,而深度学习方法则需要大量的数据.其次,机器学习方法更容易进行计算,而深度学习方法则需要借助 GPU 在合理的时间内对大量数据进行训练,代价比较大.最后,机器学习方法相比于深度学习方法更容易解释和理解,深度学习属于"黑匣子"型,其内部结构并不能完全了解.

深度学习方法相比于传统机器学习方法来说,首先可以使用数据进行有效缩放.例如深度学习方法可以通过更多的数据来提高社交关系理解的准确度.其次,利用深度学习模型可以进行表示学习,通过从原始数据中自动学习到有用的特征表示,从而更好地促进社交关系理解.最后,深度学习方法适应性强,易于转换,能够应用于不同领域.

多模态相比于单模态来说,它可以提供更多有用的信息,从而使社交关系理解更加准确.但是,多模态对数据集要求很高,数据集中应该包含多种模态的信息,并且信息的融合也需要相应的方法.

GNN 可以说是传统深度学习之上的算法,它借助于图结构的强大表征力,表现出优异的性能和可解释性,本质上是一种基于拓扑信息的特征提取过程,弥补了传统深度学习方法难以应用到非规则形态数据上的缺陷.而 GNN 的不足之处在于并不能在较短时间内得到训练和预测结果,时间方面仍然存在很大的挑战.

在方法选择方面,不管是对于关系存在判定任务还是关系类型判定任务来说,共现方法和传统图方法都是最基础的方法.当进行社交关系理解时,如果需要借助于人工标注数据,那么就需要选择监督信息进行模型训练.机器学习和深度学习方法的选择比较特殊,需要根据具体问题进行选择.例如,当解决聚类问题时可以选择机器学习中的聚类算法;当需要提取图像或视频中的特征信息时可以选择传统深度学习方法,比较常用的是 CNN;当进行社交关系识别时,既可以使用深度学习方法,又可以使用机器学习和深度学习相结合的方法,还可以将它们与监督信息相结合.特别是现在深度学习成为普遍的方法,使用的频率也越来越高,性能表现优越.多模态方法的使用需要根据数据集的形态进行选择,如果数据集中包含多种类型的数据并且相吻合,那么多模态方法就再合适不过了.GNN 主要是解决和图结构有关的问题,例如在社交关系挖掘、知识图谱构建与推理等方面都是极佳的选择.总的来说,需要根据具体问题来选择合适的方法,从而更好地解决问题.

## 3.2 基于图像信息的社交关系理解

本节将主要叙述和分析基于图像信息的社交关系理解的国内外研究现状.根据任务的不同主要从关系存在判定和关系类型判定两方面进行分析,具体的分类总结如表 4 所示.其中,"√"代表涉及该方法,"×"代表不涉及该方法.为了便于读者理解和查阅,在本章节中列出了所涉及的英文缩写以及对应的英文全称,具体内容如表 5 所示.

**表 4 基于图像信息的社交关系理解方法分类总结**

| 任务 | 方法 | 描述 | 时间 | 输出 | 方法分类 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 共现方法 | 图方法 | 监督信息 | 机器学习 | 深度学习 | 多模态 | GNN |
| 关系存在判定 | MSN[86] | 构建照片中的网络,并探究亲密度与位置的关系. | 2008 | 关系网络 | √ | √ | × | × | × | × | × |
| | FDCR[87] | 基于人脸检测和聚类结果的亲近性度量方法. | 2009 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | GCA[88] | 基于图聚类算法检测社会集群. | 2009 | 关系网络 | √ | √ | √ | √ | × | × | × |

（续　表）

| 任务 | | 方法 | 描述 | 时间 | 输出 | 方法分类 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 共现方法 | 图方法 | 监督信息 | 机器学习 | 深度学习 | 多模态 | GNN |
| 关系类型判定 | 单一关系 | TSL+AM[89] | 基于转移子空间算法和相关元数据的语义相关性预测亲属关系. | 2012 | 关系类型 | √ | √ | √ | √ | × | × | × |
| | | BoFG[78] | 基于 BoFG 和共现线索发现社交子图，预测成对关系. | 2012 | 关系类型 | √ | √ | √ | √ | × | × | × |
| | | GA[90] | 融合门控自编码器和判别神经网络层进行亲属关系验证. | 2014 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | GBKR[80] | 基于图方法进行亲属关系识别. | 2014 | 关系类型 | √ | √ | √ | √ | × | × | × |
| | | JIA[91] | 基于无监督的联合推理算法进行角色推断. | 2015 | 关系类型 | √ | √ | × | √ | × | × | × |
| | | RSBM[92] | 基于 RSBM 和空间投票的特征选择方法进行亲属关系识别. | 2015 | 关系类型 | √ | √ | √ | √ | × | × | × |
| | | CNN+SSC[93] | 基于 CNN 和半监督聚类算法进行亲属关系识别. | 2018 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | KML+DNN[94] | 基于亲属度量学习和深度神经网络进行亲属关系验证. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | LPGA+MIAN[95] | 基于自我监督和注意力机制进行亲属关系识别. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | 多元关系 | DCN+BL+SC[96] | 基于 DCN、桥接层和空间线索进行社交关系识别. | 2015 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | DSCN[97] | 基于双流 CaffeNet 进行社交关系识别. | 2017 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | DGM[98] | 基于 Dual-Glance 模型进行社交关系识别. | 2017 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | S-DCN[99] | 基于 DCN 和空间线索识别社交关系. | 2018 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | GRM[8] | 基于 GGNN 和注意力机制进行社交关系识别. | 2018 | 关系类型 | √ | √ | √ | √ | √ | × | √ |
| | | DLA+LS+SA[81] | 基于标签空间的层次结构和社会属性进行社交关系识别. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | SRG-GN[100] | 基于场景和属性上下文，利用 GRU 进行社交关系识别. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | MGR[101] | 基于多粒度语义特征，利用 GCN 进行社交关系识别. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | √ |

**表 5　英文缩写和英文全称总结**

| 英文缩写 | 英文全称 |
|---|---|
| MSN | Measuring Social Networks |
| FDCR | Face Detection and Clustering Result |
| GCA | Graph Clustering Algorithm |
| TSL+AM | Transfer Subspace Learning+Associated Metadata |
| BoFG | Bag-of-Face-Subgraphs |
| GA | Gated Autoencoder |
| GBKR | Graph-Based Kinship Recognition |
| JIA | Joint Inference Algorithm |
| RSBM | Relative Symmetric Bilinear Model |
| SSC | Semi-Supervised Clusters |
| KML+DNN | Kinship Metric Learning+Deep Neural Network |
| LPGA+MIAN | Local Parts Guided Attention+Multiple Inputs Attention Network |
| DCN+BL+SC | Deep Convolutional Network+Bridging Layer+Spatial Cue |
| DSCN | Double-Stream CaffeNet |
| DGM | Dual-Glance Model |
| S-DCN | Siamese-Deep Convolutional Network |
| GRM | Graph Reasoning Model |
| DLA+LS+SA | Deep Learning Architecture+Label Space+Social Attributes |
| SRG-GN | Social Relationship Graph Generation Network |
| MGR | Multi-Granularity Reasoning |

#### 3.2.1　关系存在判定

关系存在判定主要是判断图像中的人物之间是否存在关系，并不需要判断存在何种类型的关系，所涉及的方法主要为共现方法.

Wu 等人[88]研究了关系存在判定的问题，提出一种通过观察照片中个体的共现来构造一个加权无向图的方案. 其中，边的权重是所涉及个体（图中节点）的社交亲密性的度量. 他们首先通过检查照片中个体的共现性，从照片集中构造一个加权无向图. 其中，顶点表示照片集合中出现的人，而连边表示两个个体的共同出现. 个体之间的亲密度用权重来表示. 然后在图聚类算法基础上对社交集群进行检测，揭示社交聚类. Golder[86]利用数码照片集对社交网络进行研究，从中抽取出对人物社交关系有价值的信息. 通过对 23 名受试者照片的分析，验证了方法的可行性. 此外，对感知亲密度与网络位置的关系以及未来可能出现的问题进行讨论. 当许多照片中

都同时出现相同的两个人时，表明这两个人之间的关系十分紧密. 因此，加权图是两个节点之间的每一个链接都由一个值或权重来表示的图. Wu 等人[87]提出一种基于人脸检测和聚类结果的贴近度度量方法，设计并实现了一个计算框架，实现人脸检测和聚类等计算量较大的工作. Close & Closer 程序为了揭示人们的亲密关系，以一个全新的角度对照片集进行分析，从而推动照片共享和人们的交流.

本节对图像中的关系存在判定任务进行分析，主要借助于共现方法. 通过判断两个人是否同时出现和计算的关系权重来判断两个人之间关系存在问题以及关系的亲密程度. 关系存在问题还涉及图方法、监督信息和机器学习方法.

3.2.2 关系类型判定

关系类型判定是判断图像中的人物之间具体是何种类型的关系. 以人为中心的图像相关的最有趣的一点是图像中人与人之间的关系. 本文主要从两个角度展开，分别为单一关系和多元关系，具体内容的阐述如下所示.

(1) 单一关系

所谓的单一关系主要是指亲属关系识别. 基于图像信息的亲属关系识别是社交网络重构和分析中的一个重要问题. 因此，大量的学者将亲属关系识别作为一个研究方向进行研究.

Qin 等人[92]对关系类型判定中的单一关系问题进行了研究，提出一种新的相对对称双线性模型(RSBM)和一种空间投票的特征选择方法. 首先，使用双线性函数来模拟父母和孩子之间的相似性，通过从数据中学习的类似协方差的矩阵来获取父母和孩子之间的依赖性. 为了使模型更加健壮，引入了一个新的相对双线性相似模型，有效地整合了先验知识. 其次，使用提出的空间投票特征选择方法，在考虑局部空间信息的情况下为孩子-父母对选择最具鉴别性的特征. 与传统基于群组的特征选择方法相比，该方法本质上是允许整幅图像中的特征相互竞争，然后选择对应局部区域中个体特征较高部分的组. 最后，发布了一个针对三主体亲属关系问题的新的人脸数据库，其特征是超过 1000 个孩子-父母群体，使用该方法得到了最先进的结果. 提出方法的总体架构如图 3 所示.

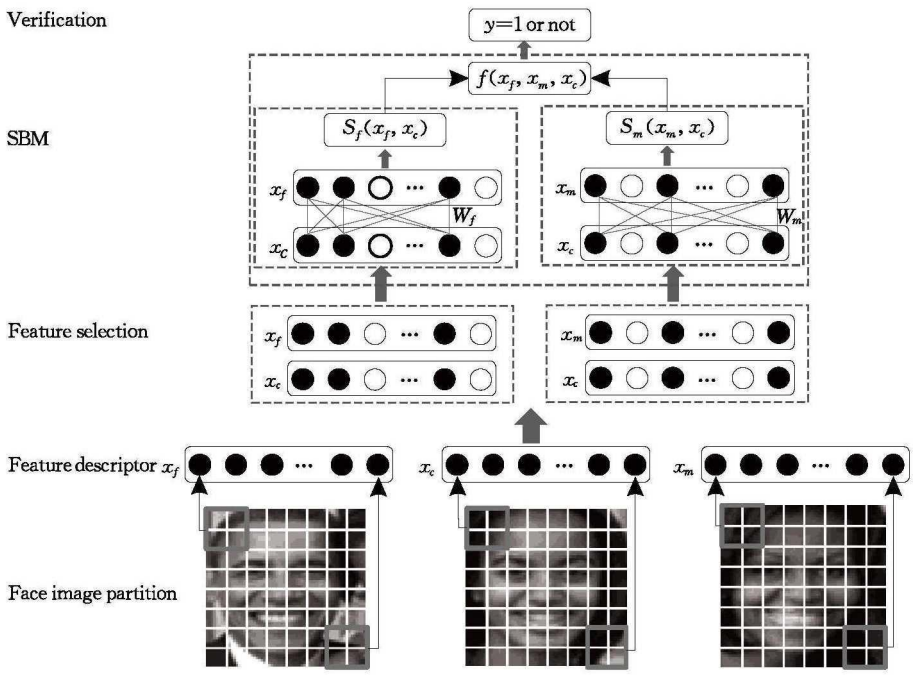


图 3 RSBM 总体框架图[92]

图 3 中的架构大致分为三个阶段. 特别是在第一阶段，将一个图像分割成多个重叠的 patch(位置)，并从每个 patch 中提取一个中层特征描述符(如 128 维的 SIFT 特征)，然后将其连接到一个特征向量中作为下一阶段的输入. 在第二阶段，为了提高鲁棒性，将最具辨别力的局部面部 patch 使用空间投票的特征选择方法来进行选择. 在第三阶段，使用双线性模型学习父母和孩子之间的相似性，并在此基础上进行最终的亲属关系验证.

Xia 等人[89]同时构建了 UB KinFace Ver2.0 和

FamilyFace 两个数据库. 在外观分布上，由于老人和儿童的面部图像存在较大差异，对此开发一种基于转移子空间的学习算法来减少差异. 此外，为了预测图像中最可能的亲属关系，在研究相关元数据语义相关性的基础上提出一种预测算法. Chen 等人[78]提出了一个新的框架. 首先，联合具有不同属性和位置的人脸，使其成为一个人脸图. 然后，设计一种方法自动发现信息子图，使用 BoFG 表示一组照片并作为一个新的特征. 最后，预测人脸图中的成对关系使用的是子图中的共现信息. Dehghan 等人[90]旨在回答两个关键问题：(1) 后代是否与父母相似？(2) 后代更像父母中的哪一方？对此提出一种将通过门控自编码器发现的特征与一个判别神经网络层相结合，学习最优特征即遗传特征，以描绘亲子关系的算法，并对模型自动发现的特征和人类学研究中发现的特征之间的相关性进行分析. Guo 等人[80]为了提高亲子关系识别的性能，在提出的基于图的方法上将所有家庭成员之间的面部相似性整合到一张照片中. 该方法使用一个全连接图来建模亲缘关系. 其中，面部由顶点表示，边由亲缘关系表示. 每一个可行的亲属关系图的总得分是通过将分类器在图的边上的得分相加来计算的，选择总分最高的图作为预测依据. 此外，文中还引入 Sibling-Face Database 和 Group-Face Database 两个数据库. Dai 等人[91]提出一种无监督的 EM-style 联合推理算法. 该算法主要是一种联合推理算法，将检测到的人脸身份和角色分配之间的相关成对关系借助于概率 CRF 算法进行建模. 实验结果表明，在所有照片中，使用身份和角色的独立估计相比于联合两者进行推断表现不佳，并且联合推理使识别许多不同照片中同一个体的能力得到了提升.

在大数据的背景下，深度学习已经成为大多数人工智能问题的首选技术[102]，尤其是已经在语音识别[103]、自然语言处理[104]、计算机视觉[105-106]和无人驾驶[107]等许多领域展现出优异的表现.

对此，Zhou 等人[94]提出一种新的带有耦合深度神经网络(DNN)模型的亲属度量学习(KML)方法，主要是为了克服度量学习中不考虑较大的年龄跨度和父母子女图像之间的性别差异导致人脸遗传相似性度量性能下降的局限性. KML 明确地对亲子对固有的跨代差异进行建模，学习耦合的深度相似性度量，拉近有亲缘关系的人脸图像对，推远没有亲缘关系但具有较高外观相似度的人脸图像对. 此外，通过在耦合 DNN 上施加连接内多样性和连接间一致性，将层次紧密度引入到耦合网络中，以便在有限的亲属关系训练数据下进行深度度量学习. KML 方法的示意图如图 4 所示.

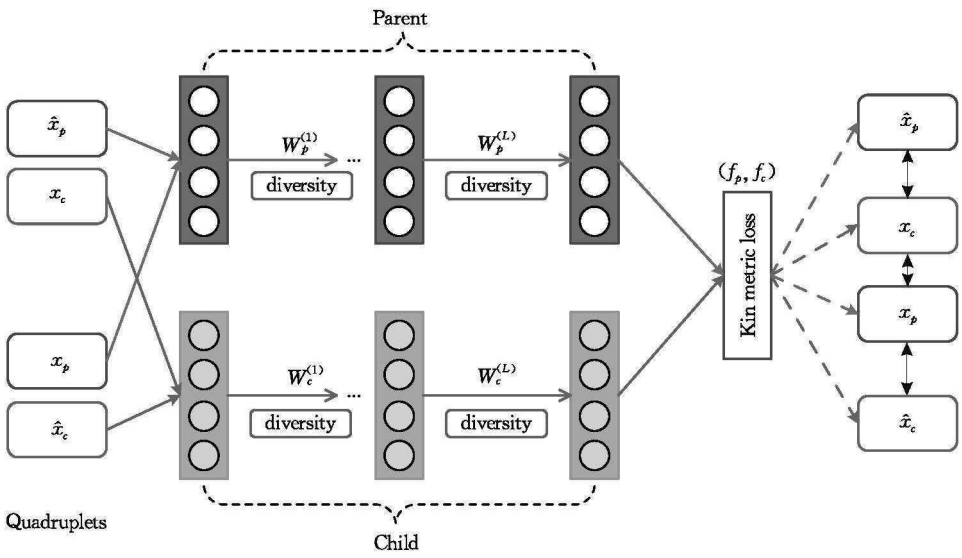


图 4　KML 方法示意图[94]

在图 4 中，KinNet 由一对神经网络(父母、孩子)组成，分别为具有不同模型参数和输出的父母、孩子神经网络而设计. 该耦合体系结构明确考虑了由于父母-孩子之间的年龄跨度大、性别差异大而不应该共享同一图像变换的领域知识. 该方法为四元输入，经过耦合网络，其结果将被输入亲属关系度量层进行损失评估. 然后，将判别(梯度)信息回传到下层以更新网络参数，从而使在某些正则化条件下的

经验损失最小化.

Joseph 等人[93]使用深度学习方法对亲属关系识别进行研究,提供一些亲属关系确认和家庭分类的基准以及一个最大的视觉亲属识别数据库(Families In the Wild,FIW),其面部编码器是使用预先训练的 CNN,比传统方法的性能更突出.通过对 FIW 上 CNN 模型的微调,还可进一步改善两种任务的结果.将测量的人类观察者的表现与计算机视觉算法进行比较,结果表明人类在识别亲属关系任务中已经不如 CNN 模型. Yan 等人[95]以提取人脸局部信息为核心,提出一种基于注意力网络的人脸亲缘关系验证方法.该方法的创新点主要是将注意力机制引入深度网络中,提取高层次特征进行人脸表示,以区别于使用低层次特征的验证方法.此外,还提出一种自我监督的方法来引导注意力网络.

(2)多元关系

所谓多元关系主要是针对图像中存在的多种类型的社交关系进行识别,不仅仅针对亲属关系.由于人物之间的交往是一个复杂的过程,从而导致社交关系的类型多种多样.如何准确地预测图像中人物之间的社交关系是一个值得研究的问题.

对于图像中的社交关系研究主要考虑场景中人物共存以及通过学习人脸表情、身体外观特征、空间语义信息来实现,并且广泛运用机器学习和深度学习方法. Li 等人[98]对关系类型判定中的多元关系问题进行研究,提出基于注意力机制的双重视角模型.该模型模拟了人类的视觉系统,以探索对社交关系分析有用且互补的视觉线索.其中第一视角着重于关注人物对,然后根据它的外观信息和几何信息进行粗略的预测.第二视角利用区域建议网络(Region Proposal Network,RPN)生成的区域上下文线索来细化粗略预测,提出使用注意力的 R-CNN,将注意力分配到每个上下文区域.通过选择性地关注相关区域,可以获得更好的性能.社交关系的推断是根据学习到的人物和空间语义特征进行的.此外,提出一个新的大规模数据集(People In Social Context,PISC).其中,包含 22 670 个图像、76 568 个注释样本以及 9 种类型的社交关系.模型框架图如图 5 所示.

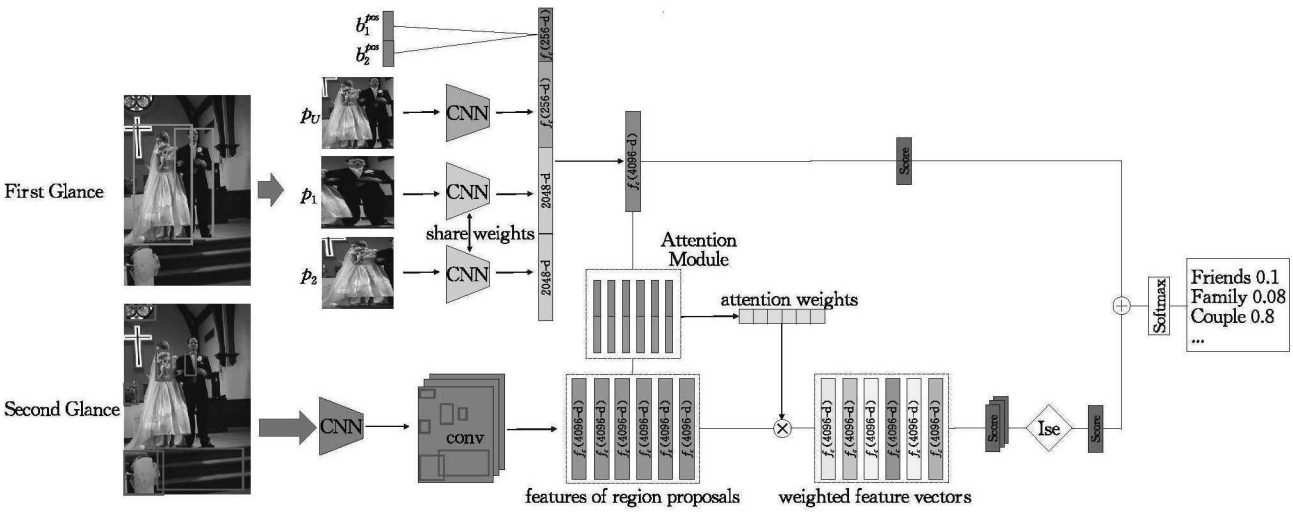


图 5 Dual-Glance 模型框架图[98]

图 5 概述了所提出的双视角模型.第一视角中的输入是一张图片以及人物边界框,包含的是两张单人照片和一张联合照片,分别输入到三个 CNN 中.最后,将 CNN 的输出以及位置信息进行拼接.第二个视角着眼于区域建议,将注意力分配到每个区域并汇总它们的输出以细化评分.注意力模块是由第一视角的信号和形成局部上下文的第二视角的信号组成,从而进行社交关系识别.

Zhang 等人[96]提出一个深度模型,主要是为了研究在自然环境的人脸图像中能否表征和量化细粒度和高层次的关系特征.该模型学习一个丰富的人脸表示来捕捉性别、表情、头部姿势和年龄相关的属性,然后执行成对人脸推理进行关系预测. Sun 等人[97]探讨了一个具有挑战性的问题,即识别照片中的社交关系.因为先前的工作对社交关系的分类存在片面性,所以采用社会学研究领域的

方法将社交关系进行全面的分类. 他们还构建了一个基于领域理论的数据集(People In Photo Album, PIPA),包含 5 个域和 16 种人物关系,并利用多种特征以及 CNN、SVM 等方法对人物关系类型进行识别. 次年,Zhang 等人[99]在此前基础上继续研究从自然环境的人脸图像中表征和量化细粒度、高层次的关系特征,提出一种用于面部表情鲁棒识别的深层网络结构. 为了突破现有模型只学习面部表情的局限性,设计了一个能够从辅助属性(如性别、年龄和头部姿势)中学习的多任务网络. 通过一种新的属性传播方法解决训练过程中要求数据集应有完整的标签问题. 为了预测人际关系,在 Siamese 模型中加入表情识别网络. Aimar 等人[81]首先构建了一个新的数据集 EgoSocialRelation,提出一种自动分类社交行为的方法. 该方法仅仅依靠用户所看到的内容. 在 Bugental 社会理论的基础上,将人际关系分为五个相关的社会领域. 他们提出的方法是一种新的深度学习架构,它提供社会交互的语义表示是结合标签空间的层次结构和在框架层中估计的一组社会属性来实现的. Goel 等人[100]提出一种能够从给定的图像中生成一个结构化和统一化的社交关系以及属性表示的端到端神经网络. 他们设计的 SRG-GN 是一种使用场景和属性上下文迭代更新社交关系状态的网络,主要借助于类似门控递归单元(Gated Recurrent Units,GRUs)的记忆单元. 该神经网络之所以比之前的社交关系识别方法有明显的改进,主要是因为图中节点与边之间的信息传递是利用 GRUs 之间的循环连接来实现的.

然而,上述只考虑人物属性以及场景中人与物的共存,忽略了认知社交关系更重要的信息,即全局信息以及人与物之间的相互作用. 再加上图被广泛应用于计算机视觉和多媒体任务中并表现出良好的性能,因此越来越多的研究者将 GNN 应用到社交关系推理中.

Wang 等人[8,108]考虑到以往研究中忽略了人周围的语境信息之间的相互作用对社交关系识别的影响,并发现人和周围的语境信息之间的相互作用可以通过一个具有适当信息传播和注意力的新的结构化知识图谱来有效地建模. 对此,提出一个端到端的图推理模型(GRM),学习一种传播机制. 通过图传播节点信息,探索感兴趣的人和上下文对象之间的交互作用,并结合注意力机制和门控图神经网络(GGNN)将结构化的知识有效地集成到深层神经网络结构中,以促进社交关系的理解. 图推理框架如图 6 所示.

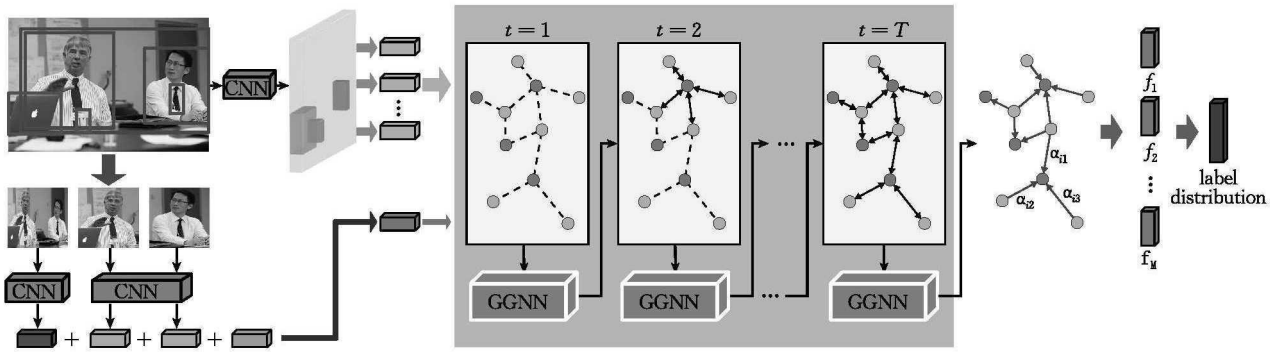


图 6　图推理框架图[8]

图是组织场景中社交关系和语义对象之间的关联关系,节点表示社交关系和语义对象,连边表示他们共同出现的概率. 由图 6 可知,首先,给定图像和感兴趣的人对,GRM 从人对区域提取特征,并使用这些特征初始化关系节点. 然后,使用预先训练好的 Faster R-CNN 检测器来搜索图像中的语义对象,并提取其特征来初始化相应的对象节点. 最后,利用 GGNN 使节点信息在图中传播,计算节点特征,并引入图注意力机制. 通过测量最具鉴别性的目标节点对各关系节点的重要性,实现人物社交关系的识别.

接着,Zhang 等人[101]考虑到现有的研究存在一定的局限性,即只对局部的多种特征进行处理,无法全面捕捉场景、人物区域线索、人与物体之间的交互等多粒度语义. 对此,提出一种基于图像的多粒度社交关系识别推理框架. 全局知识从整个场景中学到,而中层细节则从人与物的区域中学到. 更重要的是,利用人的细粒度姿势关键点建模人与物以及成对人之间的交互. 基于图,利用图卷

积网络进行社交关系推理.

本节分析了图像关系类型判定中的单一关系和多元关系问题.单一关系即亲属关系,而多元关系为日常生活中存在的多种关系类型.两者都表现出随着时间的推移,方法上不断创新,最为明显的是深度学习方法的引入.而在多元关系判定问题中,还引入了图神经网络,并且在性能方面表现优异.

### 3.3 基于视频信息的社交关系理解

本节将主要叙述和分析基于视频信息的社交关系理解的国内外研究现状.相比于图像数据,视频中人物的出现和交互更加复杂,具有抽象性、动态性和时序性,处理过程更加繁琐.根据任务的不同主要从关系存在判定和关系类型判定两方面进行分析,具体的分类总结如表 6 所示.其中,"√"代表涉及该方法,"×"代表不涉及该方法.为了便于读者理解和查阅,在本章节中列出了所涉及的英文缩写以及对应的英文全称,具体内容如表 7 所示.

**表 6 基于视频信息的社交关系理解方法分类总结**

| 任务 | | 方法 | 描述 | 时间 | 输出 | 方法分类 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 共现方法 | 图方法 | 监督信息 | 机器学习 | 深度学习 | 多模态 | GNN |
| 关系存在判定 | | RoleNet[109] | 基于 RoleNet 量化关系并构建角色社交网络. | 2009 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | | SVR+GP[110] | 基于支持向量回归和高斯过程的统计学习进行社交网络分析. | 2010 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | | 2D-PCA+CIIPA[111] | 基于上下文的演员关联层次解析方法挖掘演员关系. | 2010 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | | FEC[112] | 基于电影剪辑线索量化角色互动并构建角色社交网络. | 2012 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | | CS+TC[113] | 基于累积相似度和时间约束建立电影角色关系图. | 2013 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | | ACES[114] | 基于一种主动聚类与集成算法提取社交网络结构. | 2014 | 关系网络 | √ | √ | √ | √ | × | × | × |
| | | DCII+C-RNN[115] | 基于深度概念层次和卷积递归神经网络构建社交网络. | 2015 | 关系网络 | √ | √ | √ | √ | √ | × | × |
| | | MV+Spark[116] | 基于多视图和 Spark 构建社交网络. | 2017 | 关系网络 | √ | √ | √ | √ | √ | × | × |
| | | SRE-Net[117] | 基于 MoCNR 和场景的方法构建社交关系网络. | 2018 | 关系网络 | √ | √ | × | √ | √ | × | × |
| | | StoryRoleNet[37] | 基于 StoryRoleNet 构建社交网络. | 2018 | 关系网络 | √ | √ | √ | √ | √ | √ | × |
| 关系类型判定 | 单一关系 | SCCED[45] | 基于类似 Siamese 的卷积译码耦合网络识别亲属关系. | 2017 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | DML[118] | 基于自己的基准评估和比较不同的度量学习方法,验证亲属关系. | 2018 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | SMNAE[119] | 基于 SMNAE 的深度学习框架验证无约束视频亲属关系. | 2018 | 关系类型 | √ | √ | √ | √ | √ | × | × |
| | | DSN[120] | 基于深层 Siamese 网络实现亲属关系的多模态融合. | 2019 | 关系类型 | √ | √ | √ | √ | √ | √ | × |
| | 多元关系 | CD+TSPGM[121] | 基于概念检测器和时间平滑的概率图模型分析社交关系. | 2011 | 关系类型 | √ | √ | √ | √ | × | × | × |
| | | CRF+TVI[122] | 基于条件随机场和可处理的变分推理识别社交关系. | 2013 | 关系类型 | √ | √ | √ | √ | × | × | × |
| | | MSDL[83] | 基于多流深度学习识别社交关系. | 2018 | 关系类型 | √ | √ | √ | √ | √ | √ | × |
| | | STMV[82] | 基于多视角的时空注意力模型识别社交关系. | 2019 | 关系类型 | √ | √ | √ | √ | √ | √ | × |
| | | TSM[84] | 基于时空特征、语义对象和传播知识图识别社交关系. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | √ |
| | | MSTR[5] | 基于多尺度时空推理框架识别视频中的社交关系. | 2019 | 关系类型 | √ | √ | √ | √ | √ | × | √ |
| | | NM[123] | 基于神经模型预测角色对之间的相互作用和社交关系. | 2020 | 关系类型 | √ | √ | √ | √ | √ | √ | × |

表 7　英文缩写和英文全称总结

| 英文缩写 | 英文全称 |
|---|---|
| RoleNet | Role's social Networks |
| SVR+GP | Support Vector Regression+Gaussian Processes |
| 2D-PCA+CIIPA | 2D-Principal Component Analysis+Correlations IIierarchical Parsing Approach |
| FEC | Film-Editing Cues |
| CS+TC | Cumulative Similarity+Temporal Constraints |
| ACES | Active Clustering with Ensembles for Social structure extraction |
| DCII+C-RNN | Deep Concept IIierarchies+Convolutional-Recursive Neural Network |
| MV+Spark | Multi-View+Spark |
| SRE-Net | Social Relation Network |
| StoryRoleNet | Story Role Network |
| SCCED | Siamese-like Coupled Convolutional Encoder-Decoder |
| DML | Distance Metric Learning |
| SMNAE | Supervised Mixed Norm regularization AutoEncoder |
| DSN | Deep Siamese Network |
| CD+TSPGM | Concept Detectors+Temporal Smooth Probability Graph Model |
| TVI | Tractable Variational Inference |
| MSDL | Multi-Stream Deep Learning |
| STMV | Spatio-Temporal attention model based on Multi-View |
| TSM | Two-Stage Model |
| MSTR | Multi-scale Spatial-Temporal Reasoning |
| NM | Neural Model |

### 3.3.1　关系存在判定

关系存在判定主要是判断视频中的人物之间是否存在关系并构建人物关系网络. 当前的研究无外乎以下三种情况：第一，为了判断人物之间关系的紧密度，引入高斯加权的方法；第二，人物之间建立连边主要依靠共现的时间或场景；第三，构建人物之间的关系网络根据人物与场景目标的共现[2].

Weng 等人[109]针对关系存在判定的问题，并且为了恰当地描述电影中的角色关系，设计了一种量化关系并构建角色社交网络的方法，称为 RoleNet. 基于 RoleNet 能够执行语义分析，这超越了传统的基于特征的方法. 在这项工作中，视频场景的上下文信息是由人物之间的社交关系构成的，并且人物中的主角以及所处的社区也可以自动确定. 此外，通过对视频场景的角色语境描述，提出了基于社交关系的故事分割方法. 接着，Ding 等人[110]首先使用支持向量回归对视觉和听觉等局部特征进行估计，从而对电影角色之间的关系进行学习. 然后，为了得到角色之间的亲和力，使用高斯过程的统计学习对这些局部属性进行综合. 最后，根据学到的亲和力进行社交网络分析，找到角色所属的社区并确定社区领导者. Yuan 等人[111]介绍了一种演员关联挖掘框架，该框架是基于图方法进行设计的. 为了定位演员，使用人脸检测和跟踪方法，而预处理则使用 2D-PCA 检测器. 他们还提出一种基于上下文的参与者关联层次解析方法，目的是构建演员关系图. Yeh 等人[112]研究了电影中角色社交网络的自动构建问题. 与现有的使用共现信息来衡量两个特征之间的关系方法不同，作者认为描述特征之间的交互比共现的方法更有意义. 对此，提出一种利用电影剪辑线索量化角色互动的新方案，并在此基础上构建角色的社交网络. 此外，还展示了一个利用自动构建的社交网络来发现角色社交集群的应用程序. Chiu 等人[113]提出了在不同的头部运动和场景光照下，使用人脸轨迹来模拟同一人物的面部. 同时，作者提出了一种新的测量方法来评估不同长度的两个面部轨迹之间的相似性. 然后，在时间约束的基础上，构建电影角色关系图. Barr 等人[114]提出了一种提取一组视频剪辑中出现的人物社交网络结构的方法. 通过将来自不同视频的相似面孔分组，形成一个代表个体的身份聚类. 每个身份聚类由社交网络中的一个节点表示. 如果在一个或多个视频帧中，来自聚类的人脸同时出现，则将两个节点连接起来. 该方法结合了一种新的主动聚类技术，根据用户对模糊匹配人脸的反馈创建更精确的身份聚类. 最后的输出包括一个或多个表示社交群体的网络结构. Nan 等人[115]利用 DCH 和卷积递归神经网络(C-RNN)模型对剧中人物之间的社交网络进行分析. DCH 利用多级结构将电视剧人物的视觉语言概念转化为多样化的抽象概念，利用马尔可夫链蒙特卡洛算法提高概念空间组织的检索效率.

然而，上述方法只考虑人物视频的底层镜头和场景特征，忽略了视频的社交语义特征，因此容易造成构建人物关系网络不准确的现象.

对此，Lv 等人[37]提出了一个 StoryRoleNet 模型，用于构建一个准确完整的网络来表示角色之间的关系. 首先，在每个故事单元中采用高斯加权法对关系的权值进行度量，主要是为了避免相邻故事单元分割点之间关系的冗余计算问题. 更重要的是，通过分析视频的层次特征，提出了一种新的长视频故事分割方法. 然后，为了将一些缺失关系进行填充，把视频和字幕文本进行结合来构建关系网络. 最后，对最终的网络进行分析，从而发现社区和重要角色. 对模型的综合评价使用了三部电影和一部电视剧. StoryRoleNet 模型框架图如图 7 所示.

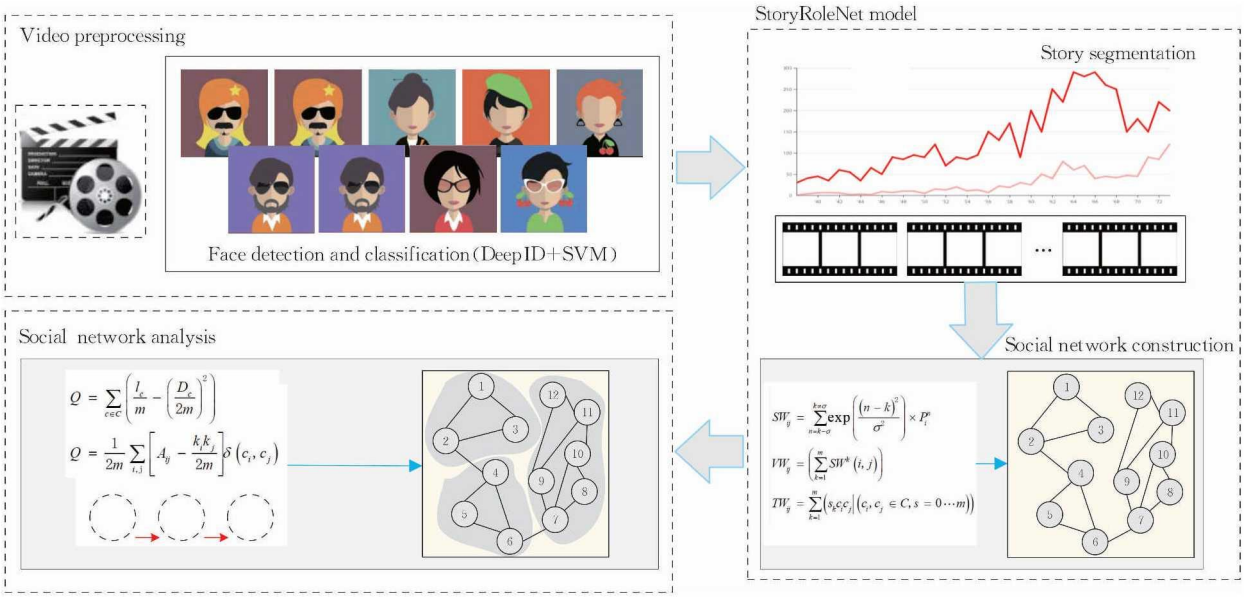


图 7 StoryRoleNet 模型框架图[37]

从图 7 中可以看出,该框架包括三个部分:视频预处理、包含故事分割和社交网络构建的 StoryRoleNet 模型、社交网络分析. 第一部分的视频预处理主要是利用 DeepID 和 SVM 方法进行面部检测和分类. 第二部分的 StoryRoleNet 模型包括两个子任务:故事分割和社交网络构建. 其中,故事分割主要采用分水岭算法,对视频内容进行分层提取,然后进行故事分割. 而社交网络的构建首先是计算角色之间关系的权重,然后使用指定的权重阈值构建社交关系网络. 第三部分的社交网络分析主要是进一步调查社区,采用 Lovain 方法发现社区,并使用模块化来度量社区结构强度,挖掘所构建网络中的重要角色.

Zhou 等人[116]提出了构建角色关系的社交网络方法. 该方法在多视图的非结构化数据基础上进行. 首先,自动提取视频中的人脸. 其次,在多视角基础上对角色社交网络进行构建. 最后,在 Spark 平台上,针对大量的非结构化数据,实现该方法的并行实现. 接着,Zhou 等人[117]提出角色节点识别方法(Method of Character Nodes Recognition, MoCNR)用于无监督特征识别. 该算法利用无监督双聚类方法识别视频中的特征节点,解决以往工作中监督学习和单聚类方法的不足. 其次,为了解决在不同视频帧中无法识别人物社交关系的问题,提出一种基于场景分割的社交关系识别方法. 最后,综合上述过程提出一种 SRE-Net 模型,实现视频自动社交关系网络构建.

本节对视频中的关系存在判定问题进行分析,发现深度学习以及多模态方法被逐渐利用,并且使用了更加丰富的特征信息. 其中,最基本的方法仍是共现方法. 判断出视频中人物之间是否存在关系并构建社交关系网络.

### 3.3.2 关系类型判定

关系类型判定主要是对视频中的人物之间属于何种类型的关系进行判断,需要借助于相应的学习模型. 这样可以对视频中的潜在语义信息进行挖掘,帮助人们更好地理解视频内容. 本文主要从两个角度进行展开,分别为单一关系和多元关系.

(1) 单一关系

所谓的单一关系主要是指亲属关系识别. 近年来,基于图像信息的亲属关系识别得到了广泛研究. 然而,由于基于视频信息的亲属关系识别是一个探索相对较少的研究领域,在安全、监视和移民控制等各种环境中具有很高的应用价值. 因此,学者们将基于视频信息的亲属关系识别作为一个具有较好前景的方向进行研究.

Hamdi[45]对关系类型判定中的单一关系问题进行了研究,探讨了利用一对亲属的面部表情的视觉转换,学习一种有效的基于视频信息的亲属关系识别的面部表征. 为此,提出一种类似 Siamese 的耦合卷积编解码网络. 为了揭示亲属关系中的相似模式,同时抛弃没有亲属关系的人之间也可以观察到的相似模式,在视觉表征空间中定义了一种新的对比损失函数. 为了进一步优化,使用基于特征的对比损失对所学习的表征进行微调. 该

模型采用了一种表情匹配方法，以最小化亲属间表情差异的负面影响. 在短期的面部动态变化基础上，使用滑动时间窗口对每个亲属视频进行分析. 整体上评估了七种不同的亲属关系，所使用的是亲属之间的微笑视频，提出的模型框架如图 8 所示.

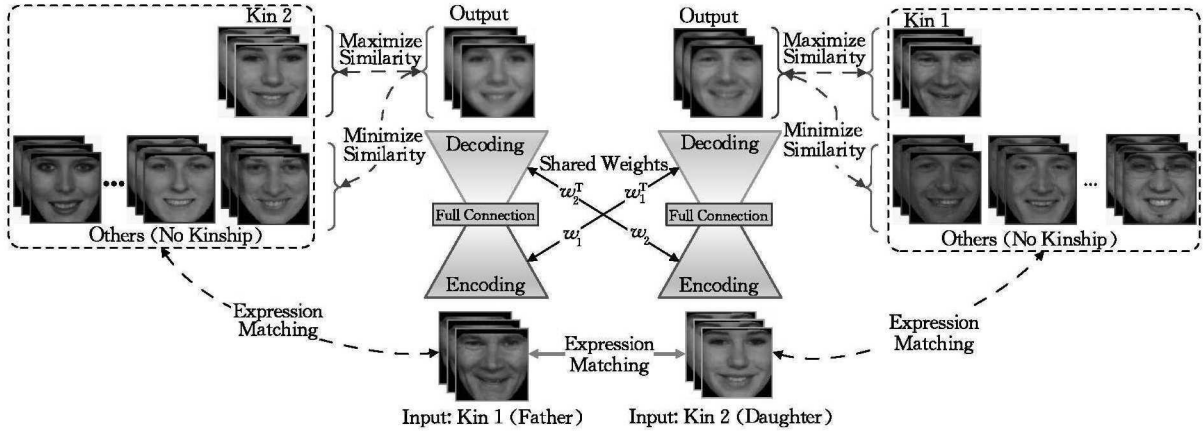


图 8　耦合卷积编解码网络[49]

在图 8 中，首先将匹配的子序列输入到耦合卷积编解码网络中，该网络学习具有亲属关系的给定对象对的面部外观之间的转换. 为了揭示亲属对之间的面部相似性，编解码网络的每个网络输出一个与其输入的亲属对相似的面部图像，同时最小化与非亲属对的相似性. 为此，定义了基于外观的对比损失. 经过训练网络后，去除解码块，并将具有对比损失的分类层连接到全连接块上. 然后，通过基于特征的对比损失函数，以监督的方式对修改后的网络进行微调. 最后，通过融合所有匹配子序列对的后验概率得到验证结果.

Yan 等人[118]主要利用人脸信息对视频中的亲属关系验证问题进行研究. 首先，提出一种新的视频人脸数据集，称为 KFVW(Kinship Face Videos in the Wild). 然后，基于自己提出的基准对几种先进的度量学习方法进行比较. 最后，对人类亲属关系识别能力进行评估，发现人类的识别能力超过基于度量学习的识别方法. Kohli 等人[119]提出一个深度学习框架，称为监督混合范数正则化自编码器(SMNAE)，主要是为了验证无约束视频中的亲属关系. 其中，引入一种特定类别的稀疏权重矩阵. 对亲属关系进行验证主要利用从视频中学习到的时空表示. Wu 等人[120]为了更加准确地验证亲属关系，首次将人脸和语音信息进行融合. 首先，提出了一个新的数据集 TALKIN(TALking KINship)，它融合了多模态信息. 然后，为了实现对亲属关系的多模态融合，提出一个深层的 Siamese 网络. 实验表明，相比于基准的单模态和多模态技术，提出的 Siamese 网络实现了更高的精确度.

(2) 多元关系

所谓的多元关系主要是针对视频中存在的多种类型社交关系进行识别，而不仅仅针对亲属关系. 相比于静态图像中的社交关系识别，视频中的情况更加复杂. 视频中的两个人不一定在同一帧中，位置信息不易获得，并且视频中交互的两个人可能只有侧脸或者背面等，导致如何确定视频中人物之间的社交关系变得更加困难.

Lv 等人[82]对关系类型判定中的多元关系问题进行了研究. 首先，为了获得从低级像素到高级社交关系空间的丰富表示，提出了一种基于多视角的时空注意力模型(STMV)，包括视频的 RGB、光流和人脸图像. 其次，为了得到更健壮的社交关系特征表示，提出了多维注意力单元模块，可以关注每个特征的多个方面. 最后，为了显式地聚合社交关系空间的多视角特征，引入张量融合层，学习多视角特征之间的相互作用. 模型的整体框架图如图 9 所示.

在图 9 中，STMV 模型主要由三部分组成，分别为时空特征提取、多维注意力单元和张量融合层. 首先，使用卷积神经网络来提取包括 RGB、光流和人脸在内的多视角特征，从而获得信息更丰富的社交关系特征表示. 其次，为了更好地理解社交关系，使用长短期记忆(LSTM)来研究多视角的时间特征，并且在注意力模块中引入多维注意力单元. 通过这种方式，可以从视频中生成一个针对社交关系特征多方面的合适特征表示，从而构建从低级视频像素到高级社交关系空间的良好映射功能. 最后，为了融合多视角特征，引入张量融合层学习特征之间的交互作用.

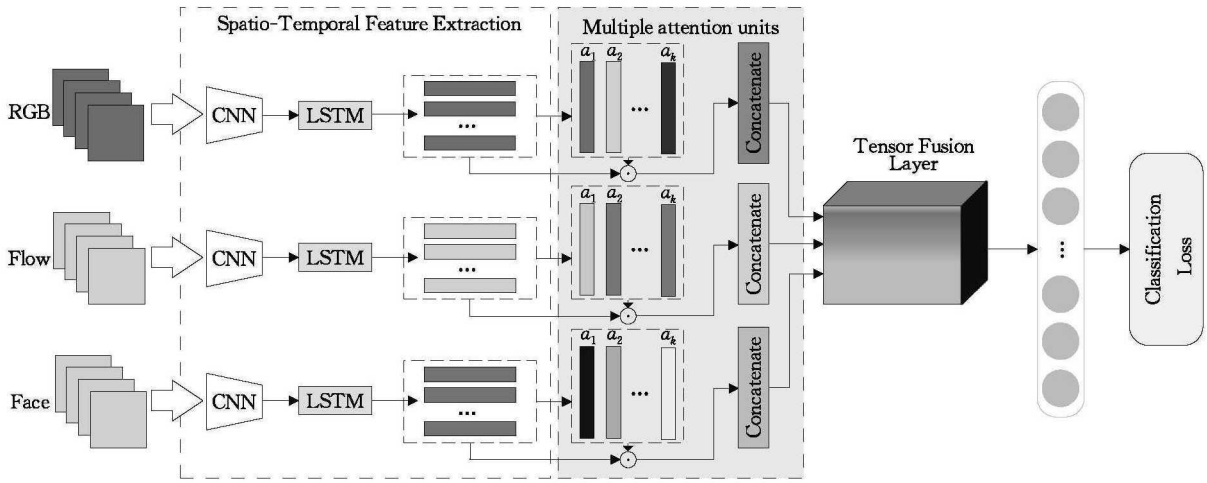


图 9 STMV 模型框架图[82]

Ding 等人[121]通过使用概念评分向量来表示视频内容,评分向量的提取主要使用大规模的概念检测器,并将它们作为分组线索去检测社交结构.为了对参与者之间的社交关系进行分析以及对社区进行检测,使用具有时间平滑的概率图模型. Ramanathan 等人[122]提出一个带有注释的新的社交角色数据集.为了模拟角色间的相互作用,提出使用条件随机场以及特定于人的社会描述符,主要是因为考虑到社交角色是由事件中人与人之间的相互作用来描述的.为了能够同时推断模型权重以及分配角色给视频中的人,开发了一种可处理的变分推断. Lv 等人[83]为了解决视频中人物社交关系识别存在的挑战,提出一个融合多模态信息的多流模型.首先,构建一个新的视频数据集(Social Relation In Videos, SRIV),这是一个带有 16 种社交关系标签的数据集.其次,为了学习视频中的时间、空间以及音频等社交语义信息,将提出的多流深度学习模型作为识别社交关系的基准.最后,将学习到的高层语义信息使用逻辑回归进行结合,准确地识别社交关系. Kukleva 等人[123]第一次尝试根据视觉和语言线索预测电影中角色之间的相互作用和关系,构建的神经模型以视频片段、对话和角色对的形式有效地编码了多模态信息,这些信息在预测交互作用时被证明是互补的.此外,该模型可以学习在视频片段中定位角色,同时在不显著降低性能的情况下使用弱标签预测交互和关系.

然而,上述方法只考虑了全局和粗粒度的特征,忽略了视频中人和物体之间的交互.在计算机视觉领域中,很多因素都可以用图来表示,例如像素、区域等,从而建模不同任务之间的关系,任务包括物体形状检测[124]、图像分割[125]、图像检索[126]、车辆搜索[127]等.近年来,大量研究者使用 GNN 这样端到端的网络对图中的消息传播等任务进行研究[128],借助于这样的思路,这些方法逐渐被应用到计算机视觉任务中.

对此,Liu 等人[6]提出了一个多尺度时空推理框架(MSTR)对社交关系进行识别,主要考虑到视频中关键人物的位置不固定甚至不在同一帧中,以及故事情节和人物的行为动作对社交关系识别的影响.对于空间表示,全局动作和场景信息的学习采用时间分段网络(TSN),而人和物体之间的视觉关系则使用设计的三重图模型.对于时间域,为了获得视频中的长、短期故事情节,提出一种金字塔图卷积网络(PGCN)进行多尺度接收域的时间推理.通过这种方法,MSTR 可以在时空维度上对多尺度行为和故事情节进行探索,更准确地推理社交关系. MSTR 模型的整体框架图如图 10 所示.

在图 10 中,MSTR 模型主要包括两部分.第一部分是三重图结构的构建.该框架以一个视频片段为输入,为了从感兴趣的区域捕获局部细节,使用 Mask R-CNN 在 MS-COCO 数据集上进行预处理,从帧中裁剪出人和物体.为了建模人和物体的时空表示,为相同的人构建 Intra-Person 图(IntraG),为不同的人构建 Inter-Person 图(InterG),并构建 Person-Object 图(POG),以捕获人与物体的共存.采用 ResNet 提取人和物体的空间特征.第二模块采用 PGCN 在每个图中通过消息传播进行关系推理.此外,采用 TSN 或 T-C3D 等全局视频分类网络,学习全局特征.最后,结合 TSN 的全局特征和 PGCN 的推理特征,对视频中的社交关系进行预测.

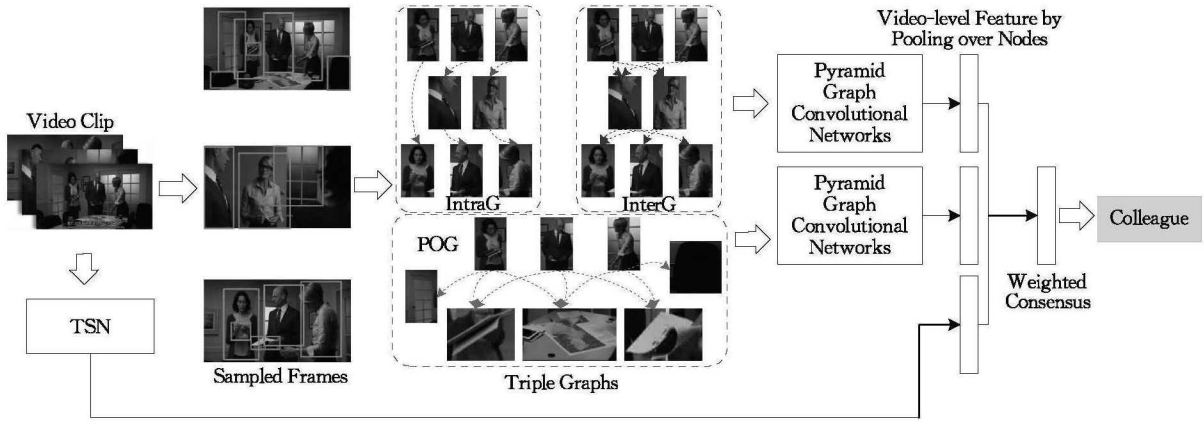


图 10　MSTR 模型框架图[5]

Dai 等人[84]提出一个两阶段模型(TSM)，主要是考虑到当前的研究大多忽略隐藏在语义对象中丰富的上下文信息. 在第一阶段，为了获得丰富的特征表示，分别对时空特征和语义对象信息进行提取. 在第二阶段，考虑到语义对象和视频场景之间的交互对社交关系识别的影响，引入 GGNN 进行表示. 由于语义对象在不同场景下的有效性存在差异，因此，构建注意力模块进行测量.

本节对视频关系类型判定中的单一关系和多元关系进行了分析，不仅考虑了深度学习和多模态，而且引入了当前比较流行的图神经网络. 尤其是在多元关系判定中，通过结合图神经网络和更丰富的特征信息，使得模型性能得到了显著的提升，帮助人们更好地理解视频内容.

## 4　实　验

在本节中，主要针对基于图像和视频信息的社交关系理解中的评测方法、数据集以及实验进行相应的介绍.

### 4.1　评测方法介绍

基于图像和视频信息的社交关系理解主要包括 5 项基本评价指标：准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、*F*1 值(*F*1 measure)、均值平均精度(mean Average Precision，*mAP*). 在介绍各个评价指标之前，首先对混淆矩阵进行介绍. 混淆矩阵如表 8 所示.

**表 8　混淆矩阵**

| | | 实际结果 | |
|---|---|---|---|
| | | 1 | 0 |
| 预测结果 | 1 | *TP* | *FP* |
| | 0 | *FN* | *TN* |

其中，T(True)代表正确、F(False)代表错误、P(Positive)代表 1(正样本)、N(Negative)代表 0(负样本). 在我们的任务中，*TP* 表示分类正确的正样本数，*FP* 表示分类错误的正样本数，*TN* 表示分类正确的负样本数，*FN* 表示分类错误的负样本数.

(1) 准确率

所谓的准确率就是在总样本中，预测正确的结果占的比例，计算公式为

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

准确率是一个常用的评价指标，但并不是最佳的评价指标，因为在样本不均衡的时候结果并不准确. 因此，引入了均衡准确率(*Balanced Accuracy*)，计算公式为

$$Balanced\ Accuracy=h\times\frac{TP}{TP+FN}+(1-h)\times\frac{TN}{TN+FP} \tag{2}$$

其中，$h$ 属于$[0,1]$，$h$ 的取值取决于灵敏度和特异度的相对重要性，通常取 1/2.

(2) 精确率

所谓的精确率就是真实样本为正样本占所有预测结果为正样本的比例，计算公式为

$$Precision=\frac{TP}{TP+FP} \tag{3}$$

准确率是对整体的预测评价，而精确率只针对局部进行评价，即对正样本的预测评价.

(3) 召回率

所谓的召回率就是被预测为正样本的数量占真实样本为正样本的比例，计算公式为

$$Recall=\frac{TP}{TP+FN} \tag{4}$$

(4) *F*1 值

$F1$ 值是以精确率和召回率为基础的，两者之间相互影响，存在互补的关系. $F1$ 值不仅考虑精确率，而且还考虑召回率，认为两者同样重要，希望都达到最高值，计算公式为

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{5}$$

上述的 $F1$ 值主要针对二分类问题，但是当遇到多分类问题时就需要用到宏 $F1(F1\_macro)$ 和微 $F1(F1\_micro)$，处理多分类问题时可以将其看成多个二分类问题. $F1\_macro$ 和 $F1\_micro$ 是以 $F1$ 值为基础的，两个指标的计算方法也存在区别. $F1\_macro$ 对每个类别的 $F1$ 值分别进行计算后求平均值. 其中，各个类别的 $F1$ 值权重相同. $F1\_micro$ 是先对 $TP$、$FN$ 和 $FP$ 的整体进行计算，然后计算 $F1$. 两者的计算公式为

$$F1\_macro=\frac{1}{C}\sum_{i-1}^{C}F_1(i) \tag{6}$$

$$F1\_micro=\frac{2\times\sum_{i-1}^{C}TP(i)}{2\times\sum_{i-1}^{C}TP(i)+\sum_{i-1}^{C}FP(i)+\sum_{i-1}^{C}FN(i)} \tag{7}$$

其中，$C$ 为类别数目.

(5) 均值平均精度

所谓的均值平均精度是在多个类别的情况下，对训练出的模型好坏进行评价，计算公式为

$$mAP=\frac{\sum_{i-1}^{c}AP_i}{C} \tag{8}$$

$mAP$ 是计算多个类别 $AP$(Average Precision)的平均，其中 $C$ 为类别数目. $AP$ 的计算公式为

$$AP_g=\frac{\sum Precision_g}{N(Total)_g} \tag{9}$$

其中，$g$ 为所有类别中的一种. 即任意一类 $g$ 的平均精度＝验证集中所有图像在类别 $g$ 中精确度的和/所有图像中包含类 $g$ 的图像数量.

### 4.2 数据集介绍

本节主要针对当前基于图像和视频信息的社交关系理解中的相关数据集进行总结.

#### 4.2.1 图像数据集

图像数据集包括多种类型，例如只有人脸的图像数据集、家庭相册数据集以及包含人物身体和场景内容的图像数据集，具体内容如表 9 所示.

**表 9 图像数据集总结**

| 作者(第一) | 数据类型、名称 | 数据集特点 | 类别 | 评测指标 | 发表期刊、会议 |
|---|---|---|---|---|---|
| Wang[79] | 图像 | 家庭相册图像，包含 276 对人物关系的 6533 个实例. | 母子、父子、祖孙、夫妻、兄弟姐妹 | *Accuracy* | ECCV 2010 |
| Xia[89] | 图像 UBKinFace | 包含 600 张 400 人的图像. | 父子、父女、母子、母女 | *Accuracy* | TMM 2012 |
| Fang[129] | 图像 Family101 | 包含 14816 张图像，206 个核心家庭，607 个公众人物. | 母子、父子、祖孙、夫妻、兄弟姐妹 | *Accuracy* | ICIP 2013 |
| Guo[80] | 图像 Sibling-Face Group-Face | 来自 Flickr 的 200 多对兄弟姐妹人脸图像和 106 个群组图像. | 父子、父女、母子、母女、兄弟姐妹 | *Accuracy* | ICPR 2014 |
| Qin[92] | 图像 TSKinFace | 来自公众人物家庭和图像共享社交网络，787 张面部图像，2589 个人 1015 个群组. | 父子、父女、母子、母女 | *Accuracy* | TMM 2015 |
| Dai[91] | 图像 | 16 个不同家庭的相册. | 孩子、父亲、母亲、祖父、祖母 | *Precision*、*Recall* | WACV 2015 |
| Zhang[96] | 图像 | 从网络和电影中选择的 8306 张图像，分为 8 种关系. | 支配、竞争、信任、温暖、友好、依附、外向、自信 | *Balanced Accuracy* | ICCV 2015 |
| Zhang[130] | 图像 PIPA | 37107 张图像和 63188 个实例以及 2356 个人物身份. | 祖孙、父子、母子、夫妻、兄弟姐妹 | *Accuracy* | CVPR 2015 |
| Li[98] | 图像 PISC | 22670 张图像，包含 9 种社交关系类型和 76568 个带标注的样本. | 无关系、亲密关系、非亲密关系、朋友、家庭、夫妻、职业、商业、无关系 | *mAP* | ICCV 2017 |
| Zhang[99] | 图像 ExpW | 91793 张手动标记的图像. | 愤怒、厌恶、恐惧、高兴、悲伤、惊讶、中性 | *Accuracy*、*Balanced Accuracy* | IJCV 2018 |
| Yoo[131] | 图像 IG | 363 张不同场景图片，2735 个人头位置标注，1439 个组标注. | 握手、拥抱、面对面交谈等 | *Precision*、*Recall*、F1 | PR 2019 |

由表 9 可知，图像数据集包括三种类型：人脸图像数据集[80,89,96,99,129]、家庭相册图像数据集[79,91-92]、包含身体和场景等更多信息的图像数据集[98,130-131]. 其中，人脸图像数据集包含的信息比较少，而且清晰

度比较低.家庭相册图像数据集相对来说清晰度比较高,并且人物之间的位置比较规则,包含的信息相对来说比较丰富.而包含身体和场景的图像数据集信息量比较大,并且图像也特别清晰,非常有利于社交关系理解.

#### 4.2.2　视频数据集

视频数据集大部分是从电影和电视剧中收集的,并且部分数据集中还进行了人工标注.不同数据集包含的类别不相同,所对应的任务也不相同,具体内容如表 10 所示.

**表 10　视频数据集总结**

| 作者(第一) | 数据类型、名称 | 数据集特点 | 类别 | 评测指标 | 发表期刊、会议 |
|---|---|---|---|---|---|
| Ding[110] | 视频 | 收集的 10 部电影数据集,包括近期和经典的戏剧电影. | 动作、冒险、科幻、戏剧等 | *Accuracy*、F1 | ECCV 2010 |
| Lan[132] | 视频 Broadcast Field Hockey Dataset | 从五场比赛中提取的 58 个视频包括 11 个动作类,5 个社交角色类和 3 个场景级别的事件类. | 动作:传球、运球、射门、接球、铲球、准备、站立、慢跑、跑步、行走、扑救<br>社会角色:进攻者、第一防守者、防守者防人、防守者防空挡、其他<br>场景级事件:进攻、任意球、罚球 | *Accuracy* | CVPR 2012 |
| Bojanowski[133] | 视频 | 两部电影,2603 个人,256 164 个人脸标注. | 走路、坐下、开门等动作 | *Accuracy*、*Precision* | ICCV 2013 |
| Barr[114] | 视频 SN-Flip | 通过 28 个人群视频记录了 190 名受试者. | 光照、面部表情、尺度、焦点和姿势 | *Precision*、*Recall*、F1 | WACV 2014 |
| Yan[118] | 视频 KFVW | 包含 418 对面部视频,每个视频大约有 100～500 帧. | 父子、父女、母子、母女 | *Accuracy* | PR 2018 |
| Lv[83] | 视频 SRIV | 从 69 部电影和电视剧中抽取的 3124 段视频,分为主观和客观两大类,每类各包含 8 种关系,共 16 种. | 支配、竞争、信任、温暖、友好、依附、压抑、自信、上下级、同事、服务、亲子、父子、兄弟姐妹、朋友、敌人 | *Accuracy*、*F1_micro*、*F1_macro* | MMM 2018 |
| Vicol[134] | 视频 MovieGraphs | 由 51 部电影分割出的 7637 个视频片段,并且包含一些相应的标注. | 家庭、朋友、工作 | *Accuracy* | CVPR 2018 |
| Liu[5] | 视频 ViSR | 包含 200 多部各种类型的电影,剪辑 8000 多个视频片段. | 领导下属、同事、服务、亲子、兄弟姐妹、夫妻、朋友、竞争 | *Accuracy*、*mAP* | CVPR 2019 |

从表 10 中可以看出,早期的视频数据集不包含人物关系标注[110,114,132-133],最近几年提出的有关社交关系分析的视频数据集对人物关系进行了标注[5,83,118,134],可以更好地促进社交关系理解.

### 4.3　基于图像数据集的实验

为了定量地分析和比较图像中的社交关系理解方法的性能,我们对 PISC 数据集总结了 14 种对比算法,对 PIPA 数据集总结了 4 种对比算法.

#### 4.3.1　各类方法在 PISC 数据集上的比较

(1)数据集介绍

PISC 数据集是 Li 等人[98]在 2017 年提出来的,它是第一个用于社交关系分析的公共数据集.它由 22 670 张图像组成,每张图像的平均人数为 3.11 人,共包含 9 种社交关系类型和 76 568 个带标注的样本.其中,9 种社交关系按照粗粒度划分可以分为 3 类:亲密关系、非亲密关系以及无关系;按照细粒度划分可以分为 6 类:朋友关系、家庭关系、夫妻关系、工作关系、商业关系以及无关系.本节主要针对细粒度的关系进行实验总结.

(2)性能比较

为了验证不同算法在 PISC 数据集上的性能,本文总结了 14 种基线算法和改进算法,算法所使用的评价指标为 *Recall* 和 *mAP*,具体的对比结果如表 11 所示.

**表 11　不同算法在 PISC 数据集上的类别 *Recall* 和 *mAP* 对比**　　(单位:%)

| 方法 | 朋友关系 | 家庭关系 | 夫妻关系 | 工作关系 | 商业关系 | 无关系 | *mAP* |
|---|---|---|---|---|---|---|---|
| Union-CNN[18] | 29.9 | 58.5 | 70.7 | 55.4 | 43.0 | 19.6 | 43.5 |
| BBox[98] | 20.7 | 36.4 | 42.7 | 31.8 | 23.2 | 32.7 | 28.8 |
| Pair-CNN[98] | 30.2 | 59.1 | 69.4 | 57.5 | 41.9 | 34.2 | 48.2 |
| Pair-CNN+BBox[98] | 30.7 | 60.2 | 72.5 | 58.1 | 43.7 | 50.7 | 54.3 |
| Pair-CNN+BBox+Union[98] | 32.5 | 62.1 | 73.9 | 61.4 | 46.0 | 52.1 | 56.9 |
| Pair-CNN+BBox+Global[98] | 32.2 | 61.7 | 72.6 | 60.8 | 44.3 | 51.0 | 54.6 |
| Pair-CNN+BBox+Scene[98] | 30.2 | 59.4 | 71.7 | 57.6 | 43.0 | 49.9 | 51.7 |
| RCNN[98] | 29.7 | 61.9 | 71.2 | 60.1 | 45.9 | 20.7 | 48.4 |
| Dual-Glance[98] | 35.4 | **68.1** | **76.3** | 70.3 | **57.6** | 60.9 | 63.2 |
| GRM[8] | 59.6 | 64.4 | 58.6 | 76.6 | 39.5 | 67.7 | 68.7 |
| MGR[101] | **64.6** | 67.8 | 60.5 | **76.8** | 34.7 | **70.4** | 70.0 |
| MN-CNN module only[100] | — | — | — | — | — | — | 60.2 |
| SRG-GN without Scene[100] | — | — | — | — | — | — | 69.2 |
| SRG-GN[100] | — | — | — | — | — | — | **71.6** |

由表 11 可知，BBox 在所有算法中性能表现最差，而 Union-CNN、Pair-CNN、Pair-CNN＋BBox、Pair-CNN＋BBox＋Union、Pair-CNN＋BBox＋Global、Pair-CNN＋BBox＋Scene、RCNN 等模型在性能上表现相当，但是从 Dual-Glance 模型开始实现了质的飞跃，性能上得到较大的提升. 而 Dual-Glance、MGR 在类别上的性能对比表现出参差不齐的现象，在多个特定类别中表现最佳，在 $mAP$ 指标上的性能逐渐提升，其中 SRG-GN 表现最好.

(3) 实验结果展示

为了使表 11 中的结果更直观地展现出来，我们从中挑选几个主要的方法进行实验结果展示. 实验结果如图 11 所示.

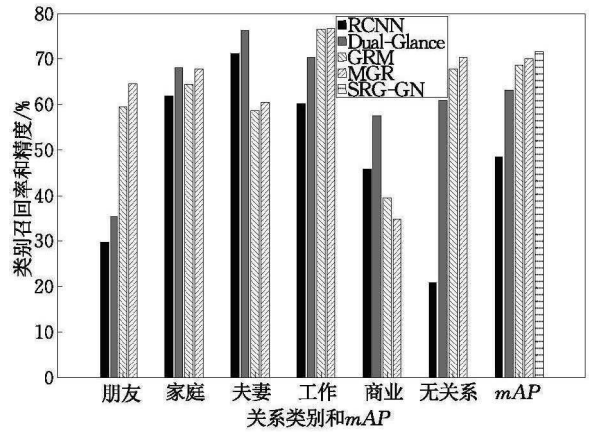


图 11 不同对比算法的实验结果展示

由于 SRG-GN 在六个类别中的评价指标为精确度，而其它算法的评价指标为召回率，所以图中并未展示 SRG-GN 在六个类别中的结果. 我们将表 11 和图 11 相结合对结果进行全面分析.

首先，可以看出 Union-CNN 和 RCNN 在识别“无关系”类别时表现最差，表明这两种方法无法清楚识别无关系的人对，能较好地识别有明显关系的人对. 相比于 Pair-CNN，Pair-CNN＋BBox 表现相对较好，表明图像中人物的位置信息对于社交关系推断产生了作用，在“无关系”类别中表现更为明显. 但是，BBox 方法在除“无关系”类别以外的其它类别上性能表现最差，说明边界框的位置并不能单独用于预测关系. 考虑局部和全局信息后，相比于 Pair-CNN＋BBox 性能得到了一定的提升，而考虑局部信息相比于全局信息来说，性能提升较小. 当考虑场景信息后，性能反而下降，这说明社交关系和场景类型是相互独立的. Dual-Glance 模型相比之前所有的基线方法，表现出最好的性能，$mAP$ 值也得到了较大的提升.

接着，将 Dual-Glance 方法与 GRM 方法进行分析. 相比于前面的方法，Dual-Glance 的效果之所以得到较大的提升，得益于它使用了更精细的局部上下文信息，而不是全局上下文信息. GRM 与 Dual-Glance 不同的是，它使用 GGNN 对相关的语义感知上下文信息进行推理，这些信息更加有利于社交关系理解，实现了更准确的识别. 相比于 Dual-Glance，GRM 将 $mAP$ 值提升了 5.5%，与其它方法相比得到更显著的提升. 原因可能是识别细粒度的社交关系更具挑战性，更加地依赖于先验知识，而 GRM 最大程度地满足了条件.

然后，将 MGR 和 Dual-Glance，GRM 进行对比. Dual-Glance 和 GRM 之所以取得比较好的结果，主要是因为它们既考虑了人物对象，也考虑了上下文语义对象之间的关系，从而反映出在社交关系识别中语义对象信息是一个非常重要的因素. Dual-Glance 在识别家庭、夫妻和商业关系中表现出最好的性能，但对于其它的关系识别表现并不理想，主要是因为使用了语义对象的注意力机制，可以更有效地发现关系较为明显的关键对象. 相反，也可能被不相关的对象误导而导致错误的关系识别. MGR 不仅使用场景级的全局信息以及人与物的外观，而且考虑人和物体之间的交互信息，从而在整体精度上得到最好的表现，在各个关系上也展现出极好的性能.

最后，只对 SRG-GN 与 MGR 的 $mAP$ 值进行分析. 从表 11 中可以看出，SRG-GN 比 MGR 提升了 1.6%，相比于仅使用 MN-CNN 模块提升了 11.4%，但只比不考虑场景的 SRG-GN 提升了 2.4%，说明场景信息对模型的性能影响不大，总体来说 SRG-GN 表现出最好的性能.

4.3.2 各类方法在 PIPA 数据集上的比较

(1) 数据集介绍

PIPA 数据集是 Zhang 等人[130]在 2015 年提出来的，它由 37107 张图像组成，包含 63188 个实例，共 2356 个身份. 接着，Sun 等人[97]对数据集进行标注，得到 134556 个人物标注. 将社会生活划分为 5 个域，16 种社会关系. 其中 5 个域分别为：依附、互惠、配偶、层级、集群，16 种社交关系分别为：父子、母子、祖孙(外公)、祖孙(外婆)、朋友、兄弟姐妹、同学、情侣/夫妻、演讲者-观众、老师-学生、讲师-学员、领导-下属、乐队成员、舞蹈队成员、体育队成员、同事.

(2) 性能比较

为了分析不同算法在 PIPA 数据集上的性能，选取 Zhang 等人[101]对比的 4 种算法，算法使用的评价指标为 *Accuracy*，具体的对比结果如表 12 所示.

**表 12　不同算法在 PIPA 数据集上的识别 *Accuracy* 对比[101]**

| 方法 | *Accuracy* /% |
|---|---|
| Two stream CNN[97] | 57.2 |
| Dual-Glance[98] | 59.6 |
| GRM[8] | 62.3 |
| MGR[101] | **64.4** |

由表 12 可知，Two stream CNN 的准确率最低，而 MGR 的准确率最高，四个模型在准确率上是逐步上升的，证明模型也是在不断地优化，以达到最好的效果. 之所以 Two stream CNN 的准确率最差，主要是因为它所考虑的特征因素相对单一，只考虑人脸、身体外观以及衣服等与人相关的特征，忽略了如全局上下文、局部上下文以及人与物体之间的交互等能够更加准确识别社交关系的影响因素. 而 MGR 相比于 Dual-Glance 和 GRM 表现出最好的性能，主要是因为 MGR 不仅使用场景级的全局信息以及人与物的外观，而且考虑人和物体之间的交互信息.

### 4.4　基于视频数据集的实验

为了定量地分析和比较视频中的社交关系理解方法的性能，我们对 SRIV 数据集总结了 6 种对比算法，对 ViSR 数据集总结了 6 种对比算法.

#### 4.4.1　各类方法在 SRIV 数据集上的比较

(1) 数据集介绍

SRIV 数据集是 Lv 等人[83]在 2018 年提出来的，它是从 69 部电影和电视剧中截取的人物交互数据集，共 3124 段视频，大约 25 个小时，每段视频的长度范围在 3 s～90 s 之间. SRIV 数据集的社交关系划分为主观和客观两大类，每类分别包括 8 种子类. 主观关系分为支配、竞争、信任、温暖、友好、依附、压抑、自信，客观关系分为上下级、同级、服务、亲子、夫妻、兄弟姐妹、友好、敌对.

(2) 性能比较

为了验证不同算法在 SRIV 数据集上的性能，本文总结了 6 种算法，算法所使用的评价指标为 *Accuracy*、$Sub_{accuracy}$ 和 *F*1. 其中，$Sub_{accuracy}$ 表示一个样本的预测标签集合与标准标签集合的完全匹配度. 具体的对比结果如表 13 所示.

**表 13　不同算法在 SRIV 数据集上的识别 *Accuracy* 和 *F*1 对比**

| 分类 | 方法 | *Accuracy* /% | $Sub_{accuracy}$ /% | *F*1 /% |
|---|---|---|---|---|
| 主观关系 | C3D[135] | 55.65 | 3.47 | 38.86 |
| | LSTM[136] | 66.67 | 27.97 | 57.76 |
| | TSN[137] | 70.89 | 34.82 | 61.42 |
| | Multi-Stream[83] | 74.36 | 52.13 | 66.83 |
| | STMV[82] | 75.35 | 52.49 | 67.95 |
| | TSM[84] | **82.74** | **59.36** | **74.12** |
| 客观关系 | C3D[135] | 55.68 | 14.51 | 30.18 |
| | LSTM[136] | 61.47 | 37.92 | 41.93 |
| | TSN[137] | 54.12 | 30.45 | 48.94 |
| | Multi-Stream[83] | 61.36 | 52.91 | 63.83 |
| | STMV[82] | 63.22 | 53.11 | 64.92 |
| | TSM[84] | **71.25** | **60.32** | **72.93** |

由表 13 可知，C3D 方法在所有算法中性能表现最差，而 LSTM 和 TSN 不管是在准确率还是 *F*1 值上表现相差不大. Multi-Stream、STMV 和 TSM 相比于前三种方法表现出更好的性能. 其中，TSM 在所有方法中性能表现最佳.

(3) 实验结果展示

为了使表 13 中的结果更直观地展现出来，我们进行了实验结果展示. 实验结果如图 12 所示.

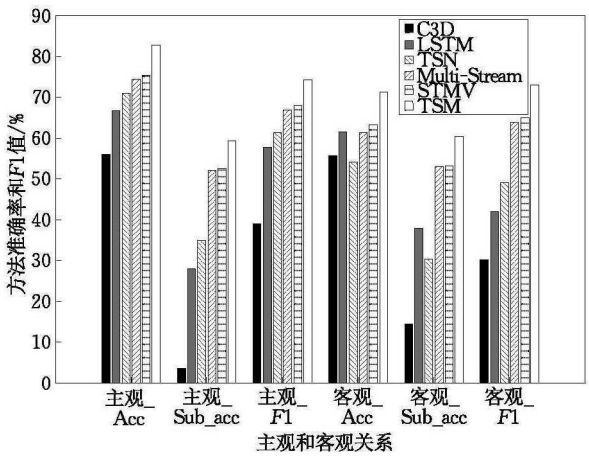


图 12　不同对比算法的实验结果展示

结合表 13 和图 12 可知，C3D、LSTM 和 TSN 三个模型的性能表现比较差. 虽然这些方法可以很好地描述视频的其它特征，但是很难提取到有效的社交关系表征. Multi-Stream 之所以较前三种方法有了明显的提升，主要是考虑到将对社交关系识别有用的空间特征(RGB 图像)、时间特征(光流图像)和声音特征进行融合. STMV 较之前的方法将准确率提高 1.3%～35.4%，主要是利用多视角特征从不同角度反映人物交互，并且使用了更能反映人物关系特征的多个注意力单元. TSM 在所有算法中表现最好，分别将两个任务的准确率提高了 9.8% 和

12.7%.其中,Multi-Stream 和 STMV 只关注了视频的时空特征,在性能上逊色于 TSM.TSM 不仅使用了注意力机制,而且考虑了物体特征、人和物体之间的交互信息.因为这些信息对于提高社交关系的准确率更加有效.

4.4.2 各类方法在 ViSR 数据集上的比较

(1)数据集介绍

ViSR 数据集是 Liu 等人[5]在 2019 年提出来的,它是基于 Lv 等人[83]在 2018 年提出的 SRIV 数据集基础上扩充出来分析人物社交关系的视频数据集.包含从 200 多部电影中剪辑出的 8000 多个人物交互视频片段,大多数剪辑长度被限制在 30 s 内.ViSR 数据集的社交关系划分为 8 个类别,分别为领导下属、同事、服务、亲子、兄弟姐妹、夫妻、朋友、竞争.

(2)性能比较

为了验证不同算法在 ViSR 数据集上的性能,本文选取 Liu 等人[5]对比的 6 种算法,算法所使用的评价指标为 Top-1 *Accuracy* 和 *mAP*.其中,Top-1 *Accuracy* 是选择每个类别中准确率最高的一类作为最终的关系类别.具体的对比结果如表 14 所示.

**表 14 不同算法在 ViSR 数据集上的识别 *Accuracy* 和 *mAP* 对比[5]** (单位:%)

| 方法 | Top-1 *Accuracy* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 领导下属 | 同事 | 服务 | 亲子 | 兄弟姐妹 | 夫妻 | 朋友 | 竞争 | *mAP* |
| GRM[8] | 48.67 | 48.67 | 6.67 | 0 | 0 | 4.17 | 0.67 | 30.13 | 16.69 |
| TSN-Spatial[83] | 55.48 | 42.93 | 30.00 | 35.20 | 34.83 | **39.78** | 48.75 | 37.07 | 42.38 |
| TSN-ST[83] | 41.05 | 33.33 | 30.00 | 32.83 | **45.78** | 29.17 | 63.76 | 32.87 | 43.23 |
| GCN[5] | 56.16 | 49.46 | 27.14 | 36.80 | 41.57 | 34.41 | 39.80 | **50.00** | 43.46 |
| PGCN[5] | 54.11 | **54.89** | 25.71 | 40.80 | 34.83 | 33.33 | 45.27 | 48.28 | 44.73 |
| MSTR[5] | **57.53** | 51.09 | **30.00** | **45.60** | 39.33 | 38.71 | **53.23** | 47.41 | **47.75** |

由表 14 可知,GRM 模型在所有算法中性能表现最差,而在各个类别的准确率对比中,TSN-Spatial、TSN-ST、GCN、PGCN、MSTR 都在某个或某几个类别上表现出最好的性能.MSTR 在类别性能表现最好的结果中占比最高.并且在 *mAP* 评价指标中表现最佳.

(3)实验结果展示

为了使表 14 中的结果更直观的展现出来,我们进行了实验结果展示.实验结果如图 13 所示.

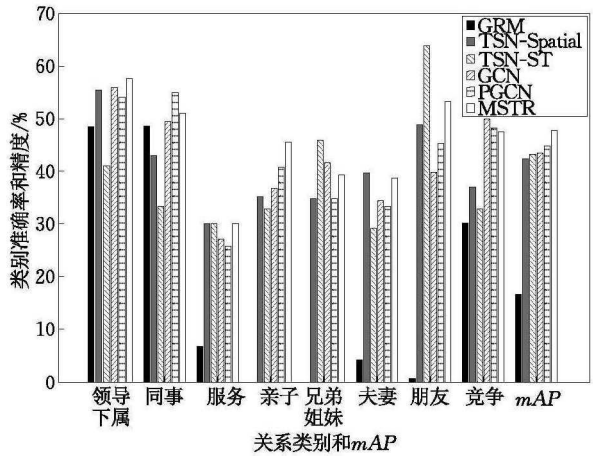


图 13 不同对比算法的实验结果展示

图 13 中的 TSN-Spatial 方法本身是用于多标签分类,由于对比的需要,将其修改成单标签分类,仅输入视频的 RGB 帧,使用 TSN 提取空间特征.TSN-ST 仅输入视频的光流,使用 TSN 提取时间和空间特征.此外,将图像实验中的 GRM 方法应用到视频数据上.GCN 为标准 GCN 方法,而 PGCN 和 MSTR 是 Liu 等人[5]提出的模型一部分和完整的金字塔时间推理框架.

结合表 14 和图 13 可知,GRM 在所有的对比算法中表现最差,可能的原因是基于图像的方法并不适用于视频数据,视频数据相比于图像数据更加的复杂.因为基于图像的方法要求一幅图像中要同时存在两个或者多个人,而视频的每一帧中可能只存在一个人,导致不能满足基于图像方法的条件.TSN-Spatial 和 TSN-ST 主要关注的是全局特征,而基于图的 GCN 和 PGCN 方法主要关注的是局部细粒度特征.通过比较发现,不管是全局特征还是局部特征都能影响性能的提升.但从整体来看,人和物体的细节外观以及行为等局部细粒度特征对性能的影响更大,更有利于社交关系识别.最后,结合 TSN 和 PGCN 的 MSTR 模型表现出最好的性能,说明融合全局特征和局部细粒度特征更有助于性能的提升,体现出互补效应.

## 5 问题与挑战

随着技术的快速发展,产生了海量的多媒体数据,尤其是图像和视频数据,研究人员在此基础上对社交关系理解进行相关的研究.尽管基于图像和视

频信息的社交关系理解研究已经取得了一定的成果，但仍然面临着巨大的问题与挑战，具体如下：

(1) 小样本训练与学习

为了在基于图像和视频信息的社交关系识别或者网络构建等任务中取得较高的准确率，一般的做法是基于大量的标注数据进行模型训练. 然而，数据的收集与标注需要耗费大量的人力物力财力. 以视频中的社交关系识别为例，为了标注人物之间的关系类型，需要标注人员对每一个视频片段进行观看，而人物之间的关系种类复杂多样，人工标注非常耗时耗力. 如果再加上其它更多的特征标注，这样人工是很难处理的. 因此，如何基于小样本进行训练与学习是一个亟待解决的问题[1]. 例如，如何通过学习方法，基于小样本的标注数据自动生成大量样本的准确标注，或者基于少量的标注数据进行模型训练，更准确地预测社交关系.

(2) 多源异构数据的关系融合

多媒体的数据类型种类繁多，例如文本、语音、图像和视频等. 基于单类别数据的社交关系网络构建已经进行了广泛的研究，但是融合多源异构数据构建的网络结构依然是当前存在的问题与挑战. 因此，如何实现跨越不同媒体数据的关系融合成为基于多媒体信息的社交关系理解的重要问题.

(3) 无监督条件下的社交关系理解

当前基于图像和视频信息的社交关系理解主要是利用人工标注的标签进行训练. 但是由于数据量过大，借助人工进行标注既费时又费力. 因此，为了使无监督的社交关系理解成为可能，就需要尽可能地摆脱对标注数据的依赖. 因为任务和需求不同，无法在监督条件下提前训练好，所以如何实现无监督条件下的社交关系理解，并将其应用到实际生活中是一个严峻的挑战.

(4) 多角色不同关系识别

现有基于视频信息的社交关系理解研究主要是针对两个角色之间的关系以及多角色之间相同关系的识别，而对于多角色之间不同关系识别的研究仍然存在一定的挑战. 例如，在视频中存在多个角色，对于人类来说很容易判断出两两角色之间的关系类型. 但是，对于机器来说，识别多角色之间不同关系与人类理解之间仍存在较大差距，如何准确识别出符合实际情况的多角色之间不同关系是一个亟待解决的问题.

(5) 高效的社交关系理解模型算法

近年来，随着深度学习的快速发展，已经在基于图像、音频和视频等媒体数据研究中广泛应用并取得了不错的效果. 例如，将卷积神经网络应用于特征提取、目标检测等计算机视觉任务中. 然而，为了获得更高的准确率以达到实际应用，模型被设计的越来越复杂. 由于模型的复杂性和参数的增加，导致训练时间急剧增长，以时间的代价换取准确率的提高. 因此，如何设计出更加高效的社交关系理解模型算法是未来研究中需要着重考虑的问题.

(6) 实时的数据反馈

随着深度学习的快速发展与应用，基于图像和视频信息的社交关系理解取得了较大进展，但是缺乏实时性，难以实时地处理现实中复杂的情况. 如何根据数据的不断变化实时准确地反馈相应的结果是一项极具挑战的任务. 此外，随着视频数据的大量出现，基于视频信息的社交关系分析将具有重要的研究和应用价值. 相比于图像，视频中的信息更加丰富，包含更多有价值的信息，同时也包含了影响模型性能的干扰信息. 如何排除干扰，充分整合有用信息进行实时的结果反馈，是一个极具挑战的问题. 以监控系统为例，当出现两个角色时，如何结合有用信息并且排除周围的干扰信息来实时地反馈出两个人物之间的关系[138]，从而为人物追踪、公共安全与反恐等提供有力的线索.

(7) 多媒体知识图谱

知识图谱是一种结构化的形式，既包含客观世界中的实体对象，也包含他们之间的复杂关系. 它有助于对海量信息进行组织、管理和理解，在搜索引擎、认知推理、问答系统等服务中得到广泛的应用. 而知识图谱的构建大多是基于文本数据，如何从多种不同来源的媒体数据中抽取知识并进行融合，从而构建知识图谱实现综合推理和动态更新是亟待解决的问题.

## 6 总　结

基于多媒体信息的社交关系理解之所以成为一个研究热点并且受到学术界和工业界的广泛关注，得益于图像、视频、音频等多媒体数据的海量出现以及具有极高的研究价值. 本文主要对近年来基于图像和视频信息的社交关系理解的分类和研究现状进行总结，整体分为六个部分：引言、问题描述、基于图像和视频信息的社交关系理解、实验、问题与挑战、总结. 其中，对基于图像和视频信息的社交关系理解面临的问题与挑战作了详细的阐述，包括小样本训

练与学习、多源异构数据的关系融合、无监督条件下的社交关系理解、多角色不同关系识别、高效的社交关系理解模型算法、实时的数据反馈、多媒体知识图谱.本文旨在为感兴趣的研究人员提供有价值的参考,促进该领域的进一步发展.

## 参 考 文 献


[1] Peng Yu-Xin, Qi Jin-Wei, Huang Xin. Current research status and prospects on multimedia content understanding. Journal of Computer Research & Development, 2019, 56(1): 187-212(in Chinese)
(彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望. 计算机研究与发展, 2019, 56(1): 187-212)

[2] Lv Jin-Na. Research on Key Technologies of Social Relationship Extraction of Video Characters [Ph. D. dissertation]. Beijing University of Posts and Telecommunications, Beijing, 2019(in Chinese)
(吕金娜. 视频人物社交关系抽取的关键技术研究[博士学位论文]. 北京邮电大学, 北京, 2019)

[3] Li Zhen-Jun, Dai Qiang-Qiang, Li Rong-Hua, et al. Social relationship mining algorithm by multi-dimensional graph structural clustering. Journal of Software, 2018, 29(3): 839-852(in Chinese)
(李振军, 代强强, 李荣华等. 多维图结构聚类的社交关系挖掘算法. 软件学报, 2018, 29(3): 839-852)

[4] Mueller S, Wagner J, Smith J, et al. The interplay of personality and functional health in old and very old age: Dynamic within-person interrelations zcross up to 13 years. Journal of Personality and Social Psychology, 2017, 115(6): 1127-1147

[5] Liu X, Liu W, Zhang M, et al. Social relation recognition from videos via multi-scale spatial-temporal reasoning// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3566-3574

[6] Andriluka M, Iqbal U, Insafutdinov E, et al. PoseTrack: A benchmark for human pose estimation and tracking// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5167-5176

[7] Vignesh R, Jonathan H, Abu-El-Haija S, et al. Detecting events and key actors in multi-person videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3043-3053

[8] Wang Z, Chen T, Ren J, et al. Deep reasoning with knowledge graph for social relationship understanding// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 1021-1028

[9] Peng Yu-Xin, Huang Xin, Zhao Yun-Zhen. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(9): 2372-2385

[10] Peng Yu-Xin, Zhu Wen-Wu, Zhao Yao, et al. Cross-media analysis and reasoning: Advances and directions. Frontiers of Information Technology and Electronic Engineering, 2017, 18(1): 44-57

[11] Li W, Hong F, Zheng W. Learning to learn relation for important people detection in still images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5003-5011

[12] Chen Z, Wei X, Wang P, et al. Multi-label image recognition with graph convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5177-5186

[13] Xia S, Shao M, Fu Y. Toward kinship verification using visual attributes//Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba, Japan, 2012: 549-552

[14] Wang J, Jiao J, Bao L, et al. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4006-4015

[15] Zeng Nian-Yin, Zhang Hong, Song Bao-Ye, et al. Facial expression recognition via learning deep sparse autoencoders. Neurocomputing, 2018, 273: 643-649

[16] Zhang H, Zawlin K, Shih-Fu C, et al. Visual translation embedding network for visual relation detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5532-5540

[17] Hamdi D, Albert A S, Theo G. Like father, like son: Facial expression dynamics for kinship verification//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 1497-1504

[18] Lu C, Ranjay K, Michael B, et al. Visual relationship detection with language priors//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 852-869

[19] Zhang C, Peng Y. Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 1135-1141

[20] Han G, Zhang X, Li C. Semi-supervised DFF: Decoupling detection and feature flow for video object detectors// Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea, 2018: 1811-1819

[21] Rao Yong-Ming, Lu Ji-Wen, Zhou Jie. Learning discriminative aggregation network for video-based face recognition and person re-identification. International Journal of Computer Vision, 2019, 127(6-7): 701-718

[22] Li Dong, Yao Ting, Duan Ling-Yu, et al. Unified spatio-temporal attention networks for action recognition in videos. IEEE Transactions on Multimedia, 2019, 21(2): 416-428

[23] Stefano A, Giuseppe S, Simone C, et al. Understanding social relationships in egocentric vision. Pattern Recognition, 2015, 48(12): 4082-4096

[24] Zhao Zhong-Qiu, Zheng Peng, Xu Shou-Tao, et al. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232

[25] Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3476-3483

[26] Javier M, David V, Antonio M L, et al. Occlusion handling via random subspace classifiers for human detection. IEEE Transactions on Cybernetics, 2013, 44(3): 342-354

[27] Huang Kai-Qi, Ren Wei-Qiang, Tan Tie-Niu. A review on image object classification and detection. Chinese Journal of Computers, 2014, 36(6): 1225-1240(in Chinese)
(黄凯奇，任伟强，谭铁牛. 图像物体分类与检测算法综述. 计算机学报, 2014, 36(6): 1225-1240)

[28] Douze M, Jegou H, Schmid C. An image-based approach to video copy detection with spatio-temporal post-filtering. IEEE Transactions on Multimedia, 2010, 12(4): 257-266

[29] Chong-Wah N, Ma Yu-Fei, Zhang Hong-Jiang. Video summarization and scene detection by graph modeling. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(2): 296-305

[30] Ji Bai-Yang, Chen Chun, Qian Ying. The development of video segmentation technology. Journal of Computer Research & Development, 2001, 38(1): 36-42(in Chinese)
(季白杨，陈纯，钱英. 视频分割技术的发展. 计算机研究与发展, 2001, 38(1): 36-42)

[31] Qian Gang, Zeng Gui-Hua. Comparison of representative methods of video shot segmentation. Computer Engineering and Applications, 2004, 32: 51-55+84(in Chinese)
(钱刚，曾贵华. 典型视频镜头分割方法的比较. 计算机工程与应用, 2004, 32: 51-55+84)

[32] Salembier P, Marques F. Region-based representations of image and video: Segmentation tools for multimedia services. IEEE Transactions on Circuits and Systems for Video Technology, 1999, 9(8): 1147-1169

[33] Nagasnlm A, Tanaka Y. Automatic video indexing and full-motion search for object appearance//Proceedings of the 2nd Working Conference on Visual Databases Systems. 1992: 113-127

[34] Ramin Z, Justin M, Kevin M. A feature-based algorithm for detecting and classifying scene breaks//Proceedings of the ACM International Conference on Multimedia. San Francisco, USA, 1995: 189-200

[35] Yang Rui-Qin, Lv Jin-Lai. Video shot segmentation method based on dual-detection. Computer Engineering and Design, 2018, 39(5): 201-206(in Chinese)
(杨瑞琴，吕进来. 基于双重检测的视频镜头分割方法. 计算机工程与设计, 2018, 39(5): 201-206)

[36] Ji Xiang, Wang Yuan, Tong Xiao-Min, et al. Video shot segmentation algorithm based on GIST feature and conditional decision. Journal of CAEIT, 2018, 13(2): 55-59(in Chinese)
(吉祥，王源，仝小敏等. 基于GIST特征和条件判定的视频镜头分割算法. 中国电子科学研究院学报, 2018, 13(2): 55-59)

[37] Lv Jin-Na, Wu Bin, Zhou Li-Li, et al. StoryRoleNet: Social network construction of role relationship in video. IEEE Access, 2018, 6: 25958-25969

[38] Zhuang Y, Rui Y, Thomas S, et al. Adaptive key frame extraction using unsupervised clustering//Proceedings of the 1998 International Conference on Image Processing. Chicago, USA, 1998: 866-870

[39] Liu Tian-Ming, Zhang Hong-Jiang, Qi Fei-Hu. A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(10): 1006-1013

[40] Chasanis V, Likas A, Nikolaos P G. Scene detection in videos using shot clustering and sequence alignment. IEEE Transactions on Multimedia, 2009, 11(1): 89-100

[41] Qi X, Liu C, Stephanie S. CNN based key frame extraction for face in video recognition//Proceedings of the 4th International Conference on Identity, Security, and Behavior Analysis. Singapore, 2018: 1-8

[42] Jia Y, Evan S, Jeff D, et al. Caffe: Convolutional architecture for fast feature embedding//Proceedings of the ACM International Conference on Multimedia. Florida, USA, 2014: 675-678

[43] Alex K, Ilya S, Geoffrey E H. ImageNet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

[44] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1302-1310

[45] Hamdi D. Visual transformation aided contrastive learning for video-based kinship verification//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2478-2487

[46] Chen X, Ma H, Wan J, et al. Multi-view 3D object detection network for autonomous driving//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6526-6534

[47] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580-587

[48] Girshick R. Fast R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440-1448

[49] Ren S, IIe K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[50] Dai J, Li Y, IIe K, et al. R-FCN: Object detection via region-based fully convolutional networks//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 379-387

[51] Lan T, Piotr D, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IIonolulu, USA, 2017: 2117-2125

[52] IIe K, Gkioxari G, Dollar P, et al. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969

[53] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6154-6162

[54] Bae S. Object detection based on region decomposition and assembly//Proceedings of the Association for the Advancement of Artificial Intelligence. IIonolulu, USA, 2019: 21-35

[55] Kim D, Lee G, Jeong J, et al. Tell me what they're holding: Weakly-supervised object detection with transferable knowledge from human-object interaction//Proceedings of the Association for the Advancement of Artificial Intelligence. New York, USA, 2020: 11246-11253

[56] Erhan D, Szegedy C, Toshev A, et al. Scalable object detection using deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2147-2154

[57] Yoo D, Park S, Lee J, et al. AttentionNet: Aggregating weak directions for accurate object detection//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2659-2667

[58] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[59] Najibi M, Rastegari M, Davis L. G-CNN: An iterative grid based object detector//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2369-2377

[60] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 21-37

[61] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IIonolulu, USA, 2017: 7263-7271

[62] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018

[63] Zhao Q, Sheng T, Wang Y, et al. M2Det: A single-shot object detector based on multi-level feature pyramid network //Proceedings of the Association for the Advancement of Artificial Intelligence. IIonolulu, USA, 2019: 9259-9266

[64] Kim S, Park S, Na B, et al. Spiking-YOLO: Spiking neural network for energy-efficient object detection//Proceedings of the Association for the Advancement of Artificial Intelligence. New York, USA, 2020: 11270-11277

[65] Peter L S, Green J D, Dimitrov B. IIigh level color similarity retrieval. International Journal of Information Theories and Applications, 2003, 10(3): 363-369

[66] Zhang Deng-Sheng, Islam M M, Lu Guo-Jun. A review on automatic image annotation techniques. Pattern Recognition, 2012, 45(1): 346-362

[67] Zhang Deng-Sheng, Lu Guo-Jun. Review of shape representation and description techniques. Pattern Recognition, 2004, 37(1): 1-19

[68] Wang W, Wang R, Shan S, et al. Exploring context and visual pattern of relationship for scene graph generation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8188-8197

[69] Zhou II, Zhang C, IIu C. Visual relationship detection with relative location mining//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 30-38

[70] Zhou Xiang-Dong. Video Feature Extraction and Video Shot Analysis [M.S. dissertation]. National University of Defense Technology, Changsha, 2002(in Chinese)
(周祥东. 视频特征提取和视频镜头分割[硕士学位论文]. 国防科技大学, 长沙, 2002)

[71] Qian X, Zhuang Y, Li Y, et al. Video relation detection with spatio-temporal graph//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 84-93

[72] Sun X, Ren T, Zi Y, et al. Video visual relation detection via multi-modal feature fusion//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 2657-2661

[73] Zheng S, Chen X, Chen S, et al. Relation understanding in videos//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 2662-2666

[74] Guo Jie, Song Bin, Zhang Peng, et al. Affective video content analysis based on multimodal data fusion in heterogeneous networks. Information Fusion, 2019, 51: 224-232

[75] Tsai Y, Divvala S, Morency L, et al. Video relationship reasoning using gated spatio-temporal energy graph//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10424-10433

[76] Xu Yuan, Xue Xiang-Yang. A method of regional high level feature extraction on video. Computer Science, 2006, 33(11): 134-138(in Chinese)
(许源, 薛向阳. 一种视频局部高层语义特征提取算法. 计算机科学, 2006, 33(11): 134-138)

[77] Zeng Zhi-Hong, Maja P, Glenn I R, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(1): 39-58

[78] Chen Y, Winston H, Liao H. Discovering informative social subgraphs and predicting pairwise relationships from group photos//Proceedings of the 20th ACM International Conference on Multimedia. Nara, Japan, 2012: 669-678

[79] Wang G, Andrew G, Luo J, et al. Seeing people in social context: Recognizing people and social relationships//Proceedings of the European Conference on Computer Vision. Crete, Greece, 2010: 169-182

[80] Guo Y, Hamdi D, Laurens V. Graph-based kinship recognition //Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden, 2014: 4287-4292

[81] Aimar E S, Radeva ES, Dimiccoli M. Social relation recognition in egocentric photostreams//Proceedings of the IEEE International Conference on Image Processing. Taipei, China, 2019: 3227-3231

[82] Lv J, Wu B. Spatio-temporal attention model based on multi-view for social relation understanding//Proceedings of the International Conference on Multimedia Modeling. Thessaloniki, Greece, 2019: 390-401

[83] Lv J, Liu W, Zhou L, et al. Multi-stream fusion model for social relation recognition from videos//Proceedings of the International Conference on Multimedia Modeling. Bangkok, Thailand, 2018: 355-368

[84] Dai P, Lv J, Wu B. Two-stage model for social relationship understanding from video//Proceedings of the IEEE International Conference on Multimedia and Expo. Shanghai, China, 2019: 1132-1137

[85] Zhou J, Cui G, Zhang Z, et al. Graph neural networks: A review of methods and applications. arXiv preprint arXiv: 1812. 08434, 2018

[86] Golder S A. Measuring social networks with digital photograph collections//Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia. Pittsburgh, USA, 2008: 43-48

[87] Wu P, Ding W, Mao Z, et al. Close & closer: Discover social relationship from photo collections//Proceedings of the IEEE International Conference on Multimedia and Expo. Cancun, Mexico, 2009: 1652-1655

[88] Wu P, Dan T. Close & closer: Social cluster and closeness from photo collections//Proceedings of the 17th ACM International Conference on Multimedia. Beijing, China, 2009: 709-712

[89] Xia Si-Yu, Shao Ming, Luo Jie-Bo, et al. Understanding kin relationships in a photo. IEEE Transactions on Multimedia, 2012, 14(4): 1046-1056

[90] Dehghan A, Ortiz E G, Villegas R, et al. Who do i look like? Determining parent-offspring resemblance via gated autoencoders//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1757-1764

[91] Dai Q, Peter C, Leonid S, et al. Family member identification from photo collections//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2015: 982-989

[92] Qin Xiao-Qian, Tan Xiao-Yang, Chen Song-Can. Tri-subject kinship verification: Understanding the core of a family. IEEE Transactions on Multimedia, 2015, 17(10): 1855-1867

[93] Joseph P, Shao Ming, Wu Yue, et al. Visual kinship recognition of families in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2624-2637

[94] Zhou Xiu-Zhuang, Jin Kai, Xu Min, et al. Learning deep compact similarity metric for kinship verification from face images. Information Fusion, 2019, 48: 84-94

[95] Yan Hai-Bin, Wang Shi-Wei. Learning part-aware attention networks for kinship verification. Pattern Recognition Letters, 2019, 128: 169-175

[96] Zhang Z, Luo P, Chen C, et al. Learning social relation traits from face images//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3631-3639

[97] Sun Q, Bernt S, Mario F. A domain based approach to social relation recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3481-3490

[98] Li J, Wong Y, Zhao Q, et al. Dual-glance model for deciphering social relationships//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2669-2678

[99] Zhang Zhan-Peng, Luo Ping, Chen C, et al. From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision, 2018, 126(5): 550-569

[100] Goel A, Ma K, Tan C. An end-to-end network for generating social relationship graphs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11186-11195

[101] Zhang M, Liu X, Liu W, et al. Multi-granularity reasoning for social relation recognition from images//Proceedings of the IEEE International Conference on Multimedia and Expo. Shanghai, China, 2019: 1618-1623

[102] Amato F, Castiglione A, Mercorio F, et al. Multimedia story creation on social networks. Future Generation Computer Systems, 2018, 86: 412-420

[103] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[104] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75

[105] Athanasios V, Nikolaos D, Anastasios D, et al. Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience, 2018, 2018: 1-13

[106] Yuan Ming-Kuan, Peng Yu-Xin. Bridge-GAN: Interpretable representation learning for text-to-image synthesis. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(11): 4258-4268

[107] Frank D, Chrysochou P, Mitkidis P, et al. Moral machines: The challenges of raising driverless cars. Science, 2018, 352(6293): 1573

[108] Chen Tian-Shui, Lin Liang. Visual expression learning and application of knowledge reasoning. Communications of the CCF, 2019, 15(11): 48-57(in Chinese)
(陈添水, 林倞. 融合知识推理的视觉表达学习及应用. 中国计算机学会通讯, 2019, 15(11): 48-57)

[109] Weng Chung-Yi, Chu Wei-Ta, Wu Ja-Ling. RoleNet: Movie analysis from the perspective of social networks. IEEE Transactions on Multimedia, 2009, 11(2): 256-271

[110] Ding L, Yilmaz A. Learning relations among movie characters: A social network perspective//Proceedings of the European Conference on Computer Vision. Crete, Greece, 2010: 410-423

[111] Yuan K, Yao H, Ji R, et al. Mining actor correlations with hierarchical concurrence parsing//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, USA, 2010: 798-801

[112] Yeh M, Tseng M, Wu W. Automatic social network construction from movies using film-editing cues//Proceedings of the IEEE International Conference on Multimedia and Expo Workshops. Melbourne, Australia, 2012: 242-247

[113] Chiu Y, Chung P, Huang C. Character relationship analysis in movies using face tracks//Proceedings of the MVA 2013 IAPR International Conference on Machine Vision Applications. Kyoto, Japan, 2013: 431-434

[114] Barr J R, Cament L A, Bowyer K W, et al. Active clustering with ensembles for social structure extraction//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Steamboat Springs, USA, 2014: 969-976

[115] Nan C, Kim K, Zhang B. Social network analysis of TV drama characters via deep concept hierarchies//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris, France, 2015: 831-836

[116] Zhou L, Lv J, Wu B. Social network construction of the role relation in unstructured data based on multi-view// Proceedings of the IEEE Second International Conference on Data Science in Cyberspace. Shenzhen, China, 2017: 382-388

[117] Zhou L, Wu B, Lv J. SRE-Net model for automatic social relation extraction from video//Proceedings of the CCF Conference on Big Data. Xi'an, China, 2018: 442-460

[118] Yan Hai-Bin, Hu Jun-Lin. Video-based kinship verification using distance metric learning. Pattern Recognition, 2018, 75: 15-24

[119] Kohli N, Yadav D, Vatsa M, et al. Supervised mixed norm autoencoder for kinship verification in unconstrained videos. IEEE Transactions on Image Processing, 2018, 28(3): 1329-1341

[120] Wu X, Eric G, Tomi H, et al. Audio-visual kinship verification in the wild//Proceedings of the 12th IAPR International Conference on Biometrics. Crete, Greece, 2019: 1-12

[121] Ding L, Yilmaz A. Inferring social relations from visual concepts//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 699-706

[122] Ramanathan V, Yao B, Li F. Social role discovery in human events//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2475-2482

[123] Kukleva A, Tapaswi M, Laptev I. Learning interactions and relationships between movie characters//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA, 2020: 9849-9858

[124] Lin Liang, Wang Xiao-Long, Yang Wei, et al. Discriminatively trained and-or graph models for object shape detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(5): 959-972

[125] Pedro F F, Daniel P H. Efficient graph-based image segmentation. International Journal of Computer Vision, 2004, 59(2): 167-181

[126] Liu W, Jiang Y, Luo J, et al. Noise resistant graph ranking for improved web image search//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 849-856

[127] Liu Xin-Chen, Liu Wu, Mei Tao, et al. PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia, 2017, 20(3): 645-658

[128] Michael D, Xavier B, Pierre V. Convolutional neural networks on graphs with fast localized spectral filtering// Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 3844-3852

[129] Fang R, Andrew C, Chen T, et al. Kinship classification by modeling facial feature heredity//Proceedings of the IEEE International Conference on Image Processing. Melbourne, Australia, 2013: 2983-2987

[130] Zhang N, Manohar P, Yaniv T, et al. Beyond frontal faces: Improving person recognition using multiple cues// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4804-4813

[131] Haanju Y, Taekyu E, Jeongmin S, Choi S I. Detection of interacting groups based on geometric and social relations between individuals in an image. Pattern Recognition, 2019, 93: 498-506

[132] Lan T, Leonid S, Greg M. Social roles in hierarchical models for human activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Rhode Island, USA, 2012: 1354-1361

[133] Bojanowski P, Bach F, Laptev I, et al. Finding actors and actions in movies//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 2280-2287

[134] Paul V, Makarand T, Lluis C, et al. MovieGraphs: Towards understanding human-centric situations from videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8581-8590

[135] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497

[136] Nicholas V F. Short note on a heuristic search strategy in long-term memory networks. Information Processing Letters, 1972, 1(5): 191-196

[137] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition //Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 20-36

[138] Wang Da-Ling, Yu Ge, Feng Shi, et al. Research on modeling entities and their relations for social media search. Chinese Journal of Computers, 2016, 39(4): 657-674(in Chinese)
(王大玲，于戈，冯时等. 面向社会媒体搜索的实体关系建模研究综述. 计算机学报, 2016, 39(4): 657-674)

**WANG Zheng**, Ph. D. candidate. His research interests include multimedia content understanding, machine learning and computer vision.

**WU Bin**, Ph. D., professor. His research interests include data mining, complex network and cloud computing.

**WANG Wen-Zhe**, M. S. candidate. His research interests include multimedia content understanding and machine learning.

**TENG Yi-Yang**, Ph. D. candidate. Her research interests include multimedia content understanding, machine learning and computer vision.

**SHUAI Jie**, Ph. D. candidate. His research interests include data mining and recommender system.

**XIAO Yun-Peng**, Ph. D., professor. His research interests include social networks and machine learning.

**BAI Ting**, Ph. D., lecturer. Her research interests include recommender systems, multidimensional data mining and network representation learning.

**Background**

As a research problem in the field of computer vision, multimedia social relation understanding aims to help people quickly understand video content, network formation mechanism and knowledge graph construction. The research is an interdisciplinary academic field that includes sociology, mathematics, computer science, psychology, biology, etc. However, previous studies mainly used textual data to study social relationships. Recent years, with the rapid development of network and multimedia technology, a large amount of image and video data have emerged. It is because of the emergence of such information-rich data that we can study the understanding of social relation based on images and videos information.

This paper provides a summary of previous research works. From the perspective of image and video, this paper summarizes two aspects of multimedia social relation understanding: the existence of relationships and the types of existing relationships (single and multiple relationships). First, we define the problem and introduce the process of understanding multimedia social relation. Second, we analyze the content from two perspectives of image and video. Then, datasets and comparative experiments are introduced. Finally, the problems and challenges in this field are summarized.

This paper is supported by the National Key Research and Development Program of China under Grant No. 2018YFC0831500, the National Natural Science Foundation of China under Grant No. 61972047, the NSFC-General Technology Basic Research Joint Funds under Grant No. U1936220, and the Fundamental Research Funds for the Central Universities under Grant No. 500420824. This paper aims to provide a beneficial reference for interested researchers to understand the research status of multimedia social relation understanding more comprehensively, and to promote the further development of this field.