计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800160

基于多特征融合的细粒度视频人物关系抽取

吕金娜　邢春玉　李　莉
北京信息科技大学信息管理学院　北京 100192

摘　要　视频人物关系抽取是信息抽取问题中的重要任务,在视频描述、视频检索,以及人物搜索、公安监察等方面具有重要价值。由于视频数据的底层像素与高层关系语义之间存在巨大的鸿沟,现有方法很难准确地抽取人物关系。现有研究大多通过粗粒度地分析人物共现等因素来抽取人物关系,忽略了具有丰富语义的视频中的细粒度信息。为解决现有算法难以准确、完整地抽取视频人物关系的问题,文中提出了一种基于多特征融合的细粒度视频人物关系抽取方法。首先,为了准确识别视频人物实体,提出了一种基于多特征融合的人物实体识别模型;然后,提出了一种基于细粒度特征的人物关系识别模型,该模型不仅融合了视频中人物的时空特征,而且考虑了与人物相关的细粒度物体信息特征,从而建立更好的映射关系来准确识别人物关系。以电影视频数据和 SRIV 人物关系识别数据集为实验数据,实验结果验证了该模型的有效性和准确性,与现有同类模型相比,所提模型的人物实体识别 F_1 值提高了约 14.4%,人物关系识别的准确率提高了约 10.1%。

关键词:视频分析;人物关系识别;关系抽取;深度学习;多特征融合
中图法分类号　TP391

Video Character Relation Extraction Based on Multi-feature Fusion and Fine-granularity Analysis

LYU Jin-na, XING Chun-yu and LI Li
School of Information Management, Beijing Information Science & Technology University, Beijing 100192, China

Abstract　Video character relation extraction is an important task of information extraction. It is valuable for video description, video retrieval, character search, public security supervision, etc. Due to the huge gap between the underlying pixels of video data and the semantics of high-level relation, it is difficult to accurately extract the relations. Most existing studies are based on coarse-granularity analysis, such as co-occurrence of characters, which ignores the fine-granularity information. In order to solve the problem that it is difficult to accurately and completely extract the relations among video characters, this paper proposes a new method for extracting relations of video characters based on multi-feature fusion and fine-granularity analysis. First, a new character entity recognition model, named CRMF (Character Recognition based on Multi-feature Fusion), is proposed. Through this manner, we can generate a more complete character set using face and body features fusion. Second, we exploit a character relationship recognition model based on fine-granularity features, named FGAG (Fine-Granularity Analysis based on GCN), which not only fuses the spatio-temporal features, but also considers the fine-granularity objects information related to the characters. Thus a better mapping can be established to accurately identify the character relations. Comprehensive evaluations are conducted on the movie video and SRIV character relationship recognition dataset, and the experimental results demonstrate that the proposed method outperforms the state-of-the-art methods on character entity and relation recognition, F_1 value increases by 14.4% and accuracy increases by 10.1%.

Keywords　Video analysis, Character relation recognition, Relation extraction, Deep learning, Multi-feature fusion

1　引言

随着计算机网络和多媒体技术的发展,视频数据数量剧增,这些视频包含丰富的内容,并且多数是以人为中心的。人物作为视频的重要实体,同时也是社会进步中最重要的角色,人们在社会生产和社会生活的直接交往中形成了人与人之间的关系,研究视频中人物关系的抽取对视频内容的理解、知识发现、角色识别等具有重要的社会意义。

视频人物关系抽取指给定一段或多段视频,研究如何利用模型或方法从视频中提取人物的关系网络图。

近年来,基于图像的实体关系研究已取得一定的研究成果[1-2],但是这些研究方法不能直接适用于具有时空变换的视频数据中。Vicol 等[3]手动构建了 MovieGraphs 数据集,抽取视频中的实体关系,为视频实体的分析提供了一定的参考。

到稿日期:2020-08-25　返修日期:2020-10-16
基金项目:北京信息科技大学校科研基金(2035012)
This work was supported by the Beijing University of Information Technology Foundation(2035012).
通信作者:吕金娜(lvjinna@bistu.edu.cn)

然而，随着海量视频的涌现，如何利用人工智能等技术从视频大数据中自动抽取有价值的人物关系是当前研究面临的挑战。

人物实体是构建人物关系网络的基本元素，如何高效、准确地识别人物实体是关系抽取的重要基础，其抽取的完整性、准确率等将直接影响关系知识库的质量。近年来，已有一些抽取视频人物关系的研究，这些研究根据人物共现时间长短[4]、共现时间远近[5]简单地抽取人物关系。这些方法中的人物实体识别都是基于人脸识别的方法，依赖于清晰的正面人脸特征，但是视频中的人物往往存在侧面，甚至背面的情况，导致现有方法在人物识别上存在不完整、不准确的现象，因此还需要研究新的方法来提高人物实体识别的准确率。

人物实体关系识别是视频人物关系抽取的核心部分，其主要任务是判定视频人物之间是何种关系。由于视频的复杂时空变换，使得关系识别研究仍存在巨大的挑战。研究人员对基于静态图像的人物关系识别进行了一些研究[6-9]，这些研究利用机器学习或深度学习算法来分析图像中人的友好、敌对等社交关系类型，利用人物的表情、年龄、动作来识别人物之间的交互关系，研究分析的图像数据具有清晰的人脸、位置关系、表情等。近年来基于视频的人物分析也取得了一些研究成果[10-12]。Bojanowski 等[10]从电影中分析人物动作，包括走路、坐下、开门等。Yan 等[12]针对视频中人物的亲属关系进行分析，包括父子、父女、母子、母女 4 种亲属关系类型。由于视频底层像素和高层关系语义之间存在巨大鸿沟，人物之间的交互关系难以刻画，因此人物关系的识别准确率还有待提高。

为了解决上述问题，本文提出了基于多特征融合的细粒度视频人物关系抽取方法。首先，本文提出了一种多特征融合的人物实体识别(Character Recognition based on Multi-feature Fusion，CRMF)模型，用于构建视频中的人物实体集合。该模型融合了人物的人脸和身体特征，通过聚类模型自动识别视频中的人物实体个数，自动标注与构建人物实体集合。其次，本文提出了一种时空特征融合的细粒度视频物体特征的人物关系的识别模型(Fine-Granularity Analysis based on GCN，FGAG)，该模型利用卷积网络(Convolutional Neural Networks，CNN)和长短期记忆网络(Long Short-Term Memory，LSTM)提取视频的时空特征，并从细粒度角度出发，融合利用图卷积网络提取的视频中与人物交互相关的物体实体特征，从而提高人物关系识别的准确率。

本文的主要贡献如下：1)提出了一种多特征融合的人物实体识别模型 CRMF，融合人脸、身体和场景特征，并利用层次聚类模型自动构建人物实体集合，解决了人物实体识别不准确和遗漏的问题；2)提出了一种利用时空特征融合的细粒度视频物体特征的人物关系识别模型 FGAG，该模型利用图卷积网络提取与人物交互相关的物体实体信息，从细粒度角度进行分析，提高了人物关系的识别准确率；3)在标准数据上集和真实数据进行实验，并与多种现有方法进行对比，验证了本文模型的有效性。

2 相关工作

本文方法主要涉及视频人物关系网络抽取中的人物实体识别和人物关系识别，因此本文主要从这两方面讨论相关工作。

2.1 视频人物关系网络抽取与人物实体识别

人物关系网络抽取是将人物实体作为网络的节点，根据一定的规则判定人物节点之间是否存在连边，从而形成人物关系网络的点边结构。最早的研究是基于结构化数据[13]，随着自然语言处理技术的不断成熟，基于文本数据的人物关系识别的研究不断涌现出来[14]。

从视频中抽取人物关系网络的研究较少，现有研究大多基于电影和电视剧视频[6]，以及角色共现抽取关系网络。Ding 等[15]分析人物是否出现在同一视频场景，从而抽取视频中的人物关系网络。Tran 等[16]统计人物在视频画面中的共现抽取人物关系。这些方法中的人物实体识别大多是基于人脸的分析，首先检测出人脸图片，然后基于人脸图片集合进行聚类或分类，并对结果进行评价。基于分类的方法需要标记一部分人脸标签，这样会大大增加人工工作量，因此有待研究智能化程度更高和更准确的视频人物实体识别方法。

近年来，关于人脸的研究有很多高质量的方法[17-19]。Zhang 等[17]利用 3 个 CNN 级联的方式实现了 coarse-to-fine 的算法结构，提高了人脸检测的精度和速度。关于行人重识别的研究成为了热点，Bartoli 等[18]提出利用场景中人的位置和尺度的高斯混合模型，该模型可以用无监督的方式从少量的视频帧中提取特征并检测行人。Wang 等[19]将单图像表示和交叉图像表示统一到一个 CNN 模型中，联合优化后得到精确的特征表示以进行行人重识别。这些研究为视频人物实体的识别研究提供了参考。

2.2 视频人物关系识别

现有研究大多抽取人物关系网络为点边的图结构，即人物作为节点，节点之间存在连边即为存在关系。现实生活中人与人之间存在着各种各样的社会关系[1]，因此识别人物之间存在何种关系有利于更进一步的知识发现。

大量研究致力于通过自然语言处理技术构建文本语料库中实体之间的关系。Wang[20]利用深度学习的方法从微博中抽取人物关系。Kang 等[21]分析微信聊天记录的时间信息，利用分层聚类和 k 均值聚类的方法对人物的关系进行抽取和分类。这些方法通过分析文本的语义来抽取人物关系，不适用于处理视频数据。基于图像数据的人物关系分析大多是通过提取人物性别、年龄、姿势和表情等特征来识别人物关系。这些研究大多分析亲子关系[22]和相册图片中的人物。Huang 等[23]利用图片中人物的社交关系，以及人物与事件的关系特征，来进一步挖掘准确的人物关系。Li 等[13]构建了 PISC 数据库，包括朋友、家庭、夫妻、工作、商业关系，提出了一种基于注意力机制的双流网络模型，通过同时学习人物和空间语义特征来推断社会关系。

近年来，已有一些基于视频数据的实体关系的研究。这些研究大多利用人脸的特征来分析人物之间的亲子关系。Ertugrul 等[24]提出了一种亲情综合框架，用于合成预测儿童的表情视频。López 等[25]利用人脸特征来分析视频中人物的血缘关系。Yan 等[26]提出了一种用于人脸亲属关系验证的弱监督特征学习方法。目前，对视频中人物社交关系分析的

研究较少。Bibi 等[27]通过分析不同摄像头产生的视频人物的动作和轨迹信息来识别人物之间的交互关系。LYU 等[28]提出了一种基于视频的人物社交关系预测方法，该方法利用视频的多角度特征来学习视频的高层语义，以预测视频中人物的关系属性。Wang 等[29]分析视频的时序特征，捕获实体对象之间的长距离依赖关系以及与人相互作用的时空关系相近的物体。

视频的时序性、复杂性、人物交互的空间不连续性，以及底层像素与高层关系语义之间的巨大鸿沟，使得视频人物关系识别问题仍面临巨大的挑战。

3 视频人物关系抽取

3.1 问题定义

视频中人物关系网络的抽取步骤如下：首先，提取出视频中的人物作为网络节点集合；其次，分析视频中人物的交互特征，从而推断节点间是否存在连边；最后，判定边的属性，即人物之间的关系属于何种关系类型。

人物节点集合表示为 C，视频人物作为关系网络的节点，通过视频人物的识别和标注形成人物节点集合，表示为 $C=\{c_1,c_2,\cdots,c_n\}$。人物关系网络 G 表示为 $G=\langle C,E,R\rangle$，其中 C 为人物节点集合，E 为人物存在关系的连边集合，R 为人物之间的关系类型。

本文重点研究两个问题：1）如何从视频 V 中构建准确的视频人物节点集合 C，即 $V->C$，利用模型从长视频中识别人物集合，并进行人物标注；2）如何准确识别人物关系的类型，即 $V->R$，利用算法或模型分析视频的特征和人物的特征来推断人物之间的关系是何种类型。

3.2 多特征融合的人物实体识别

视频中人物实体识别不同于文本，视频中人物表示的像素特征比文本的语义更加抽象。现有的人脸检测与识别算法大多基于正脸，少部分可以处理侧脸，但是在现实的视频数据中人物出现的情况多种多样，有正脸、侧脸，甚至是背面。因此，如何准确地检测和识别人物是课题研究的一个难题。

本文提出了一种多特征融合的人物实体识别模型，如图 1 所示。

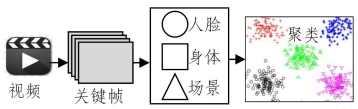

图 1 多特征融合的人物识别模型

Fig. 1 Character recognition model based on multi-feature fusion

首先，从长视频中识别人物实体，需要对长视频进行预处理，本文对视频 V 进行镜头和场景检测，每个镜头提取一帧作为关键帧，每个场景利用一组关键帧序列，表示为 $I=\{I_i\mid i=1,2,\cdots,L\}$，$L$ 表示视频帧的数目。然后，利用 ArcFace 方法[30]识别人脸，所有人脸图片利用 MTCNN 方法归一化为 112×112 像素大小，并提取特征。其次，利用 Aligned-ReID[31]进行人的识别，每个人的身体通过 Mask R-CNN 检测归一化为 128×256 像素大小，并提取特征。场景特征利用 Resnet101 网络提取。对以上 3 种特征进行融合，得到融合后的特征 $fAttr$，并将其作为聚类算法下一步的输入。

视频人物关系抽取的关键是识别视频中的人物。已有研究通常是基于分类的监督训练方法，需要标注人物的标签，因此耗费了很多人力，而基于简单聚类的方法需要指定聚类个数，因此这些方法在研究基于广义视频人物的关系抽取时存在局限性。CRMF 模型通过二次聚类的无监督方法来得到视频中人物节点的集合。该方法利用两次聚类的无监督学习将人物进行聚类，不需要指定人物的个数，同时避免了大量的人工标注。因此，该方法能够处理广义视频数据，是一种高效、通用的视频人物节点集合的识别方法。人物实体识别方法的具体流程如算法 1 所示。

算法 1 人物实体识别聚类算法

输入：人物特征 fAttr

输出：人物实体集合 C

1. 初始化聚类个数 k=0；
2. 为第一次聚类初始化聚类中心 clu_list[k]；
3. FOR f 属于 fAttr(1：) do

找到距离 f 最近的聚类中心，并记录该聚类中心的索引 index 和最小距离 min_dis；

IF 最小距离 min_dis<t do：

将 f 添加到 clu_list[index]；

ELSE：

k=k+1；

用 f 初始化新建的聚类中心 clu_list[k]；

END IF

END FOR

4. Centroids，cluster Assent=kmeans(fAttr，k+1)；
5. 为每一个聚类中心设置标签值 labels，值为 0～k；
6. 将聚类中心与 labels 按列合并，并将结果保存到 C 中。

利用人物特征提取方法提取人物的特征 $fAttr$ 并将其作为聚类方法的输入。第一次聚类得到聚类的个数，即人物的数目；第二次聚类对每个节点进行标注，以识别视频中人物对应的标签。其中，相似性计算用余弦距离来计算。

$$D_{ij}^{c}=\frac{\langle c_i,c_j\rangle}{\|c_i\|\cdot\|c_j\|}=\frac{\sum_{n=1}^{N}f_{in}\cdot f_{jn}}{\sqrt{\sum_{n=1}^{N}f_{in}^2}\cdot\sqrt{\sum_{n=1}^{N}f_{jn}^2}}$$

其中，N 表示特征的维度，$\langle c_i,c_j\rangle$ 表示 c_i 和 c_j 的内积，$\|c_i\|$ 表示 c_i 的范数，f_{in} 是 c_i 的第 n 维的值。

利用人物实体识别模型可以获得视频的人物实体集合 $C=\{c_1,c_2,\cdots,c_n\}$，本文采用人物场景共现的方法构建视频人物关系网络 $G=\langle C,E,R\rangle$。假设视频 V 用场景表示为 $V=\{s_1,s_2,\cdots,s_m\}$，因此具体描述为：

$$G=\langle C,E,R\rangle\begin{cases}\langle c_i,c_j\rangle\in s_k\Rightarrow E_{\langle c_i,c_j\rangle}=1\\ \langle c_i,c_j\rangle\notin s_k\Rightarrow E_{<c_i,c_j>}=0\\ R=0\end{cases}$$

其中，s_k 表示第 k 个场景，$\langle c_i,c_j\rangle\in s_k$ 表示人物 c_i 和 c_j 在同一场景中。

3.3 基于细粒度特征的人物关系识别模型

视频人物之间关系的抽象性与视频的像素视觉特征之间存在巨大的鸿沟，仅仅依靠人物共现场景图片的特征很难准确识别人物关系，而细粒度物体信息将有助于人物关系的分

析，如办公桌、电脑和文档资料等细粒度信息可以体现人物之间的关系可能为同事关系。本文提出了一种细粒度图卷积人物关系识别模型，模型框图如图 2 所示。FGAG 模型不仅考虑了视频的时空特征，而且融合了基于 GCN 网络分析的与人物交互相关的物体实体特征，两种特种融合后进行人物关系分类。

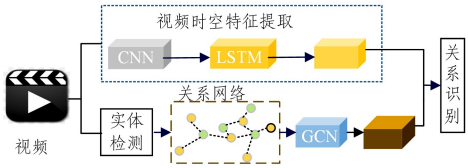

图 2 人物关系识别模型框图

Fig. 2 Framework of character relation recognition model

视频 V 经过预处理后，视频帧序列表示为 $I=\{I_i \mid i=1,2,\cdots,L\}$，$L$ 表示视频帧的数目，帧图像大小归一化为 224×224 像素大小，然后利用 CNN 和 LSTM 网络提取时空特征，得到的视频序列的时空特征表示为 $F_{st}=\{f_{st}^i \mid i=1,2,\cdots,L\}$。

物体实体检测采用 R-CNN 在 COCO 数据集预先训练的模型。输入视频帧为 I_i，检测到的物体域表示为 $P_i=\{P_i^j \mid j=1,2,\cdots,M\}$，其中 M 表示检测到的物体种类数。物体实体特征表示为 $F_o=\{f_o^j \mid j=1,2,\cdots,M\}$，然后建立视频中物体实体元素的关系图，并将其表示为邻接矩阵的形式 $\boldsymbol{A}_o \in R^{N\times N}$，判定邻接矩阵值的定义如下：

$$\boldsymbol{A}_o(o_a,o_b)\begin{cases}1, & a \text{ 和 } b \text{ 位于同一图片中}\\ 0, & \text{其他情况}\end{cases}$$

接下来，利用图卷积网络学习视频中实体元素之间的关系表示，依据人物关系标签训练模型，提取实体关系图的高层语义特征。给定 N 个节点的图，每个节点有 d 维度的特征向量，一层图卷积的操作定义如下：

$$X^{(l+1)}=\sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}X^{(l)}W^{(l)})$$

其中，$\widetilde{\boldsymbol{A}}$为图的邻接矩阵表示，$\widetilde{\boldsymbol{D}}$为矩阵$\widetilde{\boldsymbol{A}}$的度矩阵，$X^{(l)}$ 为第 $L-1$ 层的输出，$W^{(L)} \in \mathbb{R}^{d\times d'}$ 为卷积网络学习的参数矩阵，$\sigma(\cdot)$为非线性激活函数 ReLU。

GCN 网络的输出为更新后的节点特征，表示为 $X^{(L)}$。最后，融合时空语义特征和细粒度实体特征来进行人物关系的预测。损失函数采用交叉熵损失函数。

$$loss=-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}[l_k\log(p_k)+(1-l_k)\log(1-p_k)]$$

4 实验

4.1 数据集

实验数据包括两部分，即关系网络构建和人物实体识别。实验采用电影《不二情书》和《茜茜公主》的视频数据，视频和音频编码格式为 H.264 和 AC3，视频比特率为 7916 kB/s，音频比特率为 320 kB/s。

人物识别关系的数据集使用的是 SRIV 数据集[11]。它是一个视频人物关系视频数据集，包括 3124 段视频，时长约 25 h，本文训练采用客观关系分类及其子类别，共 8 类关系，如表 1 所列。

表 1 人物关系分类

Table 1 Classification of character relations

关系类别	类别	编号
工作关系	上下级关系	1
	同级关系	2
	服务关系	3
亲属关系	亲子关系	4
	夫妻关系	5
	兄弟姐妹	6
其他关系	朋友关系	7
	敌对关系	8

4.2 评价指标

实验结果利用 F1 值、准确率 Acc 和集合准确率 Sub_{acc} 来评价算法，定义如下：

$$P_i=\sum_i^m TP_i/(\sum_i^m TP_i+\sum_i^m FP_i)$$

$$R_i=\sum_i^m TP_i/(\sum_i^m TP_i+\sum_i^m FN_i)$$

$$F1=2\times P_i\times R_i/(P_i+R_i)$$

其中，TP_i 表示第 i 个类别的真正类的样本个数，FP_i 表示第 i 个类别的假正类的样本个数，P_i 为第 i 个类别的准确率，R_i 表示第 i 个类别的召回率，FN_i 表示第 i 个样本的假负类的个数。

$$Acc=\frac{1}{2}(TP/N_P+TN/N_n)$$

其中，N_P 和 N_n 分别为正负样本数。

$$sub_{acc}(h)=\frac{1}{n}\sum_{k=1}^{n}I(h(x_k)=Y_k)$$

其中，x_k 为第 k 类的预测值，Y_k 为第 k 类的真实值。

4.3 结果分析

4.3.1 人物实体识别结果分析

通过电影《茜茜公主》和《不二情书》的视频来分析人物实体识别的准确率，并进行对比评价来验证多特征融合的人物实体识别方法的有效性。

两部电影共检测到人物图片 1324 张，其中人脸图片 390 张。表 2 列出了不同方法的人物识别结果的 $F1$ 值、准确率和召回率的对比结果。

表 2 人物识别结果

Table 2 Results of character recognition

（单位：%）

方法	召回率 R	准确率 P	$F1$
FaceNet[32]	76.4	87.3	71.3
Body	34.5	83.2	49.2
CRMF(Ours)	83.9	90.5	89.1

从表 2 可以看出，本文方法的 3 个评价指标均高于其他方法，人脸识别 $F1$ 值达到了 89.1%。其中，$F1$ 值比 FaceNet 和 Body 方法分别提高了 16.7% 和 38.4%，这说明融合人脸和身体的图片可以更加准确地识别人物实体。从表中还可以看出，单使用人脸来识别人物时准确率达到了 87.3%，但是召回率却只有 76.4%。Body 方法的召回率低于其他方法，这是因为电影中人物身体显示不完整，多数情况下只有半身，或者只显示头部。融合人脸和身体特征之后，召回率提高了

7.5%。因此，实验结果证明融合多种特征的 CRMF 方法是有效的。

为了验证多特征融合的人物实体识别模型对提高人物关系网络的准确性的影响，实验对比了引入该模型前后的人物关系抽取结果，如表 3 所列。实验利用了文献[28]中的基于故事分割的人物关系抽取方法。从表 3 可以看出，CRMF 模型的引入，使得两部电影数据上的 $F1$ 值分别提高了 2.6%和 1.2%，其中《不二情书》的提高值低于《茜茜公主》，原因在于其视频中的人物关系复杂度低于后者。

表 3 人物关系网络抽取的 $F1$ 值的对比结果

Table 3 $F1$ values of relationship network extraction

方法	茜茜公主	不二情书
PlotNet	0.1670	0.3529
RoleNet[5]	0.1481	0.2795
ICASSP10[33]	0.2105	0.7451
StoryRoleNet[28]	0.6947	0.8727
StoryRoleNet(+CRMF)	**0.7132**	**0.8831**

4.3.2 人物关系识别结果分析

本文人物关系数据集 SRIV 上进行实验分析，本文方法与其他方法的结果的对比情况如表 4 所列。

表 4 人物关系识别结果对比

Table 4 Results comparison of character relation recognition

方法	$F1$	Acc	Subacc
C3D[34]	0.3886	0.0556	0.0347
LSTM[35]	0.5776	0.6667	0.2797
TSN[36]	0.6142	0.7089	0.3482
Multi-Stream[11]	0.6383	0.6136	0.5291
STMV[37]	0.6492	0.6322	0.5311
Ours(no CRMF)	0.6714	0.6526	0.5243
Ours(no FGAG)	0.7185	0.6824	0.5302
Ours	**0.7432**	**0.7013**	**0.5382**

从表 4 可以看出，本文方法取得了最好的效果，$F1$ 值达到 0.7432，Acc 提高了 10.1%。这是由于采用基于图卷积网络的细粒度分析模型可以捕捉到更有利于人物关系分析的细粒度线索。从表 4 可以看出，C3D 模型的 Acc 值较低，可能是因为该模型虽然对段视频的分类效果较好，但是不适用于人物交互的细粒度分析。

为了验证 CRMF 和 FGAG 模型的有效性，分别用人脸特征的人物实体识别(no CRMF)和未融合细粒度特征的时空特征(no FGAG)模型来替换这两部分。从表 4 可以看出，多特征融合比单使用人脸特征的人物实体识别得到了更高的 $F1$ 值，为 0.6714，说明融合身体和场景的特征可以更准确地识别人物实体。另外，使用 FGAG 模型比单使用时空特征模型(no FGAG)时 Acc 指标提高得比较明显，提高了 7.9%，说明细粒度信息在考虑人物时空关系的同时还提取了与人物共现的事物细粒度特征，可以提高人物关系识别准确率。

实验运行环境和参数设置为 Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz 处理器，GPU 处理器为 TITAN X(Pascal)，操作系统为 Ubuntu 16.04。模型训练采用随机梯度下降优化，损失函数如图 3 所示，设置 batch_size 为 32，当迭代次数为 200 时达到最优值。

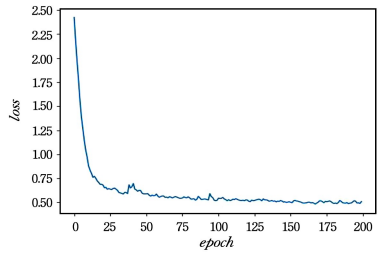

图 3 FGAG 模型损失函数变化图

Fig. 3 Loss function graph of FGAG model

4.3.3 结果可视化

实验从电影视频《茜茜公主》和《不二情书》中抽取人物关系网络的可视化结果，如图 4 所示。节点为人物 C，在图 4 中用圆形人脸图片表示，不同人物关系用直线连线表示，边上的编号为关系类型(见表 1)。

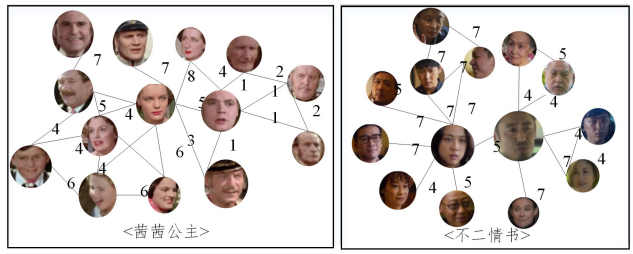

图 4 人物关系抽取结果可视化

Fig. 4 Visualization of character relationship extraction

结束语 为了解决视频人物关系网络抽取中人物实体识别和关系识别不准确的问题，本文提出了一种基于多特征融合的细粒度视频人物关系抽取方法。首先，本文提出了一种多特征融合的人物实体识别模型，其不仅使用人脸特征，而且检测人的身体并提取特征进行融合，从而提高了人物实体识别准确率。其次，本文提出了一种时空特征融合的细粒度视频物体特征的人物关系识别模型，不其仅提取视频的时空特征，而且从细粒度角度提取与人物交互相关的物体实体的特性，从而提高人物关系识别准确率。实验结果证明，与对比方法相比，本文方法在 $F1$ 值和准确率指标上都达到了最优值。

参 考 文 献

[1] ZHANG Z, LUO P, LOY C C, et al. Learning Social Relation Traits from Face Images[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3631-3639.

[2] LI J, WONG Y, ZHAO Q, et al. Dual-Glance Model for Deciphering Social Relationships[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2669-2678.

[3] VICOL P, TAPASWI M, CASTREJON L, et al. MovieGraphs: Towards Understanding Human-Centric Situations from Videos [C]// Proceedings of the CVPR. 2018: 8581-8590.

[4] TRAN Q D, JUNG J E. CoCharNet: Extracting Social Networks using Character Co-occurrence in Movies [J]. Journal of Universal Computer Science, 2015, 21(6): 796-815.

[5] WENG C, CHU W, WU J. RoleNet: Movie Analysis from the Perspective of Social Networks[J]. IEEE Transactions on Multimedia, 2009, 11(2): 256-271.

[6] YUAN K, YAO H, JI R, et al. Mining actor correlations with hierarchical concurrence parsing[C]// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Dallas, Texas, USA, 2010: 798-801.

[7] WANG G, GALLAGHER A C, LUO J, et al. Seeing People in Social Context: Recognizing People and Social Relationships [C]// Proceedings of the Computer Vision- ECCV 2010-11th European Conference on Computer Vision. Crete, Greece, 2010: 169-182.

[8] DAI Q, CARR P, SIGAL L, et al. Family Member Identification from Photo Collections[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA, 2015: 982-989.

[9] SAPRU A, BOURLARD H. Automatic Recognition of Emergent Social Roles in Small Group Interactions [J]. IEEE Trans. Multimedia, 2015, 17(5): 746-760.

[10] BOJANOWSKI P, BACH F R, LAPTEV I, et al. Finding Actors and Actions in Movies[C]// Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 2280-2287.

[11] LV J, LIU W, ZHOU L, et al. Multi-stream Fusion Model for Social Relation Recognition from Videos[C]// MultiMedia Modeling 24th International Conference. Bangkok, Thailand, 2018: 355-368.

[12] YAN H, HU J. Video-based kinship verification using distance metric learning [J]. Pattern Recognition, 2018, 75: 15-24.

[13] HE X M, CHN Y D, LI D. A Construction for Social Network on the Basis of Project Cooperation[J]. Journal of Computer Research and Development, 2016, 53(4): 776-784.

[14] MIKA P. Flink: Semantic Web technology for the extraction and analysis of social networks [J]. SSRN Electronic Journal, 2005, 3(2/3): 211-223.

[15] DING L, YILMAZ A. Learning Relations among Movie Characters: A Social Network Perspective[C]// Computer Vision - ECCV 2010, 11th European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 410-423.

[16] TRAN Q D, JUNG J E. CoCharNet: Extracting Social Networks using Character Co-occurrence in Movies [J]. Journal of Universal Computer, 2015, 21(6): 796-815.

[17] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[18] BARTOLI F, LISANTI G, KARAMAN S, et al. Scene-dependent proposals for efficient person detection [J]. Pattern Recognition, 2019, 87: 170-178.

[19] WANG F, ZUO W, LIN L, et al. Joint learning of single-image and cross-image representations for person re-identification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1288-1296.

[20] WANG C. Research and implementation of Chinese microblog character relationship map based on deep learning [D]. Wuhan: Wuhan Institute of Posts and Telecommunications, 2018.

[21] KANG Y R, ZHAO L, FAN W, et al. Digital Profiling: Relationships Analysis Based on Time Information of WeChat [J]. Journal of Criminal Technique, 2018, 43(3): 187-192.

[22] QIN X, TAN X, CHEN S. Tri-Subject Kinship Verification: Understanding the Core of a Family[J]. IEEE Trans on Multimedia, 2015, 17(10): 1855-1867.

[23] HUANG Q, XIONG Y, LIN D. Unifying Identification and Context Learning for Person Recognition[C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018: 2217-2225.

[24] ERTUGRUL I O, JENI L A. Modeling and synthesis of kinship patterns of facial expressions [J]. Image Vision Computer, 2018, 79: 133-143.

[25] LÓPEZ M B, HADID A, BOUTELLAA E, et al. Kinship verification from facial images and videos: human versus machine [J]. Mach. Vis. Appl., 2018, 29(5): 873-890.

[26] YAN H. Learning discriminative compact binary face descriptor for kinship verification [J]. Pattern Recognition Letters, 2019, 117: 146-152.

[27] BIBI S, ANJUM N, SHER M. Automated multi-feature human interaction recognition in complex environment [J]. Computers in Industry, 2018, 99: 282-293.

[28] LV J, WU B, ZHOU L, et al. StoryRoleNet: Social Network Construction of Role Relationship in Video [J]. IEEE Access, 2018, 6: 25958-25969.

[29] WANG X, GUPTA A. Videos as Space-Time Region Graphs [C]// Proceedings of the Computer Vision - ECCV 2018 -15th European Conference. Munich, Germany, 2018: 413-431.

[30] JDENG J K, GUO J, ZAFEIRIOU S F. Arcface: Additive angular margin loss for deep face recognition [J]. arXiv: 1801.07698, 2018.

[31] ZHANG X, LUO H, FAN X, et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification [J]. arXiv: 1711.08184, 2017.

[32] FLORIAN S, DMITRY K, JAMES P. FaceNet: A unified embedding for face recognition and clustering[C]// Proceedings of the CVPR 2015. Boston, Massachusetts, 2015: 815-823.

[33] YUAN K, YAO H. Mining actor correlations with hierarchical concurrence parsing[C]// IEEE ICASSP. 2010: 798-801.

[34] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 4489-4497.

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

[36] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]// Proceedings of the ECCV 2016. Amsterdam, Netherlands, 2016: 20-36.

[37] LYU J N, WU B. Spatio-Temporal Attention Model Based on Multi-view for Social Relation Understanding [C]// International Conference on Multimedia Modeling. Springer, Cham, 2019: 390-401.

LYU Jin-Na, born in 1981, Ph. D, lecture, is a member of China Computer Federation. Her main research interests include multimedia content analysis, social network analysis and so on.