(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112347965 A
(43)申请公布日 2021.02.09

(21)申请号 202011280036.1

(22)申请日 2020.11.16

(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号
申请人 同盾控股有限公司

(72)发明人 庄越挺 肖俊 汤斯亮 吴飞 杨易 李晓林 谭炽烈 蒋韬

(74)专利代理机构 杭州求是专利事务所有限公司 33200
代理人 郑海峰

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)

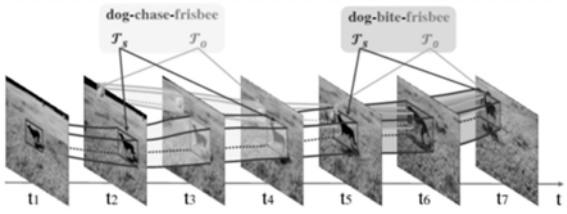
权利要求书3页 说明书9页 附图3页

(54)发明名称

一种基于时空图的视频关系检测方法和系统

(57)摘要

本发明公开了一种基于时空图的视频关系检测方法和系统。首先,将视频中的实体集合和它们之间的关系建模为一个全连接的时空图,该图包括时间和空间维度邻域中的实体节点。对于关系检测,本发明提出了一种视频关系检测图卷积网络模型(VRD-GCN),用于从上下文中聚合信息并在此时空图中进行推理。一方面,VRD-GCN通过捕捉实体在时空维度上的几何和外观的相对变化来检测实体之间的动态关系。另一方面,通过将时空图中邻域中的节点和上下文的消息传递给目标实体,VRD-GCN能够产生更准确和完整的检测结果。在检测到视频片段中的关系实例后,用一种使用孪生网络的在线关联方法将整个视频中的短期关系实例合并。该方法对于视频中的关系检测准确率高。



CN 112347965 A

1.一种基于时空图的视频关系检测方法,其特征在于包括如下步骤:

1)获取视频片段的帧级别的实体特征和视频片段的实体轨迹特征;

2)将前一个片段和当前片段的实体特征和实体轨迹特征,以及当前片段和下一个片段的实体特征和实体轨迹特征分别拼接,作为两个分支输入全连接的时空图卷积网络模块;对全连接的时空图卷积网络模块两个分支的输出按元素相加的方式提取出当前片段中实体的特征图;

3)获取用于预测实体分类的向量和预测谓词分布的向量;

4)将每个用于预测实体分类的向量和预测谓词分布的向量相乘,对于每个视频片段,取相乘结果中得分最高的L个关系实例作为带有孪生网络的关联模块的输入;使用孪生网络的在线关联方法将整个视频中的短期关系实例合并;获取关联置信度分数;

5)将检测结果按置信度得分降序排列,得到视频关系检测结果。

2.根据权利要求1所述的基于时空图的视频关系检测方法,其特征在于,所述的步骤1)包括:

把视频分割成多个片段,每个片段包含多帧;对于每个片段,每一帧上产生实体检测框,提取实体特征,并将每个片段中帧级别的实体框连接起来,生成实体轨迹特征;对生成的实体轨迹按vIoU值降序排序,将前N个轨迹作为该片段的实体轨迹特征。

3.根据权利要求2所述的基于时空图的视频关系检测方法,其特征在于,所述的步骤1)中生成实体轨迹特征后,还包括:设置vIoU阈值,去掉低于阈值的实体轨迹的步骤。

4.根据权利要求1所述的基于时空图的视频关系检测方法,其特征在于,所述步骤2)中的时空图卷积网络模块由几何图卷积网络和外观图卷积网络组成;

在几何图卷积网络中,将vIoU值作为仿射矩阵中的值,然后对仿射矩阵的每一行用曼哈顿范数进行归一化;对于几何图卷积网络的输出X^g进行ReLU激活和Layer-Norm,使得几何图卷积网络的输入输出维度保持一致;

在外观图卷积网络中,将两个不同的线性变换应用于输入到外观图卷积网络的实体特征,然后相乘以获得外观相关性值,这些外观相关性值组成外观相关性矩阵A^a,对每一行用softmax重新缩放;对外观图卷积网络的输出X^a进行ReLU激活和Layer-Norm,使得外观图卷积网络的输入输出维度保持一致。

5.根据权利要求4所述的基于时空图的视频关系检测方法,其特征在于,对于输入到时空卷积网络模块的实体特征X,几何图卷积网络的输出X^g和外观卷积网络的输出X^a,按如下公式相加:

$$X' = \mathrm{norm}(\sigma(X^a + X + X^g))$$

对于计算输出结果X',进行ReLU激活和归一化,然后输入下一个时空图卷积网络模块。

6.根据权利要求1所述的基于时空图的视频关系检测方法,其特征在于,所述的步骤3)中,

将步骤2)得到的当前片段中实体的特征图Z输入线性变换层和softmax层,来获取用于预测实体分类的向量V°,公式如下:

$$V_i^\circ = \mathrm{softmax}(\Phi^\circ(Z_i)) \quad (i \in [1,N])$$

其中,Z_i表特征图Z中第i行的特征向量;$\Phi^\circ(Z_i)$表示对Z_i进行线性变换;特征图Z的维度为(N,d),V_i°表示向量V°中的第i个元素。

7.根据权利要求1所述的基于时空图的视频关系检测方法，其特征在于，所述的步骤3)中，

将步骤2)得到的当前片段中实体的特征图Z中每两个特征向量配对组成一个新的<主语，宾语>的特征图，维度为(N*(N-1)，2d)；获取相对运动特征图Z^{rm}；然后这个<主语，宾语>的特征图与相对运动特征图$Z^{rm} \in \mathbb{R}^{(N\times(N-1))\times d'}$进行拼接，生成维度为(N*(N-1)，2d+d')的特征图Z'，之后特征图Z'通过一个线性变换层和sigmoid层生成预测谓词分布的向量V^p，公式如下：

$$V_{ij}^{p}=\mathrm{sigmoid}\left(\phi^{p}\left(Z_i \left\| Z_j \right\| Z_{ij}^{rm}\right)\right) \quad (i,j\in[1,N] \text{ and } i\neq j)$$

其中，ϕ^p表示线性变换，sigmoid()表示一个sigmoid层，Z_i，Z_j表示特征图Z中第i行向量和第j行向量，Z_{ij}^{rm}表示相对运动特征图Z^{rm}的第i行第j列元素，$Z_i \left\| Z_j \right\| Z_{ij}^{rm}$表示$Z_i$，$Z_j$，$Z_{ij}^{rm}$三者进行首尾拼接。

8.根据权利要求1所述的基于时空图的视频关系检测方法，其特征在于，所述的步骤4)中，带有孪生网络的关联模块的处理过程如下：

4.1)将来自两个相邻片段中的任意两个轨迹的特征向量输入孪生网络中，孪生网络是由三个线性变换层组成嵌入网络，然后通过余弦相似度函数计算两个实体的外观相似度的置信度α，公式如下：

$$\alpha\left(\mathcal{T},\mathcal{T}'\right)=\varphi\left(emb\left(f_{\mathcal{T}}\right),emb\left(f_{\mathcal{T}'}\right)\right)$$

其中，emb()表示嵌入网络，$\varphi(\ )$表示余弦相似度函数，$\mathcal{T}$和$\mathcal{T}'$是两个相邻片段中的任意两个轨迹，$f_{\mathcal{T}}$和$f_{\mathcal{T}'}$分别是轨迹$\mathcal{T}$和$\mathcal{T}'$的特征；

4.2)同时考虑几何信息和外观信息，将vIoU值和置信度α再与对应的权重值相乘再相加，得到最后的关联置信度分数$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)$，公式如下：

$$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)=W_g * v\mathrm{IoU}\left(\mathcal{T},\mathcal{T}'\right)+W_a * \alpha\left(\mathcal{T},\mathcal{T}'\right)$$

4.3)当前时刻T所对应的片段中的所有短期关系实例集合为$\mathcal{A}=\left\{\left(\hat{c},\langle\hat{s},\hat{p},\hat{o}\rangle,\left(\hat{\mathcal{T}}_s,\hat{\mathcal{T}}_o\right)\right)\right\}$，其中，$\hat{c}$是短期实例的置信度得分，为预测实体分类的向量$V^{\circ}$的和预测谓词分布的向量$V^p$相乘的结果，$\langle\hat{s},\hat{p},\hat{o}\rangle$是短期实例对应的<主语，谓词，宾语>三元组，$\hat{\mathcal{T}}_s$和$\hat{\mathcal{T}}_o$分别是短期实例中主语对应实体的轨迹和宾语对应的实体轨迹；在时刻T之前的片段已经检测到的所有长期关系实例集合为$\mathcal{S}=\left\{\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)\right\}$，其中c是长期实例的置信度得分，＜s,p,o＞是长期实例对应的<主语，谓词，宾语>三元组，$\mathcal{T}_s$和$\mathcal{T}_o$分别是长期实例中主语对应实体的轨迹和宾语对应的实体轨迹；对集合$\mathcal{A}$和集合$\mathcal{S}$按$\hat{c}$和c降序排序；

然后进行两层循环计算，外层循环遍历集合$\mathcal{S}$和内层循环遍历集合$\mathcal{A}$，对于短期关系

实例$\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)$属于集合$\mathcal{S}$和长期关系实例$\left(\hat{c},<\hat{s},\hat{p},\hat{o}>,\left(\widehat{\mathcal{T}}_s,\widehat{\mathcal{T}}_o\right)\right)$属于集合$\mathcal{A}$，$\mathcal{T}_s$和$\widehat{\mathcal{T}}_s$，$\mathcal{T}_o$和$\widehat{\mathcal{T}}_o$分别按(8)～(9)计算关联置信度分数，只有当短期关系实例和长期关系实例对应的三元组相同，且两个关联置信度分数均大于阈值y时，才会将两者关联合并；对于从第m个片段到第n个片段的长期关系实例$p=\left\{\left(c^t,<s,p,o>,\left(\mathcal{T}_s^t,\mathcal{T}_o^t\right)\right)\right\}(t\in[m,n])$的置信度得分$c_p$，用p中所有的短期关系实例的最高分来更新，公式如下：

$c_p=\max(c^t)\ \ (t\in[m,n])$。

9.一种基于时空图的视频关系检测系统，其特征在于包括：

特征提取模块，用于获取视频片段的帧级别的实体特征，并将每个片段中帧级别的实体框连接起来，生成实体轨迹特征；

特征拼接模块，将前一个片段和当前片段的实体特征和实体轨迹特征，以及当前片段和下一个片段的实体特征和实体轨迹特征分别拼接，作为全连接的时空图卷积网络模块的两个分支输入；

全连接的时空图卷积网络模块，其具有两个分支，包括多个时空图卷积网络模块；每个时空图卷积网络模块由几何图卷积网络和外观图卷积网络组成；输入到时空卷积网络模块的实体特征与时空卷积网络模块中几何图卷积网络的输出、外观卷积网络的输出相加得到该时空卷积网络模块的输出结果，输出结果进行ReLU激活和归一化后作为下一个时空图卷积网络模块的输入；

特征图提取模块，对全连接的时空图卷积网络模块两个分支的输出按元素相加的方式提取出当前片段中实体的特征图；

第一特征向量生成单元，用于获得预测实体分类的向量；

第二特征向量生成单元，用于获取预测谓词分布的向量；

关系实例模块，将每个用于预测实体分类的向量和预测谓词分布的向量相乘，对于每个视频片段，取相乘结果中得分最高的L个关系实例作为带有孪生网络的关联模块的输入；

带有孪生网络的关联模块，使用孪生网络的在线关联方法将整个视频中的短期关系实例合并；获取关联置信度分数；

检测结果输出模块，将带有孪生网络的关联模块的检测结果按置信度得分降序排列，输出视频关系检测结果。

一种基于时空图的视频关系检测方法和系统

技术领域

[0001] 本发明涉及机器学习与计算机视觉研究中的视频关系检测、时空图卷积神经网络、孪生关联网络这几个主题，尤其涉及一种基于时空图的视频关系(视觉关系)检测方法和系统。

背景技术

[0002] 了解视觉信息是计算机视觉的主要目标。视觉内容中的关系检测需要捕获细粒度的视觉线索，包括定位实体的位置以及它们之间的交互方式，这是一项充满挑战但有意义的任务。虽然视频中对象之间的关系是深入理解动态视觉内容的重要组成部分，但是视频中的关系检测和推理却很少被研究。成功检测视频关系的尝试不仅将帮助我们为某些高级视觉理解任务(例如视觉问题解答和视觉字幕)建立更有效的模型，而且还将促进计算机视觉其他领域的发展，例如：视频检索，视频动作检测和视频活动识别。

[0003] 大量的最新研究在静态图像关系检测中获得了令人兴奋的重要成果。视频中关系检测的自然解决方案是将这些方法直接扩展到视频。但是，由于图像和视频之间的内在差异，无法获得令人满意的结果。为静态图像关系检测和推理设计的方法往往会忽略实体之间的动态交互，而动态交互始终在视频中发生。考虑到视频的特性，视频中的关系检测和推理解决方案应该能够捕获实体之间的动态和时变关系。Xindi Shang等人的论文《Video Visual Relation Detection》是迄今为止重点检测视频中的关系的唯一尝试，然而该方法的表现有限，部分原因是它缺乏从周围环境中收集线索的能力。

[0004] 针对上述问题，本发明提出了一种基于时空图的视频关系检测方法。与前面所提到的方法不同，本方法利用实体之间的消息通信来进行视频关系预测。此外，为了解决由于场景改变或者轨迹漂移问题的产生而导致仅依靠几何重叠不能确定连续段中的两个轨迹是否属于同一实体的问题，本发明提出了一种新的使用孪生网络的在线关联方法，该方法同时考虑了外观相似度和关系实例的几何重叠，准确率大有提升。

发明内容

[0005] 本发明的目的是克服现有技术的不足，提供一种基于时空图的视频关系检测方法和系统。

[0006] 本发明首先公开了一种基于时空图的视频关系检测方法，其包括如下步骤：

[0007] 1)获取视频片段的帧级别的实体特征和视频片段的实体轨迹特征；

[0008] 2)将前一个片段和当前片段的实体特征和实体轨迹特征，以及当前片段和下一个片段的实体特征和实体轨迹特征分别拼接，作为两个分支输入全连接的时空图卷积网络模块；对全连接的时空图卷积网络模块两个分支的输出按元素相加的方式提取出当前片段中实体的特征图；

[0009] 3)获取用于预测实体分类的向量和预测谓词分布的向量；

[0010] 4)将每个用于预测实体分类的向量和预测谓词分布的向量相乘，对于每个视频片

段，取相乘结果中得分最高的L个关系实例作为带有孪生网络的关联模块的输入；使用孪生网络的在线关联方法将整个视频中的短期关系实例合并；获取关联置信度分数；

[0011] 5)将检测结果按置信度得分降序排列，得到视频关系检测结果。

[0012] 优选的，带有孪生网络的关联模块的处理过程如下：

[0013] 4.1)将来自两个相邻片段中的任意两个轨迹的特征向量输入孪生网络中，孪生网络是由三个线性变换层组成嵌入网络，然后通过余弦相似度函数计算两个实体的外观相似度的置信度α，公式如下：

[0014] $$\alpha\left(\mathcal{T},\mathcal{T}'\right)=\varphi\left(emb\left(f_{\mathcal{T}}\right),emb\left(f_{\mathcal{T}'}\right)\right)$$

[0015] 其中，emb()表示嵌入网络，$\varphi(\ )$表示余弦相似度函数，$\mathcal{T}$和$\mathcal{T}'$是两个相邻片段中的任意两个轨迹，$f_{\mathcal{T}}$和$f_{\mathcal{T}'}$分别是轨迹$\mathcal{T}$和$\mathcal{T}'$的特征；

[0016] 4.2)同时考虑几何信息和外观信息，将vIoU值和置信度α再与对应的权重值相乘再相加，得到最后的关联置信度分数$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)$，公式如下：

[0017] $$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)=W_g*v\mathrm{IoU}\left(\mathcal{T},\mathcal{T}'\right)+W_a*\alpha\left(\mathcal{T},\mathcal{T}'\right)$$

[0018] 4.3)当前时刻T所对应的片段中的所有短期关系实例集合为$\mathcal{A}=\left\{\left(\hat{c},\langle\hat{s},\hat{p},\hat{o}\rangle,\left(\hat{\mathcal{T}}_s,\hat{\mathcal{T}}_o\right)\right)\right\}$，其中，$\hat{c}$是短期实例的置信度得分，为预测实体分类的向量$V^o$的和预测谓词分布的向量$V^p$相乘的结果，$\langle\hat{s},\hat{p},\hat{o}\rangle$是短期实例对应的<主语，谓词，宾语>三元组，$\hat{\mathcal{T}}_s$和$\hat{\mathcal{T}}_o$分别是短期实例中主语对应实体的轨迹和宾语对应的实体轨迹；在时刻T之前的片段已经检测到的所有长期关系实例集合为$\mathcal{S}=\left\{\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)\right\}$，其中c是长期实例的置信度得分，＜s,p,o＞是长期实例对应的<主语，谓词，宾语>三元组，$\mathcal{T}_s$和$\mathcal{T}_o$分别是长期实例中主语对应实体的轨迹和宾语对应的实体轨迹；对集合$\mathcal{A}$和集合$\mathcal{S}$按$\hat{c}$和c降序排序；

[0019] 然后进行两层循环计算，外层循环遍历集合$\mathcal{S}$和内层循环遍历集合$\mathcal{A}$，对于短期关系实例$\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)$属于集合$\mathcal{S}$和长期关系实例$\left(\hat{c},<\hat{s},\hat{p},\hat{o}>,\left(\widehat{\mathcal{T}}_s,\widehat{\mathcal{T}}_o\right)\right)$属于集合$\mathcal{A}$，$\mathcal{T}_s$和$\widehat{\mathcal{T}}_s$，$\mathcal{T}_o$和$\widehat{\mathcal{T}}_o$分别按(8)～(9)计算关联置信度分数，只有当短期关系实例和长期关系实例对应的三元组相同，且两个关联置信度分数均大于阈值y时，才会将两者关联合并；对于从第m个片段到第n个片段的长期关系实例$p=\left\{\left(c^t,<s,p,o>,\left(\mathcal{T}_s^t,\mathcal{T}_o^t\right)\right)\right\}(t\in[m,n])$的置信度得分$c_p$，用p中所有的短期关系实例的最高分来更新，公式如下：

[0020] $c_p=\max(c^t)\ (t\in[m,n])$。

[0021] 本发明还公开了一种基于时空图的视频关系检测系统，其包括：

[0022] 特征提取模块，用于获取视频片段的帧级别的实体特征，并将每个片段中帧级别的实体框连接起来，生成实体轨迹特征；

[0023] 特征拼接模块，将前一个片段和当前片段的实体特征和实体轨迹特征，以及当前

片段和下一个片段的实体特征和实体轨迹特征分别拼接，作为全连接的时空图卷积网络模块的两个分支输入；

[0024] 全连接的时空图卷积网络模块，其具有两个分支，包括多个时空图卷积网络模块；每个时空图卷积网络模块由几何图卷积网络和外观图卷积网络组成；输入到时空卷积网络模块的实体特征与时空卷积网络模块中几何图卷积网络的输出、外观卷积网络的输出相加得到该时空卷积网络模块的输出结果，输出结果进行ReLU激活和归一化后作为下一个时空图卷积网络模块的输入；

[0025] 特征图提取模块，对全连接的时空图卷积网络模块两个分支的输出按元素相加的方式提取出当前片段中实体的特征图；

[0026] 第一特征向量生成单元，用于获得预测实体分类的向量；

[0027] 第二特征向量生成单元，用于获取预测谓词分布的向量；

[0028] 关系实例模块，将每个用于预测实体分类的向量和预测谓词分布的向量相乘，对于每个视频片段，取相乘结果中得分最高的L个关系实例作为带有孪生网络的关联模块的输入；

[0029] 带有孪生网络的关联模块，使用孪生网络的在线关联方法将整个视频中的短期关系实例合并；获取关联置信度分数；

[0030] 检测结果输出模块，将带有孪生网络的关联模块的检测结果按置信度得分降序排列，输出视频关系检测结果。

[0031] 因为本发明采用了带有孪生网络的关联方法，因此克服了现有技术中采用的贪婪关联算法仅利用几何信息，当轨迹生成不准确或出现轨迹漂移问题时，算法结果不准确地问题，从而有效提高了轨迹关联结果的准确性，提高了关联算法的性能。另外，本发明提出的基于时空图的视频关系检测模型VRD-GCN，将视频抽象为全连接的时空图，在时空图中传递消息并进行推理，方法新颖，视频关系检测结果优秀。

附图说明

[0032] 图1VidVRD视频视觉关系数据样例；

[0033] 图2使用VRD-GCN在VidVRD数据集上的准确率随训练epoch变化曲线；

[0034] 图3算法迭代收敛曲线；

[0035] 图4VRD-GCN视频关系检测结果与基准结果对比。

[0036] 图5本发明的方法流程图。

具体实施方式

[0037] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明的一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0038] 如图5所示，本发明基于时空图的视频关系检测方法包括如下步骤：

[0039] 1)获取视频片段的帧级别的实体特征和视频片段的实体轨迹特征；

[0040] 把视频分割成多个片段，每个片段包含多帧；对于每个片段，每一帧上产生实体检

测框，提取实体特征，并将每个片段中帧级别的实体框连接起来，生成实体轨迹特征；对生成的实体轨迹按vIoU值降序排序，将前N个轨迹作为该片段的实体轨迹特征。

[0041] 2)将前一个片段和当前片段的实体特征和实体轨迹特征，以及当前片段和下一个片段的实体特征和实体轨迹特征分别拼接，作为两个分支输入全连接的时空图卷积网络模块；对全连接的时空图卷积网络模块两个分支的输出按元素相加的方式提取出当前片段中实体的特征图；

[0042] 3)获取用于预测实体分类的向量和预测谓词分布的向量；

[0043] 4)将每个用于预测实体分类的向量和预测谓词分布的向量相乘，对于每个视频片段，取相乘结果中得分最高的L个关系实例作为带有孪生网络的关联模块的输入；使用孪生网络的在线关联方法将整个视频中的短期关系实例合并；获取关联置信度分数；

[0044] 具体的，所述的步骤4)具体如下：

[0045] 4.1)将来自两个相邻片段中的任意两个轨迹的特征向量输入孪生网络中，孪生网络是由三个线性变换层组成嵌入网络，然后通过余弦相似度函数计算两个实体的外观相似度的置信度α，公式如下：

[0046] $$\alpha\left(\mathcal{T},\mathcal{T}'\right)=\varphi\left(emb\left(f_{\mathcal{T}}\right),emb\left(f_{\mathcal{T}'}\right)\right)$$

[0047] 其中，emb()表示嵌入网络，$\varphi(\ )$表示余弦相似度函数，$\mathcal{T}$和$\mathcal{T}'$是两个相邻片段中的任意两个轨迹，$f_{\mathcal{T}}$和$f_{\mathcal{T}'}$分别是轨迹$\mathcal{T}$和$\mathcal{T}'$的特征；

[0048] 4.2)同时考虑几何信息和外观信息，将vIoU值和置信度α再与对应的权重值相乘再相加，得到最后的关联置信度分数$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)$，公式如下：

[0049] $$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)=W_g*v\mathrm{IoU}\left(\mathcal{T},\mathcal{T}'\right)+W_a*\alpha\left(\mathcal{T},\mathcal{T}'\right)$$

[0050] 4.3)当前时刻T所对应的片段中的所有短期关系实例集合为$\mathcal{A}=\left\{\left(\hat{c},\langle\hat{s},\hat{p},\hat{o}\rangle,\left(\hat{\mathcal{T}}_s,\hat{\mathcal{T}}_o\right)\right)\right\}$，其中，$\hat{c}$是短期实例的置信度得分，为预测实体分类的向量$V^o$的和预测谓词分布的向量$V^p$相乘的结果，$\langle\hat{s},\hat{p},\hat{o}\rangle$是短期实例对应的<主语，谓词，宾语>三元组，$\hat{\mathcal{T}}_s$和$\hat{\mathcal{T}}_o$分别是短期实例中主语对应实体的轨迹和宾语对应的实体轨迹；在时刻T之前的片段已经检测到的所有长期关系实例集合为$\mathcal{S}=\left\{\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)\right\}$，其中c是长期实例的置信度得分，＜s,p,o＞是长期实例对应的<主语，谓词，宾语>三元组，$\mathcal{T}_s$和$\mathcal{T}_o$分别是长期实例中主语对应实体的轨迹和宾语对应的实体轨迹；对集合$\mathcal{A}$和集合$\mathcal{S}$按$\hat{c}$和c降序排序；

[0051] 然后进行两层循环计算，外层循环遍历集合$\mathcal{S}$和内层循环遍历集合$\mathcal{A}$，对于短期关系实例$\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)$属于集合$\mathcal{S}$和长期关系实例$\left(\hat{c},<\hat{s},\hat{p},\hat{o}>,\left(\hat{\mathcal{T}}_s,\hat{\mathcal{T}}_o\right)\right)$属于集合$\mathcal{A}$，$\mathcal{T}_s$和$\hat{\mathcal{T}}_s$，$\mathcal{T}_o$和$\hat{\mathcal{T}}_o$分别按(8)～(9)计算关联置信度分数，只有当短期关系实例和长期关系实例对应的三元组相同，且两个关联置信度分数均大于阈值y时，才会将两者关联合并；

对于从第m个片段到第n个片段的长期关系实例$p=\left\{\left(c^t,<s,p,o>,\left(\mathcal{T}_s^t,\mathcal{T}_o^t\right)\right)\right\}(t\in[m,n])$的置信度得分$c_p$,用p中所有的短期关系实例的最高分来更新,公式如下:

[0052] $c_p=\max(c^t)\ (t\in[m,n])$ (10);

[0053] 5)将检测结果按置信度得分降序排列,得到视频关系检测结果。

[0054] 在本发明的一个优选实施例中,所述的步骤1)中生成实体轨迹特征后,还包括:设置vIoU阈值,去掉低于阈值的实体轨迹(减少相似的轨迹)的步骤。

[0055] 在本发明的一个优选实施例中,所述步骤2)中的时空图卷积网络模块由几何图卷积网络和外观图卷积网络组成;

[0056] 在几何图卷积网络中,将vIoU值作为仿射矩阵中的值,然后对仿射矩阵的每一行用曼哈顿范数进行归一化,公式如下:

[0057] $$A_{ij}^{g}=\frac{vIoU\left(\mathcal{T}_i,\mathcal{T}_j\right)}{\sum_{j=0}^{N-1}\left(v\,\mathrm{IoU}\left(\mathcal{T}_i,\mathcal{T}_j\right)\right)}$$

[0058] 其中,$\mathcal{T}_i$表示第i个轨迹,$\mathcal{T}_j$表示第j个轨迹,$vIoU\left(\mathcal{T}_i,\mathcal{T}_j\right)$表示第i个轨迹和第j个轨迹的vIoU值,N表示轨迹的总数量,A_{ij}^{g}表示仿射矩阵中第i行,第j列的值。

[0059] 几何图卷积网络计算公式如下:

[0060] $X^g=\mathrm{norm}(\sigma(A^gXW^g))$

[0061] 其中,A^g是几何图卷积网络的仿射矩阵,$\mathbf{X}\in\mathbb{R}^{2N\times d}$是输入到几何图卷积网络的实体特征,$W^g\in R^{d\times d}$是几何图卷积网络的自适应参数矩阵,σ是非线性激活函数,norm是归一化函数,X^g是几何图卷积网络的输出。

[0062] 同时,对于几何图卷积网络的输出X^g进行ReLU激活和Layer-Norm,使得几何图卷积网络的输入输出维度保持一致;

[0063] 在外观图卷积网络中,首先将两个不同的线性变换应用于输入到外观图卷积网络的实体特征,然后相乘以获得外观相关性值,这些值组成外观相关性矩阵A^a,对每一行用softmax重新缩放,公式如下:

[0064] $$A_{ij}^{a}=\frac{\exp\left(\phi\left(\mathrm{X_i}\right)^T\phi'\left(\mathrm{X_j}\right)\right)}{\sum_{j=0}^{N-1}\left(\exp\left(\phi\left(\mathrm{X_i}\right)^T\phi'\left(\mathrm{X_j}\right)\right)\right)}$$

[0065] 其中,X_i表示第i个实体特征,X_j表示第j个实体特征,$\Phi(X_i)^T$表示对第i个实体特征做线性变换后再进行转置,$\Phi'(X_j)$表示对第j个实体特征做另一种线性变换,exp()表示以自然常数e为底的指数函数,N表示实体特征的个数,A_{ij}^{a}表示外观相关性矩阵第i行,第j列的值。

[0066] 外观图卷积网络计算公式如下:

[0067] $X^a=\mathrm{norm}(\sigma(A^aXW^a))$

[0068] 其中,W^a是外观图卷积网络的自适应参数矩阵,X^a是外观图卷积网络的输出。

[0069] 然后,同样对于外观图卷积网络的输出X^a进行ReLU激活和Layer-Norm,使得外观

图卷积网络的输入输出维度保持一致。

[0070] 在本发明的一个优选实施例中，对于输入到时空卷积网络模块的实体特征X，几何图卷积网络的输出X^g和外观卷积网络的输出X^a，按如下公式相加：

[0071] $X' = \mathrm{norm}(\sigma(X^a + X + X^g))$

[0072] 对于计算输出结果X'，进行ReLU激活和归一化，然后输入下一个时空图卷积网络模块。

[0073] 在本发明的一个优选实施例中，所述的步骤3)中，将步骤2)得到的当前片段中实体的特征图Z输入线性变换层和softmax层，来获取用于预测实体分类的向量V^o，公式如下：

[0074] $V_i^o = \mathrm{softmax}(\phi^o(Z_i))\ (i \in [1,N])$

[0075] 其中，Z_i表特征图Z中第i行的特征向量；$\phi^o(Z_i)$表示对Z_i进行线性变换；特征图Z的维度为(N,d)，V_i^o表示向量V^o中的第i个元素。

[0076] 在本发明的一个优选实施例中，所述的步骤3)中，将步骤2)得到的当前片段中实体的特征图Z中每两个特征向量配对组成一个新的<主语，宾语>的特征图，维度为(N*(N-1)，2d)；获取相对运动特征图Z^{rm}；然后这个<主语，宾语>的特征图与相对运动特征图$Z^{rm} \in \mathbb{R}^{(N\times(N-1))\times d'}$进行拼接，生成维度为(N*(N-1)，2d+d')的特征图Z'，之后特征图Z'通过一个线性变换层和sigmoid层生成预测谓词分布的向量V^p，公式如下：

[0077] $$V_{ij}^p = \mathrm{sigmoid}\left(\phi^p\left(Z_i \| Z_j \| Z_{ij}^{rm}\right)\right) \quad (i,j \in [1,N] \text{ and } i \neq j)$$

[0078] 其中，ϕ^p表示线性变换，sigmoid()表示一个sigmoid层，Z_i，Z_j表示特征图Z中第i行向量和第j行向量，Z_{ij}^{rm}表示相对运动特征图Z^{rm}的第i行第j列元素，$Z_i \| Z_j \| Z_{ij}^{rm}$表示$Z_i$，$Z_j$，$Z_{ij}^{rm}$三者进行首尾拼接。

[0079] 实例1

[0080] 使用视频视觉关系数据集VidVRD测试本方法的视频关系检测能力。VidVRD数据集包含总共1000个视频，这些视频用对象类别和相应的轨迹进行了很好的标记。视觉关系被标记在35个对象类别和132个谓词类别下，标记为<主语，谓词，宾语>。图1展示了VidVRD视频视觉关系数据样例，一个视觉关系实例由关系三元组<主语，谓词，宾语>以及主语和宾语的轨迹表示。

[0081] 下面结合前面所述的具体技术方案说明该实例实施的步骤，如下：

[0082] 1)首先将VidVRD数据集中的80％用来做预采集的训练数据集，将剩余的20％视频数据作为测试视频数据。把每个视频分割成多个片段，每个片段30帧；

[0083] 2)使用微调的Faster R-CNN做为实体检测器，在每一帧上产生实体检测框，并将每个片段中帧级别的实体框连接起来，生成实体轨迹特征。设置vIoU阈值，减少相似的轨迹，并使用前N个轨迹作为关联网络的输入，这里设置N为5；

[0084] 3)取三个相邻的片段(前一个，当前和下一个)，将前一个和当前片段特征和当前和下一个片段特征分别拼接，作为两个分支输入全连接的时空图卷积网络模块。每一个时空图卷积网络模块由几何图卷积网络和外观图卷积网络两部分计算组成；

[0085] 4)在几何图卷积网络中，将vIoU的值作为仿射矩阵中的值，几何图卷积网络计算公式如下：

[0086] $X^g=\mathrm{norm}(\sigma(A^gXW^g))$ (1)

[0087] 其中，A^g是几何图卷积网络的仿射矩阵，$X\in\mathbb{R}^{2N\times d}$是输入到几何图卷积网络的实体特征，$W^g\in R^{d\times d}$是几何图卷积网络的自适应参数矩阵，σ是非线性激活函数，norm是归一化函数，X^g是几何图卷积网络的输出。

[0088] 然后对仿射矩阵的每一行用曼哈顿范数进行归一化，公式如下：

[0089] $$A_{ij}^{g}=\frac{vIoU\left(\mathcal{T}_i,\mathcal{T}_j\right)}{\sum_{j=0}^{N-1}\left(v\,\mathrm{IoU}\left(\mathcal{T}_i,\mathcal{T}_j\right)\right)} \tag{2}$$

[0090] 同时，对于几何图卷积网络的输出X^g进行ReLU激活和Layer-Norm，使得几何图卷积网络的输入输出维度保持一致；

[0091] 5)在外观图卷积网络中，首先将两个不同的线性变换应用于实体特征，然后将它们相乘以获得外观相关性值，这些值组成外观相关性矩阵A^a，对每一行用softmax重新缩放，公式如下：

[0092] $$A_{ij}^{a}=\frac{\exp\left(\phi(\mathrm{X_i})^T\phi'(\mathrm{X_j})\right)}{\sum_{j=0}^{N-1}\left(\exp\left(\phi(\mathrm{X_i})^T\phi'(\mathrm{X_j})\right)\right)} \tag{3}$$

[0093] 外观图卷积网络计算公式如下：

[0094] $X^a=\mathrm{norm}(\sigma(A^aXW^a))$ (4)

[0095] 其中，W^a是外观图卷积网络的自适应参数矩阵，X^a是外观图卷积网络的输出。

[0096] 然后，同样对于外观图卷积网络的输出X^a进行ReLU激活和Layer-Norm，使得外观图卷积网络的输入输出维度保持一致。

[0097] 6)对于原始特征X，几何图卷积网络的输出X^g和外观卷积网络的输出X^a，按如下公式相加：

[0098] $X'=\mathrm{norm}(\sigma(X^a+X+X^g))$ (5)

[0099] 对于计算输出结果X'，进行ReLU激活和归一化，然后输入下一个时空图卷积网络模块。

[0100] 7)对两个分支的输出按元素相加的方式提取出当前片段中实体的特征图$Z\in\mathbb{R}^{N\times d}$。一方面，将Z输入线性变换层和softmax层，来获取用于预测对象分类的向量V^o，公式如下：

[0101] $V_i^o=\mathrm{softmax}(\phi^o(Z_i))(i\in[1,N])$ (6)

[0102] 另一方面，Z中每两个特征向量可以配对组成一个新的<主语，宾语>的特征图，维度为(N*(N-1)，2d)。然后这个<主语，宾语>的特征图与相对运动特征图$Z^{rm}\in\mathbb{R}^{(N\times(N-1))\times d'}$进行拼接，生成维度为(N*(N-1)，2d+d')的特征图，之后特征图通过一个线性变换层和sigmoid层生成预测谓词分布的特征向量V^p，公式如下：

[0103] $$V_{ij}^{p}=\mathrm{sigmoid}\left(\phi^p\left(Z_i\left\|Z_j\right\|Z_{ij}^{rm}\right)\right)\quad(i,j\in[1,N]\ \mathrm{and}\ i\neq j) \tag{7}$$

[0104] 最后，将每个关系实例三元组<主语，谓词，宾语>的V^o和V^p置信度相得分相乘，对于每个片段，取得分最高的n个关系实例作为关联网络的输入；

[0105] 8)将来自两个相邻片段的主语或者宾语的特征向量输入一个嵌入网络,然后通过余弦相似度函数计算两个实体的外观相似度的置信度得分α,公式如下:

[0106] $$\alpha\left(\mathcal{T},\mathcal{T}'\right)=\varphi\left(emb\left(f_{\mathcal{T}}\right),emb\left(f_{\mathcal{T}'}\right)\right) \quad (8)$$

[0107] 其中,$\mathcal{T}$和$\mathcal{T}'$是连续段中的任意两个轨迹;

[0108] 9)为了同时考虑几何信息和外观信息,将vIoU和置信度α在与对应的权重值相乘再相加,得到最后的关联置信度分数$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)$,公式如下:

[0109] $$S_{asso}\left(\mathcal{T},\mathcal{T}'\right)=W_g*v\mathrm{IoU}\left(\mathcal{T},\mathcal{T}'\right)+W_a*\alpha\left(\mathcal{T},\mathcal{T}'\right) \quad (9)$$

[0110] 10)当前时刻T所对应的片段中的所有短期关系实例集合为$\mathcal{A}=\left\{\left(\hat{c},\langle\hat{s},\hat{p},\hat{o}\rangle,\left(\hat{\mathcal{T}}_s,\hat{\mathcal{T}}_o\right)\right)\right\}$,其中,$\hat{c}$是短期实例的置信度得分,为预测实体分类的向量$V^o$的和预测谓词分布的向量$V^p$相乘的结果,$\langle\hat{s},\hat{p},\hat{o}\rangle$是短期实例对应的<主语,谓词,宾语>三元组,$\hat{\mathcal{T}}_s$和$\hat{\mathcal{T}}_o$分别是短期实例中主语对应实体的轨迹和宾语对应的实体轨迹;在时刻T之前的片段已经检测到的所有长期关系实例集合为$\mathcal{S}=\left\{\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)\right\}$,其中c是长期实例的置信度得分,<s,p,o>是长期实例对应的<主语,谓词,宾语>三元组,$\mathcal{T}_s$和$\mathcal{T}_o$分别是长期实例中主语对应实体的轨迹和宾语对应的实体轨迹。对集合$\mathcal{A}$和集合$\mathcal{S}$按$\hat{c}$和c降序排序。

[0111] 然后进行两层循环计算,外层循环遍历集合$\mathcal{S}$和内层循环遍历集合$\mathcal{A}$,对于短期关系实例$\left(c,<s,p,o>,\left(\mathcal{T}_s,\mathcal{T}_o\right)\right)$属于集合$\mathcal{S}$和长期关系实例$\left(\hat{c},<\hat{s},\hat{p},\hat{o}>,\left(\widehat{\mathcal{T}}_s,\widehat{\mathcal{T}}_o\right)\right)$属于集合$\mathcal{A}$,$\mathcal{T}_s$和$\widehat{\mathcal{T}}_s$,$\mathcal{T}_o$和$\widehat{\mathcal{T}}_o$分别按(8)~(9)计算关联置信度分数,只有当短期关系实例和长期关系实例对应的三元组相同,且两个关联置信度分数均大于阈值y时,才会将两者关联合并。对于从第m个片段到第n个片段的长期关系实例$p=\left\{\left(c^t,<s,p,o>,\left(\mathcal{T}_s^t,\mathcal{T}_o^t\right)\right)\right\}(t\in[m,n])$的置信度得分$c_p$,用p中所有的短期关系实例的最高分来更新,公式如下:

[0112] $c_p=\max(c^t)\ (t\in[m,n])$ (10)

[0113] 11)将检测结果按置信度得分降序排列后,使用Recall@K(K=50或100)作为视频视觉关系检测的评估指标,它表示在前K个检测结果中检测到的正确视频视觉关系实例的比例。取得分最高的5个结果来评估结果的准确性。

[0114] 在实施实例中,将本发明方法的检测结果与VidVRD数据集提供的基准结果进行对比。实施结果参见图2~4。图2为使用VRD-GCN在VidVRD数据集上的准确率随训练epoch变化曲线,随着训练epoch数量的增加,准确率逐渐上升,在25个epoch后趋于稳定,图3展示了算法迭代收敛曲线,随着训练epoch数量增加,损失函数值逐渐下降,在25个epoch后下降幅度缓慢,损失值接近于0,图4为VRD-GCN视频关系检测结果与基准结果对比,对于相同一个视频片段,本发明的检测结果相对于基准方法VidVRD的检测结果,检测结果更加丰富,更加准确。

[0115] 以上所述实施例仅表达了本发明的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。因此，本发明专利的保护范围应以所附权利要求为准。

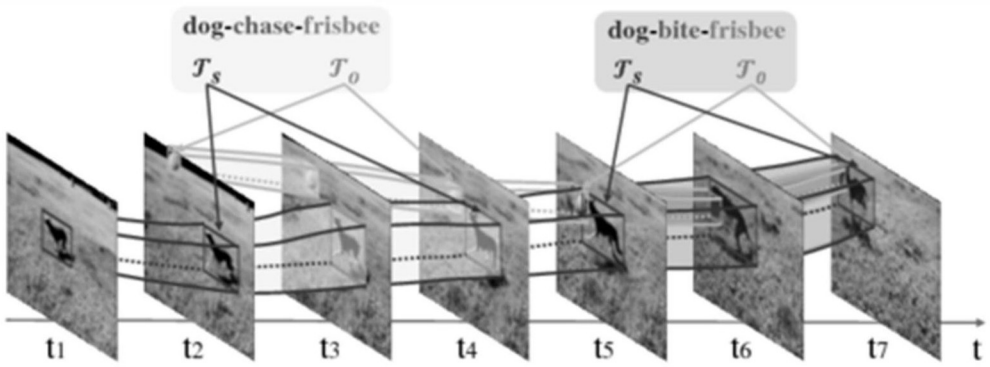

dog-chase-frisbee
$\mathcal{T}_s$
$\mathcal{T}_0$
dog-bite-frisbee
$\mathcal{T}_s$
$\mathcal{T}_0$
t1
t2
t3
t4
t5
t6
t7
t

图1

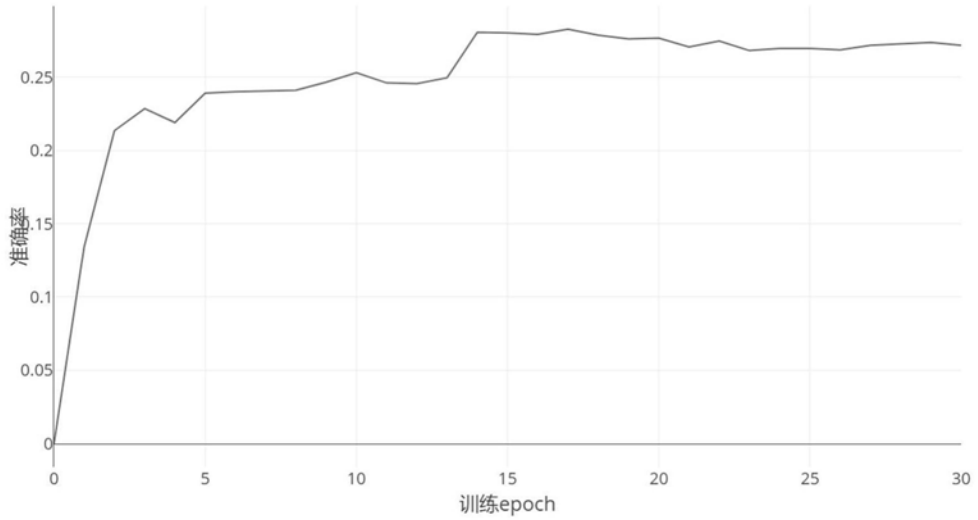

准确率
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
训练epoch

图2

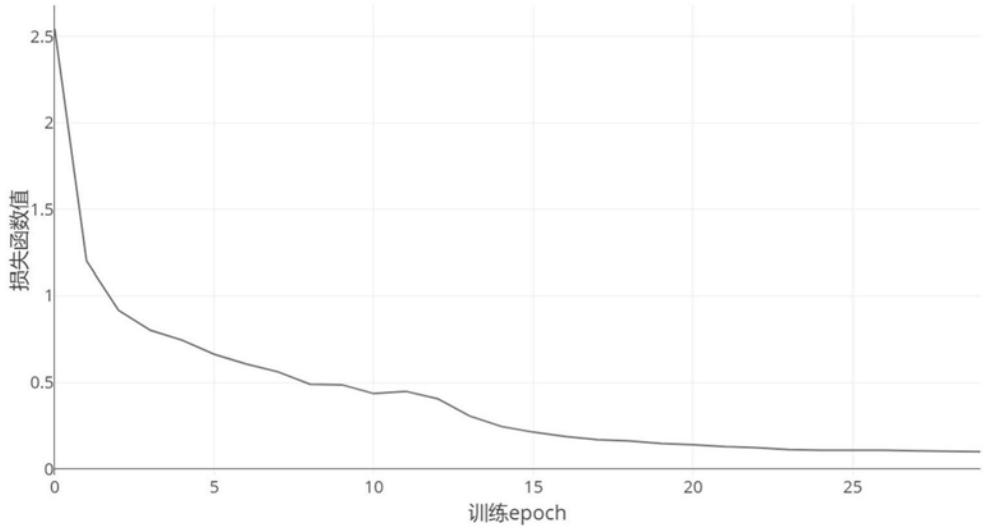

损失函数值
2.5
2
1.5
1
0.5
0
0
5
10
15
20
25
训练epoch

图3


VidVRD
airplane-move_front-watercraft
VRD-GCN
watercraft-move_right-airplane
watercraft-move_behind-airplane
airplane-move_front-watercraft
airplane-move_left-watercraft
airplane-move_past-watercraft

图4

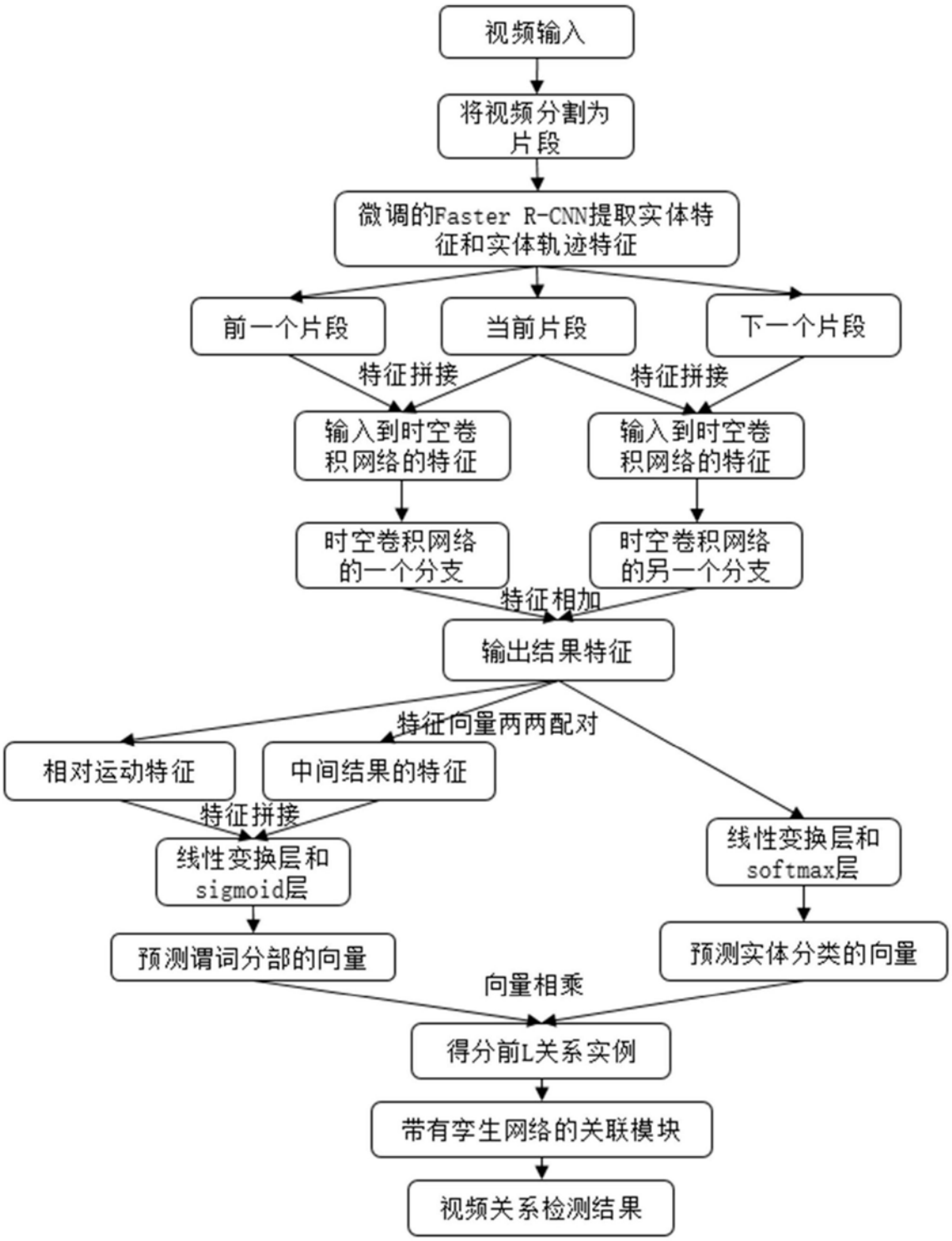

视频输入
将视频分割为片段
微调的Faster R-CNN提取实体特征和实体轨迹特征
前一个片段
当前片段
下一个片段
特征拼接
特征拼接
输入到时空卷积网络的特征
输入到时空卷积网络的特征
时空卷积网络的一个分支
时空卷积网络的另一个分支
特征相加
输出结果特征
特征向量两两配对
相对运动特征
中间结果的特征
特征拼接
线性变换层和sigmoid层
线性变换层和softmax层
预测谓词分部的向量
预测实体分类的向量
向量相乘
得分前L关系实例
带有孪生网络的关联模块
视频关系检测结果

图5