

一种基于多模态异质特征融合的视频密集事件描述方法

技术领域

本发明涉及计算机视觉，涉及视频描述领域，具体涉及一种基于多模态异质特征融合的视频密集事件描述方法。

背景技术

计算机视觉的应用范围与规模是目前人工智能应用中最为广泛与普遍的，且早已深入日常生活与工作的多方面，涉及到网络安全、系统评价、监控、智能机器等，在人类文明发展中有着重要的推动作用，推动社会的发展和进步。计算机视觉人工智能的图像识别中，主要分为静态图像和动态图像，静态图像主要包含图片等内容，动态图像主要是视频等内容。对于视频描述任务来说，可以识别视频中这些事件之间的关系，并把所有事件描述出来，是目前比较重要的研究领域，尤其是涉及到视频监控、网络安全、系统评价等具体应用中。

目前，由于互联网以及终端应用的高度发展，涉及的开放域、长视频、密集事件的视频往往数不胜数，而且处于不断的更新状态。又因为在开放域的长视频中通常包括多个事件，其中牵涉多个行为以及实体，因此合理的描述方式应是对每个事件进行定位并描述。一般的视频描述任务针对一个视频只会生成一句文本描述，当视频中包含多个事件或者复杂场景时，这样的描述显然是不充分的。

此外，由于图像描述任务会对每一张图片生成一句文本描述，如果将图像描述方法应用在视频中则是对特定的视频帧进行描述，如果对每一帧都描述，则会需要较高的计算成本，并且由于生成的描述也过于稠密，还需要使用后处理对描述进行整合，增大了很多工作量。

在图像或视频描述领域，传统机器学习方法已经不满足要求，现在利用卷积神经网络(Convolutional Neural Networks)是比较常用的。其中最大的一个原因就是，图像数据的特征设计，即特征描述一直是计算机视觉研究的痛点。

在特征描述中，重点在于特征融合，特征级融合即对抽象特征进行融合，可分早期融合和晚期融合。早期融合是先将特征融合后再送入模型预测，晚期融合是先将特征送入模型预测再进行融合后统一打分。传统特征融合算法多为早期融合，一般就是对多模态特征进行相加或者级联操作，但是这些简单的线性融合方法无法具备多模态特征间复杂的特征交互能力。目前，融合时参数过多以及多模态特征无法充分交互成了多模态融合领域的主要问题，而且传统融合算法大多达不到现实中多场景应用的要求，缺少模型的泛化能力。

虽然已有诸多视频描述相关的研究工作，但目前来看，视频描述仍然面临着许多挑战。不仅是因为视频描述任务本身对模型的编码解码能力要求较高，还在于当前的一些客观因素限制了视频描述的发展，具体可以总结如下：（1）目前对于开放域、长视频、密集事件的视频描述方法的需求正在日益提升，而目前的大量研究工作集中在图像描述或是仅包含单个事

件的短视频描述；(2)目前的视频描述方法太依赖对已知数据的直观理解，生成简单的共现描述，缺乏对潜在信息的推理过程，导致模型的理解能力局限性较大；(3)大多数视频描述方法还未能完全利用已有数据，仅利用单一模态数据进行分析限制了模型的结构和性能。

发明内容

为了有效解决长视频、密集事件的视频描述方法中频繁生成共现描述等无意义描述的弊端，同时满足多场景的应用的要求，本发明提供一种基于多模态异质特征融合的视频密集事件描述方法。

本发明请求保护以下技术方案，

本发明提供一种基于多模态异质特征融合的视频密集事件描述方法，其特征在于，包括以下步骤：

特征提取，利用 I3D 卷积网络选取开放域中长视频进行剪辑得到对应的视频帧，并对所述视频帧中的动态视觉特征 F_m 进行提取，获取动态视觉特征向量；利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量；

场景图特征嵌入，将视频帧中的场景对象信息进行语义化表征，生成场景图谱，得到实体编码，属性编码和关系编码，通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入，得到场景图特征嵌入向量；

三重多模态循环融合，将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入，并将其进行三重多模态循环融合，再将输出的向量使用规范化函数沿通道做归一化，并且利用 Dropout 进行正则化后输入到解码器中，并为重构解码器做准备；

描述重构解码器，包括预训练的语言字典解码器以及描述重构算法，通过描述重构算法对视频密集事件进行检测；并利用预训练的语言字典解码器来解码生成视频场景事件的描述；所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

进一步的，所述动态视觉特征 F_m ，其中 $F_m = [f_1^m, f_2^m, \dots, f_N^m]$ ，其中 N 为抽取的视频剪辑的数量，整体对应的维度是 $T_v \times d_v$ ；所述音频韵律特征向量对应的维度是 $T_a \times d_a$ ，音频特征向量被转化为具有语义的 128 维的特征向量。

进一步的，在场景图特征嵌入过程中，输入为视频帧，首先经过 Mask R-CNN 检测视频帧图片中的物体包围框坐标等物体位置信息，确定物体目标所在位置，得到物体区域；再将所述视频帧原图和所述物体区域分别送入视觉场景图生成器生成场景图，将所述视频帧图片信息以及被检测出来的物体位置信息等视觉信息进行语义化表征，得到实体编码、属性编码和关系编码，之后将这些编码送入联合嵌入层进行联合表征得到语义信息，将所述语义信息嵌入所述实体编码、属性编码和关系编码中对应的特征向量；再通过图卷积对特征向量进行

图嵌入，得到场景图特征嵌入向量；在此过程中对实体之间的关联进行强化推理。

进一步的，本发明在三重多模态循环融合之后，还包括：

自适应多模态数据平衡，利用全卷积网络对所述动态视觉特征向量和所述音频韵律特征向量在时序维度上进行特征提取，并使用特定值来预测事件的时序边界以及置信度分数；其次，将预测得到的事件送入公共池中，并按照降序排列并选取高置信度分数的事件；最终对所述高置信度分数的事件进行聚类，选择各个对应的聚类中心来组成最终的事件集合。

进一步的，所述特定值是指中心位置 $center = p + \sigma(c)$ ，其中 $\sigma(\cdot)$ 是 *sigmoid* 函数，作用是把 c 的值缩放到 $[0, 1]$ 区间内，用来确定中心点 p 的右邻域；事件时长 $length = anchor \cdot e$ ，其中 e 是缩放系数；置信度分数 $confidence = \sigma(o)$ 。

进一步的，所述全卷积网络皆采用一维卷积对时序维度上进行特征提取，第一层的卷积核大小为 k ，第二层和第三层的卷积核都为 1，前两层的激活函数是 *Relu*，最后一层的激活函数是 *sigmoid*。

进一步的，在三重多模态循环融合过程中，所述规范化函数指是 *Softmax* 函数。

在三重多模态循环融合过程中，维度为 $T_v \times d_v$ 以及 $T_a \times d_a$ 的动态视觉特征向量和音频特征向量先被送入自注意力层进行特征表示。

进一步的，所述高置信度分数取前 100 的事件，聚类算法采用根据欧式距离的 *K-Means* 算法。

进一步的，所述语言字典的预训练方法为，将已标注的描述文本转换为场景图，再通过图卷积网络进行场景图特征嵌入，将得到的向量与所述预训练的语言字典进行计算对场景图特征进行重编码，再将重编码后的向量送入解码器重构后进行文本描述，通过比较重构后的文本描述与所述已标注的描述文本的差异来监督模型学习；所述标注描述文本是存于数据集中，用于预训练的数据；

所述预训练的语言字典，是指预训练的语言字典 D 是一个维度为 $d \times L$ 的矩阵；

所述重编码，是指进行重编码 $\varphi(\cdot)$ 的计算，过程如下：

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

其中 d_l 表示 D 中第 l 列向量，此处 D 的作用就是将人类语言习惯进行编码嵌入。

进一步的，在所述描述重构解码器步骤中，所述的描述重构算法具体为：

描述重构算法

输入：视频的开始时间 P 和结束时间 Q ；事件集合 E ，每个事件包含属性：事件的中心时间点 m ，事件时长 l ，置信度分数 s ，事件对应的描述 z

输出：候选描述集合 Z

1 根据 m 和 l 计算出每个事件的开始时间 s 和结束时间 e ：

$$s = m - \frac{l}{2}, e = m + \frac{l}{2};$$

- 2 计算出事件之间的 tIOU，将重叠事件去重；
- 3 将事件按照开始时间升序排列并选择开始时间最早的事件作为基准事件 k ；
- 4 从剩余事件集合 $E - k$ 中选择开始时间在 $[k_s, k_s + \varepsilon]$ 中的事件一同组成起始事件集合 E_1 ；
- 5 从剩余事件集合 $E - E_1$ 中选择开始时间在 $[E_{1e} - \varepsilon, E_{1e} + \varepsilon]$ 中的事件与 E_1 中对应的事件组成事件序列集合 $\hat{E}$ ；
- 6 重复步骤 5 并更新 $\hat{E}$ 直到选择的事件的结束时间在 $[Q - \varepsilon, Q]$ 范围内，更新 $\hat{E}$ ，选择结束；
- 7 将 $\hat{E}$ 中每个事件序列对应的文本描述进行合并后加入到候选描述集合 Z 。

本发明还提供一种兼顾参数量与交互强度的融合方法，其特征在于，可以用于实现权利要求 1 所述的三重多模态循环融合，在不产生高维特征的前提下，让多种特征进行深层次融合。

本发明还提供一种基于多模态异质特征融合的视频密集事件描述的装置，其特征在于，特征提取模块，用于利用 I3D 卷积网络选取开放域中长视频进行剪辑得到对应的视频帧，并对所述视频帧中的动态视觉特征 F_m 进行提取，获取动态视觉特征向量；利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量；

场景图特征嵌入模块，用于将视频帧中的场景对象信息进行语义化表征，生成场景图谱，得到实体编码，属性编码和关系编码，通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入，得到场景图特征嵌入向量；

三重多模态循环融合模块，用于将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入，并将其进行三重多模态循环融合，再将输出的向量使用规范化函数沿通道做归一化，并且利用 Dropout 进行正则化后输入到解码器中，并为重构解码器做准备；所述三重多模态循环融合模块包括预先建立的一种兼顾参数量与交互强度的融合方法，在不产生高维特征的前提下，让多种特征进行深层次融合；

描述重构解码器模块，用于通过描述重构算法对视频密集事件进行检测；并利用预训练的语言字典解码器来解码生成视频场景事件的描述，所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

进一步的，还包括自适应多模态数据平衡模块，用于利用全卷积网络对所述动态视觉特征向量和所述音频韵律特征向量在时序维度上进行特征提取，并使用特定值来预测事件的时序边界以及置信度分数；其次，将预测得到的事件送入公共池中，并按照降序排列并选取高置信度分数的事件；最终对所述高置信度分数的事件进行聚类，选择各个对应的聚类中心来组成最终的事件集合。

与现有技术相比，本发明的优势在于：

(1) 本发明针对视频密集事件的描述，通过利用 I3D 卷积网络以及 VGGish 模型提取特征，对视频描述任务的进一步细化，解决了长视频的描述的对每个事件进行定位问题。并在视频密集事件描述方法中首次提出使用场景图特征嵌入的方式，并通过本发明设计的三重多模态循环融合模块实现了多模态循环融合，使得视频密集描述任务描述了事件对象之间的关联关系、

效果更好，解决了目前的大量研究工作集中在图像描述或是仅包含单个事件的短视频描述的缺陷。

(2) 本发明搭载了自适应多模态数据平衡，使得动态视觉和音频韵律两个特征的相互契合，保证事件提取的合理性，提高了模型的泛化能力。

(3) 本发明还重构了解码器，设计了一种描述重构算法对视频事件进行检测，并根据预训练得到的语言字典生成视频场景事件的描述，有效的解决了视频描述方法中频繁生成共现描述等无意义描述的弊端，并且有效利用了多模态信息挖掘出场景事件的关系。

附图说明

图 1. 本发明提供了一种基于多模态异质特征融合的视频密集事件描述方法的步骤示意图。

图 2. 本发明提供了一种基于多模态异质特征融合的视频密集事件描述装置的配置程序图。

图 3. 本发明提供的另一种基于多模态异质特征融合的视频密集事件描述装置的配置程序图。

图 4. 本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的步骤示意图。

图 5. 本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的流程图。

图 6. 本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的场景图嵌入生成示意图。

图 7. 本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的三重多模态循环融合模块示意图。

图 8. 本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的语言字典训练示意图。

具体实施方式

为了使本技术领域的人员更好地理解本发明方案，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，所描述的实施例仅是本发明一部分的实施例，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做出创造性劳动前提下所获得的所有其他实施例，都应当属于本发明保护的范围。

为刚便于理解本发明的技术方案，下面对本发明涉及的专业名词做详细说明。

视频密集事件描述任务是对视频描述任务的进一步细化，刚好解决了长视频的描述的每个事件进行定位问题。视频密集事件描述任务是先定位视频中包含关键事件的视频片段，

包括事件发生和结束的时间戳，再对获取到的片段进行描述，最后再对多句描述进行整合。该任务能够保证生成的描述在时序上的正确性，且满足对多事件描述的需求，同时也比一般的视频描述任务更具有挑战性。

场景图是有向图的一种数据结构，可以定义为元组（三元组） $G=(O,E)$ ，其中 O 是图像中检测到的一组对象，在图中表示为节点。每个对象都有 $O_i = (C_i, A_i)$ 的形式，其中 C_i 和 A_i 分别是对象的类别和属性。而 E 是一组有向边，表示宾语之间的成对关系，可以是动词、空间、介词和比较级。场景图能够包含图像的结构化语义信息，包括当前对象、对象属性和成对关系的知识。因此，场景图可以为图像检索、视频问题回答和图像生成等其他视觉任务提供有益的先验知识。

图卷积是图卷积神经网络结构中的一种计算方法。

图嵌入指的是将高维的特征向量进行计算映射成低维的向量，这样的好处是可以抽取向量之间更深层的关系，并且减轻计算量。

开放域可以理解为公开域，比如不侵权的视频领域。开放域视频是指包括教育、医疗、科学等多个领域的视频，本发明的方法并不局限于某一领域。

为使本发明技术方案的优点更加清楚，下面结合附图和实施例对本发明做详细说明。

本发明提供了一种基于多模态异质特征融合的视频密集事件描述方法，其中，包括以下步骤，如图 1 所示：

步骤 S11，特征提取，利用 I3D 卷积网络选取开放域中长视频进行剪辑得到对应的视频帧，并对所述视频帧中的动态视觉特征 F_m 进行提取，获取动态视觉特征向量；利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量；

步骤 S12，场景图特征嵌入，将视频帧中的场景对象信息进行语义化表征，生成场景图谱，得到实体编码，属性编码和关系编码，通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入，得到场景图特征嵌入向量；

步骤 S13，三重多模态循环融合，将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入，并将其进行三重多模态循环融合，再将输出的向量使用规范化函数沿通道做归一化，并且利用 Dropout 进行正则化后输入到解码器中，并为重构解码器做准备；

步骤 S14，描述重构解码器，包括预训练的语言字典解码器以及描述重构算法，通过描述重构算法对视频密集事件进行检测，这里视频密集事件是视频帧的事件的集合；并利用预训练的语言字典解码器来解码生成视频场景事件的描述；所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

基于上述方法，本发明还提供了一种基于多模态异质特征融合的视频密集事件描述装置，图 2 为其配置程序图。如图 2 所示，包括：

特征提取模块 101，用于对开放域中长视频进行剪辑得到对应的视频帧，并对所述视频帧中的动态视觉特征 F_m 进行提取，获取动态视觉特征向量；利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量。

进一步的，在特征提取模块 101 中，具体包括利用 I3D 卷积网络选取开放域中长视频进行剪辑得到对应的视频帧，并对所述视频帧中的动态视觉特征 F_m 进行提取，获取动态视觉特征向量；所述动态视觉特征 F_m ，其中 $F_m = [f_1^m, f_2^m, \dots, f_N^m]$ ，其中 N 为抽取的视频剪辑的数量，整体对应的维度是 $T_v \times d_v$ ；所述音频韵律特征向量对应的维度是 $T_a \times d_a$ ，音频特征向量被转化为具有语义的 128 维的特征向量。VGGish 模型的网络结构本质上就是卷积神经网络中的 VGG 网络，该模型以音频数据作为输入，从中提取具有语义信息的特征向量。

场景图特征嵌入模块 102，用于将视频帧中的场景对象信息进行语义化表征，生成场景图谱，得到实体编码，属性编码和关系编码，通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入，得到场景图特征嵌入向量。

进一步的，在场景图特征嵌入模块 102 中，输入为视频帧，首先经过 Mask R-CNN 检测视频帧图片中的物体包围框坐标等物体位置信息，确定物体目标所在位置，得到物体区域；再将所述视频帧原图和所述物体区域分别送入视觉场景图生成器生成场景图，将所述视频帧图片信息以及被检测出来的物体位置信息等视觉信息进行语义化表征，得到实体编码、属性编码和关系编码，之后将这些编码送入联合嵌入层进行联合表征得到语义信息，将所述语义信息嵌入所述实体编码、属性编码和关系编码中对应的特征向量；再通过图卷积对特征向量进行图嵌入，得到场景图特征嵌入向量；在此过程中对实体之间的关联进行强化推理。

三重多模态循环融合模块 103，用于将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入，并将其进行三重多模态循环融合，再将输出的向量使用规范化函数沿通道做归一化，并且利用 Dropout 进行正则化后输入到解码器中，并为重构解码器做准备；

进一步的，所述三重多模态循环融合模块包括预先建立的一种兼顾参数量与交互强度的融合方法，在不产生高维特征的前提下，让多种特征进行深层次融合。所述规范化函数指的是 Softmax 函数。其中，维度为 $T_v \times d_v$ 以及 $T_a \times d_a$ 的动态视觉特征向量和音频特征向量先被送入自注意力层进行特征表示。

这里需要注意的是，本发明还提供了一种兼顾参数量与交互强度的融合方法，其中，可以用于实现上述方法中的所述的三重多模态循环融合，在不产生高维特征的前提下，让多种特征进行深层次融合。

这里当然还涉及其他多重融合，比如可以同时三个特征向量或者是两个特征向量、一个特征图做融合。包括在不考虑计算机性能的前提下，对任意数量的模态特征进行循环融合。但是由于随着特征融合数量的增多，必然会导致计算性能的下降，而且有可能还会引发模型

的过拟合问题，因此，三种特征融合效果是良好的、可接受的。

描述重构解码器模块 104，用于通过描述重构算法对视频密集事件进行检测；并利用预训练的语言字典解码器来解码生成视频场景事件的描述，所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

除上述实施例图 2 中的功能模块外，本发明还提供了另一种包含自适应多模态数据平衡模块的基于多模态异质特征融合的视频密集事件描述装置，如图 4 所示，是本发明提供的另一种基于多模态异质特征融合的视频密集事件描述装置的配置程序图，包括：特征提取模块 201、场景图特征嵌入模块 202、三重多模态循环融合模块 203、自适应多模态数据平衡模块 204 以及描述重构解码器模块 205，所述自适应多模态数据平衡模块 204，用于利用全卷积网络对所述动态视觉特征向量和所述音频韵律特征向量在时序维度上进行特征提取，并使用特定值来预测事件的时序边界以及置信度分数；其次，将预测得到的事件送入公共池中，并按照降序排列并选取高置信度分数的事件；最终对所述高置信度分数的事件进行聚类，选择各个对应的聚类中心来组成最终的事件集合。

图 3 是本发明基于图 4 中的程序功能模块的装置，提供的另一种基于多模态异质特征融合的视频密集事件描述方法的步骤示意图。其中描述了本发明的另一种基于多模态异质特征融合的视频密集事件描述方法的步骤。具体如下：

步骤 S21 特征提取，对开放域中长视频进行剪辑得到对应的视频帧，并对所述视频帧中的动态视觉特征 F_m 进行提取，获取动态视觉特征向量；利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量；

步骤 S22 场景图特征嵌入，将视频帧中的场景对象信息进行语义化表征，生成场景图谱，得到实体编码，属性编码和关系编码，通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入，得到场景图特征嵌入向量；

步骤 S23 三重多模态循环融合，将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入，并将其进行三重多模态循环融合，再将输出的向量使用规范化函数沿通道做归一化，并且利用 Dropout 进行正则化后输入到解码器中，并为重构解码器做准备；

步骤 S24 自适应多模态数据平衡，利用全卷积网络对所述动态视觉特征向量和所述音频韵律特征向量在时序维度上进行特征提取，并使用特定值来预测事件的时序边界以及置信度分数；其次，将预测得到的事件送入公共池中，并按照降序排列并选取高置信度分数的事件；最终对所述高置信度分数的事件进行聚类，选择各个对应的聚类中心来组成最终的事件集合。

步骤 S25 描述重构解码器，包括预训练的语言字典解码器以及描述重构算法，通过描述重构算法对视频密集事件进行检测；并利用预训练的语言字典解码器来解码生成视频场景事件的描述，所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

为了便于更清楚的了解本发明的具体的实施过程，在图 3 提供的另一种基于多模态异质特征融合的视频密集事件描述方法的步骤的基础上，这里提供了另一种包含步骤 S24 自适应多模态数据平衡的基于多模态异质特征融合的视频密集事件描述方法的具体的实施流程示意图。如图 5 所示，本发明可以通过以下的步骤实现对视频流中密集事件的描述，有效解决了视频描述方法中频繁生成共现描述等无意义描述的弊端，并且有效利用了多模态信息挖掘出场景事件的关系。其中，将输入的视频流中开放域长视频经过 S21 进行剪辑后，分别提取动态视觉特征和音频韵律特征；通过 S22 将提取的动态视觉特征进行场景图特征嵌入，得到场景图特征；通过 S23 三个特征进行三重多模态循环特征融合，再经过 S24 自适应多模态数据平衡后，再通过 S25 描述重构解码器，得到生成视频场景事件的描述。

具体的分步骤过程，如下所示：

步骤 S21 特征提取包括：提取动态视觉特征，利用 I3D 卷积网络选取一定数量视频剪辑（clip），这里是指选取的开放域或者实际监控视频中的长视频逐个视频进行剪辑；并对其中的动态视觉特征 F_m 进行提取，在获取动态视觉特征的同时还能获取平滑的流动信息，这里平滑的流动信息指的是光流信息。I3D 是 C3D 模型和双流网络的结合体，它不仅通过对二维网络中的所有卷积核和池化核额外赋予了一个时间维度来将二维卷积扩展到了三维，还使用了 Inception（INC）结构来提升运算效率。

步骤 S21 特征提取还包括：音频韵律信息抽取，使用预训练的 VGGish 模型来提取视频源中音频韵律特征向量。具体来说，利用 VGGish 模型将音频韵律特征向量转化为具有语义的 128 维的特征向量，而 128 维的特征向量与动态视觉特征向量融合起来作为下游模型的输入。首先将输入音频重采样为 1.6 万赫兹的单声道音频；其次，使用 20 毫秒的海宁时窗和 15 毫秒的帧移对音频进行短时傅里叶变换得到频谱图；然后，将频谱图映射到 64 阶的梅尔滤波器组中计算梅尔声谱；最后，将这些特征以 0.96 秒的时长进行组帧，每一帧都包含 64 个梅尔频带，时长 10 毫秒（即总共 96 帧）。模型最终输出为 128 维的特征向量，即 $d_a=128$ 。

在本实施例中，所述 I3D 卷积网络模型具体为：I3D 中有两个支路并行对输入的视频以及视频的光流图进行处理，光流图通过 opencv 提取得到，最终将两条支路的结果进行线性相加得到最终的视觉特征作为输出。本文使用的 I3D 模型在 Kinetics 数据集上进行预训练，其输入为 64 帧 RGB 视频和光流视频，每帧图片大小为 224*224，视频帧率为 25。在提取过程中，本文将输入的视频取中心位置截取为 224*224 的尺寸，同时按照 64 帧进行分段。I3D 模型每 64 帧输出 1024 维 RGB 特征向量和光流特征向量，最后再对这两个向量做加和得到针对 64 帧视频段的 1024 维输出向量。

步骤 S22 场景图特征嵌入，是将场景对象信息进行语义化表征，生成场景图谱，得到实体编码，属性编码和关系编码，通过图卷积对特征向量进行图嵌入。这里的场景对象指视频帧中的对象，如“人”，“马”，“医院”，“手表”等实体对象。具体可见图 7 中绘制出的场景

图。具体来说,首先经过 Mask R-CNN 检测视频帧图片中的物体包围框坐标等物体位置信息,确定物体目标所在位置,得到物体区域;再将所述视频帧原图和所述物体区域分别送入视觉场景图生成器生成场景图,将所述视频帧图片信息以及被检测出来的物体位置信息等视觉信息进行语义化表征,得到实体编码、属性编码和关系编码,之后将这些编码送入联合嵌入层进行联合表征得到语义信息,将所述语义信息嵌入所述实体编码、属性编码和关系编码中对应的特征向量;再通过图卷积对特征向量进行图嵌入,得到场景图特征嵌入向量;在此过程中对实体之间的关联进行强化推理。这里所述的实体编码,属性编码,关系编码中对应的特征向量,其中,如图 7 所示,实体是人,马,乡村小路,关系是 riding 和 on,属性是红色;此处是为了配合理解举例,实际过程中这些都是被编码的向量。

步骤 S23 是三重多模态循环融合,将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入,并将其进行三重多模态循环融合,再将输出的向量使用规范化函数沿通道做归一化,并且利用 Dropout 进行正则化后输入到解码器中,并为重构解码器做准备。

进一步的,首先将所述场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量,这三种特征投影到同一低维空间中,并将同一维度的特征向量分别进行循环矩阵操作,将其转化为循环矩阵,从第二行向量开始,把每一个行向量都定义为上一个行向量将整体循环右移一个元素。其次,将每一种向量与除此种特征信息之外所有的循环矩阵做元素积,以探索各个特征向量信息之间所包含的复杂关联。最后,对三种不同模态交互后的特征向量进行归一化和正则化后送入解码器中,并为重构解码器做准备。

步骤 S24 是自适应多模态数据平衡,使所述动态视觉和音频韵律两个特征的相互契合,保证事件提取的合理性。具体来说,首先利用全卷积网络对动态视觉特征向量和音频特征向量在时序维度上进行特征提取,并使用三个特定值来预测事件的时序边界以及置信度分数。其次,将得到的事件送入公共池中,并按照降序排列并选取高置信度分数的事件。最终对这些事件进行聚类,选择各个聚类中心组成最终的事件集合。例如:一段音频特征中全是噪音,这里认为在这个时间维度上的音频韵律特征向量严重影响此维度的动态视觉特征向量,导致多模态融合的无效性,影响后期事件的描述。而多模态平衡主要体现在这里,使得动态视频特征和音频特征相互契合,保证事件提取的合理性。

进一步的,这里使用的三个特征值来预测事件的时序边界以及置信度分数具体为:

事件发生时间段的中心位置 $center = p + \sigma(c)$, 其中 $\sigma(\cdot)$ 是 sigmoid 函数,作用是把 c 的值缩放到 $[0, 1]$ 区间内,用来确定中心点 p 的右邻域;事件时长 $length = anchor \cdot e$, 其中 e 是缩放系数;置信度分数 $confidence = \sigma(o)$ 。这里利用置信度分数来筛选合理的事件,剔除无意义的事件,如:同一事件中的空白期,无法对这种事件生成其描述。

在本实施例中,所述的多模态数据平衡,本质上是一个全卷积网络,由三层卷积层组成。

其中卷积皆采用一维卷积对时序维度上进行特征提取，来确定哪些时间段包含重要的事件。第一层的卷积核大小为 k ，第二层和第三层的卷积核都为 1，前两层的激活函数是 Relu，最后一层的激活函数是 *sigmoid*。该模块的输入是由编码模块输出的维度分别为 $T_v \times d_v$ 以及 $T_a \times d_a$ 的动态视觉特征向量和音频韵律特征向量。这里我们分别对动态视觉和音频两条线路使用该机制来提取时序维度上的事件，再将提取到的事件输入到公共池进行排序，最终选择出合适的事件集合。

进一步的，所述的卷积操作具体为：动态视觉特征向量和音频韵律特征向量需要分别做 K_v 和 K_a ($K_v = K_a$) 次全卷积操作，以便能够动态捕捉不同时长的事件，且每次操作中第一层卷积核的 k 值都不同，以保证卷积核具有丰富的感受野。

进一步的，所述预测事件的数量具体为：这里定义 φ 来决定在做第 m 次全卷积操作时需要预测的事件的数量。由于动态视觉特征向量的维度 T_v 和音频韵律特征向量的维度 T_a 是不同的，因此本文还分别定义了 φ_v 和 φ_a ($\varphi_v \neq \varphi_a$) 来保证 $T_v \cdot |\varphi_v|$ 和 $T_a \cdot |\varphi_a|$ 尽可能接近，使得两条线路生成的事件在数量上尽可能匹配。

进一步的，所述最终事件集合具体为：时序维度上提取到的事件被送入公共池，在公共池中将所有事件按照降序排列并选取置信度分数前 100 的事件，最后使用 K-Means 算法根据欧氏距离对这些事件进行聚类，选择各个聚类中心组成最终的事件集合。

步骤 S25 是描述重构解码器，包括预训练的语言字典解码器以及描述重构算法。其中，语言字典的预训练方法为，将已标注的描述文本转换为场景图，再通过图卷积网络进行特征嵌入，将得到的向量与预训练的语言字典 D (D 是一个维度为 $d \times L$ 的矩阵) 进行计算对特征进行重编码，再将重编码后的向量送入解码器重构文本描述，通过比较重构后的描述与已标注的描述文本的差异来监督模型学习。所述重编码 $\varphi(\cdot)$ 的计算过程如下：

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

其中 d_l 表示 D 中第 l 列向量，此处 D 的作用就是将人类语言习惯进行编码嵌入。

进一步的，在 S25 步骤中，描述重构算法可以将生成的多事件描述整合成一段完整的描述，以此来保证描述在逻辑上的时序正确性。

所述的描述重构算法具体为：

描述重构算法

输入：视频的开始时间 P 和结束时间 Q ；事件集合 E ，每个事件包含属性：事件的中心时间点 m ，事件时长 l ，置信度分数 s ，事件对应的描述 z

输出：候选描述集合 Z

2 根据 m 和 l 计算出每个事件的开始时间 s 和结束时间 e ：

$$s = m - \frac{l}{2}, e = m + \frac{l}{2};$$

2 计算出事件之间的 tIOU，将重叠事件去重；

3 将事件按照开始时间升序排列并选择开始时间最早的事件作为基准事件 k ；

- 4 从剩余事件集合 $E - k$ 中选择开始时间在 $[k_s, k_s + \varepsilon]$ 中的事件一同组成起始事件集合 E_1 ;
- 5 从剩余事件集合 $E - E_1$ 中选择开始时间在 $[E_{1e} - \varepsilon, E_{1e} + \varepsilon]$ 中的事件与 E_1 中对应的事件组成事件序列集合 $\hat{E}$;
- 6 重复步骤 5 并更新 $\hat{E}$ 直到选择的事件的结束时间在 $[Q - \varepsilon, Q]$ 范围内, 更新 $\hat{E}$, 选择结束;
- 7 将 $\hat{E}$ 中每个事件序列对应的文本描述进行合并后加入到候选描述集合 Z 。

为了更便于理解本发明设计的步骤 S22 场景图特征嵌入的具体实施过程, 这里列出图 6, 是本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的场景图嵌入生成示意图。

如图 6 所示, 根据 Mask R-CNN 检测出图中对象的位置, 并依据对象的局部信息和对象间的全局信息生成场景图 G , $G = \langle e_o, e_a, e_r \rangle$, e_o, e_a, e_r 分别代表场景图中实体、属性和关系对应的向量表示。图 6 中 a_{ij} 表示第 i 个实体的第 j 个属性, o_i 表示第 i 个实体, r_{ij} 表示 o_i 和 o_j 之间的关系, 箭头头部的实体代表关系主体, 尾部的实体代表关系客体。生成场景图后利用四个空间图卷积层 g_r, g_a, g_s, g_o 对实体、关系和属性向量进行图嵌入, 四个图卷积层都具有相同的结构: 一层用 Relu 函数激活的全连接层, 可将其形式化表达为 $g(\cdot) = \text{ReLU}(fc(\text{concat}(e_1, e_2, \dots, e_n)))$, e_i 表示所有输入的向量。对于关系特征嵌入 $x_{r_{ij}}$ 定义如下: 对于场景图中每一对关系 $\langle o_i, r_{ij}, o_j \rangle$,

$$x_{r_{ij}} = g_r(e_{o_i}, e_{r_{ij}}, e_{o_j})$$

上式能将 o_i 和 o_j 上下文关系合并起来。对于属性特征嵌入 x_{a_i} 定义如下: 对于场景图中的某个实体 o_i 及其属性 a_{ij} ,

$$x_{a_i} = \frac{\sum_{k=1}^j g_a(e_{o_i}, e_{a_{ik}})}{j}$$

上式能将 o_i 和 o_i 的所有属性合并起来。对于实体特征嵌入 x_{o_i} 定义如下: 对于场景图中的每个实体 o_i , 由于在某个关系中, 实体既有可能是关系主体又有可能是关系客体, 因此需要将不同的情况都囊括进来, 故

$$x_{o_i} = \frac{1}{N_{r_i}} \left[\sum_{o_j \in \text{sbj}(o_i)} g_s(e_{o_i}, e_{r_{ij}}, e_{o_j}) + \sum_{o_k \in \text{obj}(o_i)} g_s(e_{o_i}, e_{r_{ik}}, e_{o_k}) \right]_r$$

sbj 是指以 o_i 为关系主体的关系客体集合, obj 是指以 o_i 为关系客体的关系主体集合, N_{r_i} 是指与 o_i 相关的关系总数。通过对场景图生成的实体、属性、关系向量分别进行嵌入, 利用图卷积网络在嵌入过程中推理图中相邻节点之间的关联, 强化节点之间的语义关系和语义约束, 即关联推理过程。

为了更便于理解本发明设计的步骤 S3 多重模态循环融合的具体实施过程, 这里列出图 7, 是本发明提供的另一种基于多模态异质特征融合的视频密集事件描述方法的三重模态循环

融合模块示意图。

如图 7 所示, c_T^q 与 c_T^a 分别是动态视觉特征向量和音频韵律特征向量, 而 v_T 是特征图, 所以以 v_T 为基准进行投影, 如下公式所示:

$$V_q = c_T^q W_q^T$$

$$V_a = c_T^a W_a^T$$

$$V_v = W_v v_T$$

其中矩阵 $W_i, i \in \{1, 2, 3\}$ 的上标 T 代表转置操作, $W_q \in \mathbb{R}^{d \times a}$ 、 $W_a \in \mathbb{R}^{d \times b}$ 、 $W_v \in \mathbb{R}^c$, 映射后的特征向量 $V_q \in \mathbb{R}^d$ 、 $V_a \in \mathbb{R}^d$ 、 $V_v \in \mathbb{R}^d$ 。此时特征向量 V_q 、 V_a 、 V_v 都在同一维度上, 接着分别对其进行循环矩阵的改造, 具体做法是从第二行向量开始, 把每一个行向量都定义为上一个行向量将整体循环右移一个元素, 改造后的循环矩阵如下公式所示:

$$C_q = \text{Ciro}(V_q)$$

$$C_a = \text{Ciro}(V_a)$$

$$C_v = \text{Ciro}(V_v)$$

其中, $C_q \in \mathbb{R}^{d \times d}$ 、 $C_a \in \mathbb{R}^{d \times d}$ 、 $C_v \in \mathbb{R}^{d \times d}$ 分别为转化后的矩阵, $\text{Ciro}(V)$ 为循环矩阵改造函数。最后, 分别将每种特征向量与另外两种不同向量改造而成的循环矩阵做元素积, 以探索三种模态信息之间的充分交互, 如下公式所示:

$$F_q = \frac{1}{d} \sum_{i=1}^d f_i^q \odot C_a \odot C_v$$

$$F_a = \frac{1}{d} \sum_{i=1}^d f_i^a \odot C_q \odot C_v$$

$$F_v = \frac{1}{d} \sum_{i=1}^d f_i^v \odot C_q \odot C_a$$

其中, f_i 为对应模态特征的循环矩阵的第 i 行特征向量, $\odot$ 为元素积操作, $F_q \in \mathbb{R}^d$ 、 $F_a \in \mathbb{R}^d$ 、 $F_v \in \mathbb{R}^d$ 为交融后的特征向量。

在本实施例中, 本发明还提供了另一种基于多模态异质特征融合的视频密集事件描述方法的语言字典训练示意图。

如图 8 所示, 为了避免生成的描述停留在浅显的共现描述, 如: 图 8 中有两个人, 而无法描述出此二人的关系, 活动等; 需要将语言先验知识提炼成一个知识字典, 该字典能够与视觉输入信息进行融合计算, 并指导模型给出更具深层理解的有意义描述。为获取该字典, 本实例中设计了一个自监督的代理任务: 将已标注的文本描述 S 转换成场景图 G , 其目的是为了从文本中提取实体、属性和关系。接着将得到的向量与预训练的语言字典 D (D 是一个维度为 $d \times L$ 的矩阵) 进行计算对特征进行重编码 ψ , 再将重编码后的向量送入解码器重构文

本描述，通过比较重构后的描述与标注描述的差异来监督模型学习。其中重编码 $\varphi(\cdot)$ 的计算过程如下：

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

式中 d_l 表示 D 中第 l 列向量，此处 D 的作用就是将人类语言习惯进行编码嵌入。对语言知识提取模型进行训练实际上就是在训练 D 中的参数。

用于比较重构后的描述与标注描述的差异来监督模型学习的损失函数如下：

$$\text{loss} = \mathbb{E}_{S^s \sim p(s)} [\text{diff}(S^s; S)]$$

其中 S^s 表示模型输出的描述结果，是一句句子， S 为人为标注的数据集标签， diff 表示衡量模型输出与标签之间差距的算法，在本实例中所使用的是 CIDEr-D 度量算法。

本发明还提供一种电子装置，其中，该装置包括存储器、处理器，所述存储器上存储有可在所述处理器上运行本发明所述装置的配置程序系统，所述配置程序被所述处理器执行时可以实现本发明所述视频密集事件描述方法。

本发明还提供一种计算机可读存储介质，其中，所述计算机可读的存储介质上存储有本发明所述装置的配置程序，所述配置程序可以被一个或多个处理器执行，以实现如本发明所述视频密集事件描述方法。

以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

权利要求书

1、一种基于多模态异质特征融合的视频密集事件描述方法,其特征在于,包括以下步骤:

特征提取,利用 I3D 卷积网络选取开放域中长视频进行剪辑得到对应的视频帧,并对所述视频帧中的动态视觉特征 F_m 进行提取,获取动态视觉特征向量;利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量;

场景图特征嵌入,将视频帧中的场景对象信息进行语义化表征,生成场景图谱,得到实体编码,属性编码和关系编码,通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入,得到场景图特征嵌入向量;

三重多模态循环融合,将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入,并将其进行三重多模态循环融合,再将输出的向量使用规范化函数沿通道做归一化,并且利用 Dropout 进行正则化后输入到解码器中,并为重构解码器做准备;

描述重构解码器,包括预训练的语言字典解码器以及描述重构算法,通过描述重构算法对视频密集事件进行检测;并利用预训练的语言字典解码器来解码生成视频场景事件的描述;所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

2、根据权利要求 1 所述的视频密集事件描述方法,其特征还在于,

所述动态视觉特征 F_m ,其中 $F_m = [f_1^m, f_2^m, \dots, f_N^m]$,其中 N 为抽取的视频剪辑的数量,整体对应的维度是 $T_v \times d_v$;

所述音频韵律特征向量对应的维度是 $T_a \times d_a$,音频特征向量被转化为具有语义的 128 维的特征向量。

3、根据权利要求 1 或 2 所述的视频密集事件描述方法,其特征还在于,

在场景图特征嵌入过程中,输入为视频帧,首先经过 Mask R-CNN 检测视频帧图片中的物体包围框坐标等物体位置信息,确定物体目标所在位置,得到物体区域;再将所述视频帧原图和所述物体区域分别送入视觉场景图生成器生成场景图,将所述视频帧图片信息以及被检测出来的物体位置信息等视觉信息进行语义化表征,得到实体编码、属性编码和关系编码,之后将这些编码送入联合嵌入层进行联合表征得到语义信息,将所述语义信息嵌入所述实体编码、属性编码和关系编码中对应的特征向量;再通过图卷积对特征向量进行图嵌入,得到场景图特征嵌入向量;在此过程中对实体之间的关联进行强化推理。

4、根据权利要求 1 所述的视频密集事件描述方法,其特征还在于,

在三重多模态循环融合之后,还包括:

自适应多模态数据平衡,利用全卷积网络对所述动态视觉特征向量和所述音频韵律特征向量在时序维度上进行特征提取,并使用特定值来预测事件的时序边界以及置信度分数;其次,将预测得到的事件送入公共池中,并按照降序排列并选取高置信度分数的事件;最终对

所述高置信度分数的事件进行聚类，选择各个对应的聚类中心来组成最终的事件集合。

5、根据权利要求4所述的视频密集事件描述方法，其特征还在于，

所述特定值是指中心位置 $center = p + \sigma(c)$ ，其中 $\sigma(\cdot)$ 是 *sigmoid* 函数，作用是把 c 的值缩放到 $[0, 1]$ 区间内，用来确定中心点 p 的右邻域；事件时长 $length = anchor \cdot e$ ，其中 e 是缩放系数；置信度分数 $confidence = \sigma(o)$ 。

6、根据权利要求1或4或5所述的视频密集事件描述方法，其特征还在于，

所述全卷积网络皆采用一维卷积对时序维度上进行特征提取，第一层的卷积核大小为 k ，第二层和第三层的卷积核都为1，前两层的激活函数是 *Relu*，最后一层的激活函数是 *sigmoid*。

7、根据权利要求1或4所述的视频密集事件描述方法，其特征还在于，

在三重多模态循环融合过程中，所述规范化函数指是 *Softmax* 函数。

8、根据权利要求1或4所述的视频密集事件描述方法，其特征还在于，在三重多模态循环融合过程中，维度为 $T_v \times d_v$ 以及 $T_a \times d_a$ 的动态视觉特征向量和音频特征向量先被送入自注意力层进行特征表示。

9、根据权利要求4所述的视频密集事件描述方法，其特征还在于，所述高置信度分数取前100的事件，聚类算法采用根据欧式距离的 *K-Means* 算法。

10、根据权利要求1所述的视频密集事件描述方法，其特征还在于，所述语言字典的预训练方法为，将已标注的描述文本转换为场景图，再通过图卷积网络进行场景图特征嵌入，将得到的向量与所述预训练的语言字典进行计算对场景图特征进行重编码，再将重编码后的向量送入解码器重构后进行文本描述，通过比较重构后的文本描述与所述已标注的描述文本的差异来监督模型学习；所述标注描述文本是存于数据集中，用于预训练的数据；

所述预训练的语言字典，是指预训练的语言字典 D 是一个维度为 $d \times L$ 的矩阵；

所述重编码，是指进行重编码 $\varphi(\cdot)$ 的计算，过程如下：

$$\hat{x} = \varphi(x, D) = \sum_{l=1}^L \text{softmax}(D^T x_l) d_l$$

其中 d_l 表示 D 中第 l 列向量，此处 D 的作用就是将人类语言习惯进行编码嵌入。

11、根据权利要求1或10所述的视频密集事件描述方法，其特征还在于，在所述描述重构解码器步骤中，所述的描述重构算法具体为：

描述重构算法

输入：视频的开始时间 P 和结束时间 Q ；事件集合 E ，每个事件包含属性：事件的中心时间点 m ，事件时长 l ，置信度分数 s ，事件对应的描述 z

输出：候选描述集合 Z

1 根据 m 和 l 计算出每个事件的开始时间 s 和结束时间 e ：

$$s = m - \frac{l}{2}, e = m + \frac{l}{2};$$

2 计算出事件之间的 *tiou*，将重叠事件去重；

-
- 3 将事件按照开始时间升序排列并选择开始时间最早的事件作为基准事件 k ;
 - 4 从剩余事件集合 $E - k$ 中选择开始时间在 $[k_s, k_s + \varepsilon]$ 中的事件一同组成起始事件集合 E_1 ;
 - 5 从剩余事件集合 $E - E_1$ 中选择开始时间在 $[E_{1_e} - \varepsilon, E_{1_e} + \varepsilon]$ 中的事件与 E_1 中对应的事件组成事件序列集合 $\hat{E}$;
 - 6 重复步骤 5 并更新 $\hat{E}$ 直到选择的事件的结束时间在 $[Q - \varepsilon, Q]$ 范围内, 更新 $\hat{E}$, 选择结束;
 - 7 将 $\hat{E}$ 中每个事件序列对应的文本描述进行合并后加入到候选描述集合 Z 。
-

12、一种兼顾参数量与交互强度的融合方法, 其特征在于, 可以用于实现权利要求 1 所述的三重多模态循环融合, 在不产生高维特征的前提下, 让多种特征进行深层次融合。

13、一种基于多模态异质特征融合的视频密集事件描述的装置, 其特征在于,

特征提取模块, 用于利用 I3D 卷积网络选取开放域中长视频进行剪辑得到对应的视频帧, 并对所述视频帧中的动态视觉特征 F_m 进行提取, 获取动态视觉特征向量; 利用 VGGish 模型来提取所述视频片段中的音频韵律特征向量;

场景图特征嵌入模块, 用于将视频帧中的场景对象信息进行语义化表征, 生成场景图谱, 得到实体编码, 属性编码和关系编码, 通过图卷积对所述实体编码、属性编码、关系编码中对应的特征向量进行图嵌入, 得到场景图特征嵌入向量;

三重多模态循环融合模块, 用于将场景图特征嵌入向量、所述动态视觉特征向量、所述音频特征向量作为输入, 并将其进行三重多模态循环融合, 再将输出的向量使用规范化函数沿通道做归一化, 并且利用 Dropout 进行正则化后输入到解码器中, 并为重构解码器做准备; 所述三重多模态循环融合模块包括预先建立的一种兼顾参数量与交互强度的融合方法, 在不产生高维特征的前提下, 让多种特征进行深层次融合;

描述重构解码器模块, 用于通过描述重构算法对视频密集事件进行检测; 并利用预训练的语言字典解码器来解码生成视频场景事件的描述, 所述预训练的语言字典解码器是指利用预训练的语言字典来对视频场景事件的描述进行解码的模块。

14、根据权利要求 13 所述的视频密集事件描述的装置, 其特征在于,

还包括自适应多模态数据平衡模块, 用于利用全卷积网络对所述动态视觉特征向量和所述音频韵律特征向量在时序维度上进行特征提取, 并使用特定值来预测事件的时序边界以及置信度分数; 其次, 将预测得到的事件送入公共池中, 并按照降序排列并选取高置信度分数的事件; 最终对所述高置信度分数的事件进行聚类, 选择各个对应的聚类中心来组成最终的事件集合。

15、一种电子装置, 其特征在于, 该装置包括存储器、处理器, 所述存储器上存储有可在所述处理器上运行如权利要求13-14所述装置的配置程序系统, 所述配置程序被所述处理器执行时可以实现如权利要求1-11所述视频密集事件描述方法。

16、一种计算机可读存储介质, 其特征在于, 所述计算机可读的存储介质上存储有如权利要求13-14所述装置的配置程序, 所述配置程序可以被一个或多个处理器执行, 以实现如权

利要求1-11所述视频密集事件描述方法。