

检索方法及装置、存储介质

相关申请的交叉引用

本公开基于申请号为 201910934892.5、申请日为 2019 年 09 月 29 日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本公开作为参考。

5 技术领域

本公开涉及计算机视觉技术领域，具体涉及一种检索方法及装置、存储介质。

背景技术

在现实生活中，根据一段文字描述，在视频数据库中检索符合文字描述的视频这项功能有着广泛的需求。传统的检索方法通常将文字编码为词向量，同时将视频编码成视频特征向量。

10 发明内容

本公开提供一种检索方法的技术方案。

根据本公开的第一方面，提供了一种检索方法，所述方法包括：确定文本和至少一个视频之间的第一相似度，所述文本用于表征检索条件；确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图；确定所述第一人物互动图和所述第二人物互动图之间的第二相似度；根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频。

如此，相对于传统的基于特征的检索算法，本公开通过确定文本和至少一个视频之间的第一相似度，所述文本的第一人物互动图和所述至少一个视频的第二人物互动图之间的第二相似度，可以利用文字本身的语法结构以及视频本身的事件结构等信息，进行视频检索，从而能提高根据文本描述检索视频如电影的准确率。

20 在一种可能的实现方式中，所述确定文本和至少一个视频之间的第一相似度，包括：确定所述文本的段落特征；确定所述至少一个视频的视频特征；根据所述文本的段落特征和所述至少一个视频的视频特征，确定所述文本和所述至少一个视频之间的第一相似度。

如此，通过分析文本的段落特征和视频的视频特征来确定第一相似度，可得到视频和文本直接匹配的相似度，为后续确定与检索条件相匹配的视频提供参考依据。

25 在一种可能的实现方式中，所述段落特征包括句子特征和句子的数量；所述视频特征包括镜头特征和镜头的数量。

如此，通过将句子特征和句子的数量作为文本的段落特征，将镜头特征和镜头的数量作为视频的视频特征，对文本和视频进行了量化，进而能够为分析文本的段落特征和视频的视频特征提供分析依据。

在一种可能的实现方式中，所述确定所述文本的第一人物互动图，包括：检测所述文本中包含的人名；在数据库中搜索到所述人名对应的人物的肖像，并提取所述肖像的图像特征，得到所述人物的角色节点；解析确定所述文本的语义树，基于所述语义树得到所述人物的运动特征，得到所述人物的动作节点；将每个所述人物对应的角色节点和动作节点连接；其中，所述人物的角色节点用肖像的图像特征表征；所述人物的动作节点采用语义树中的运动特征表征。

如此，文本中的句子通常遵循与事件中的情景相似的顺序，每一段文本都描述了视频中的一个事件，通过构建文本的人物交互图来捕捉视频的叙事结构，为后续确定与检索条件相匹配的视频提供参考依据。

在一种可能的实现方式中，所述方法还包括：将连接同一动作节点的角色节点相互连接。

10 如此，有助于更好地构建文本的人物交互图，进而更好地捕捉视频的叙事结构。

在一种可能的实现方式中，所述检测所述文本中包含的人名，包括：将所述文本中的代词替换为所述代词所代表的所述人名。

如此，防止漏掉文本中用非人名表示的人物，能够对文本中描述的所有人物进行分析，进而提高确定文本的人物互动图的准确率。

15 在一种可能的实现方式中，所述确定所述至少一个视频的第二人物互动图，包括：检测出所述至少一个视频的每个镜头中的人物；提取所述人物的人体特征与运动特征；将所述人物的人体特征附加到所述人物的角色节点上，将所述人物的运动特征附加到所述人物的动作节点上；将每个人物对应的角色节点和动作节点相连。

20 如此，由于人物之间的相互作用经常在文本中描述，角色之间的互动在视频故事中扮演着重要的角色，为了结合这一点，本公开提出了一个基于图表表示的人物交互图，通过确定视频的人物交互图和文本的人物交互图之间的相似度，为后续确定与检索条件相匹配的视频提供参考依据。

在一种可能的实现方式中，所述确定所述至少一个视频的第二人物互动图，还包括：将同时出现在一个镜头中的一组人物作为同组人物，将所述同组人物中的人物的角色节点两两相连。

如此，有助于更好地构建视频的人物交互图，进而更好地捕捉视频的叙事结构。

25 在一种可能的实现方式中，所述确定所述至少一个视频的第二人物互动图，还包括：将一个镜头中的一位人物和其相邻镜头的每个人物的角色节点都相连。

如此，有助于更好地构建视频的人物交互图，进而更好地捕捉视频的叙事结构。

30 在一种可能的实现方式中，所述根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频，包括：对每个视频的所述第一相似度和所述第二相似度加权求和，得到每个视频的相似度值；将相似度值最高的视频，确定为与所述检索条件相匹配的视频。

如此，结合第一相似度和第二相似度来确定与检索条件相匹配的视频，能提高根据文本描述检索视频的准确率。

在一种可能的实现方式中，所述检索方法通过检索网络实现，所述方法还包括：确定文本和训练样本集中的视频之间的第一相似度预测值，所述文本用于表征检索条件；确定所述文本的第一人

物互动图和所述训练样本集中的视频的第二人物互动图之间的第二相似度；根据所述第一相似度预测值与所述第一相似度真值确定所述第一相似度的损失；根据所述第二相似度预测值与所述第二相似度真值确定所述第二相似度的损失；根据所述第一相似度的损失以及所述第二相似度的损失，结合损失函数确定总损失值；根据所述总损失值调整所述检索网络的权重参数。

5 如此，通过检索网络实现检索，有助于快速检索出与文本描述相匹配的视频。

在一种可能的实现方式中，所述检索网络包括第一子网络以及第二子网络；所述第一子网络用于确定文本与视频的第一相似度，所述第二子网络用于确定所述文本的第一人物互动图和所述视频的第二人物互动图之间的相似度；所述根据所述总损失值调整所述检索网络的权重参数，包括：基于所述总损失值调整所述第一子网络以及所述第二子网络的权重参数。

10 如此，通过不同的子网络分别确定不同的相似度，有助于快速得到与检索条件相关的第一相似度和第二相似度，进而能够快速检索出与检索条件相适应的视频。

根据本公开的第二方面，提供了一种检索装置，所述装置包括：第一确定模块，被配置为确定文本和至少一个视频之间的第一相似度，所述文本用于表征检索条件；第二确定模块，被配置为确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图；确定所述第一人物互动图和
15 所述第二人物互动图之间的第二相似度；处理模块，被配置为根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频。

在一种可能的实现方式中，所述第一确定模块，被配置为：确定所述文本的段落特征；确定所述至少一个视频的视频特征；根据所述文本的段落特征和所述至少一个视频的视频特征，确定所述文本和所述至少一个视频之间的第一相似度。

20 在一种可能的实现方式中，所述段落特征包括句子特征和句子的数量；所述视频特征包括镜头特征和镜头的数量。

在一种可能的实现方式中，所述第二确定模块，被配置为：检测所述文本中包含的人名；在数据库中搜索到所述人名对应的人物的肖像，并提取所述肖像的图像特征，得到所述人物的角色节点；解析确定所述文本的语义树，基于所述语义树得到所述人物的运动特征，得到所述人物的动作节点；
25 将每个所述人物对应的角色节点和动作节点连接；其中，所述人物的角色节点用肖像的图像特征表征；所述人物的动作节点采用语义树中的运动特征表征。

在一种可能的实现方式中，所述第二确定模块，还被配置为：将连接同一动作节点的角色节点相互连接。

在一种可能的实现方式中，所述第二确定模块，被配置为：将所述文本中的代词替换为所述代
30 词所代表的所述人名。

在一种可能的实现方式中，所述第二确定模块，被配置为：检测出所述至少一个视频的每个镜头中的人物；提取所述人物的人体特征与运动特征；将所述人物的人体特征附加到所述人物的角色节点上，将所述人物的运动特征附加到所述人物的动作节点上；将每个人物对应的角色节点和动作节点相连。

在一种可能的实现方式中，所述第二确定模块，还被配置为：将同时出现在一个镜头中的一组人物作为同组人物，将所述同组人物中的人物的角色节点两两相连。

在一种可能的实现方式中，所述第二确定模块，还被配置为：将一个镜头中的一位人物和其相邻镜头的每个人物的角色节点都相连。

- 5 在一种可能的实现方式中，所述处理模块，被配置为：对每个视频的所述第一相似度和所述第二相似度加权求和，得到每个视频的相似度值；将相似度值最高的视频，确定为与所述检索条件相匹配的视频。

- 在一种可能的实现方式中，所述检索装置通过检索网络实现，所述装置还包括：训练模块，被配置为：确定文本和训练样本集中的视频之间的第一相似度预测值，所述文本用于表征检索条件；
- 10 确定所述文本的第一人物互动图和所述训练样本集中的视频的第二人物互动图之间的第二相似度；根据所述第一相似度预测值与所述第一相似度真值确定所述第一相似度的损失；根据所述第二相似度预测值与所述第二相似度真值确定所述第二相似度的损失；根据所述第一相似度的损失以及所述第二相似度的损失，结合损失函数确定总损失值；根据所述总损失值调整所述检索网络的权重参数。

- 在一种可能的实现方式中，所述检索网络包括第一子网络以及第二子网络；所述第一子网络用于确定文本与视频的第一相似度，所述第二子网络用于确定文本的第一人物互动图和所述视频的第二人物互动图之间的相似度；所述训练模块，被配置为：基于所述总损失值调整所述第一子网络以及所述第二子网络的权重参数。
- 15

- 根据本公开的第三方面，提供了一种检索装置，所述装置包括：存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现本公开实施例所述的检索
- 20 方法的步骤。

根据本公开的第四方面，提供了一种存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行本公开实施例所述的检索方法的步骤。

根据本公开的第五方面，提供了一种计算机程序，包括计算机可读代码，当所述计算机可读代码在电子设备中运行时，所述电子设备中的处理器执行用于实现本公开实施例所述的检索方法。

- 25 本公开提供的技术方案，确定文本和至少一个视频之间的第一相似度，所述文本用于表征检索条件；确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图；确定所述第一人物互动图和所述第二人物互动图之间的第二相似度；根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频。如此，相对于传统的基于特征的检索算法，本公开通过确定文本和至少一个视频之间的第一相似度，所述文本的第一人物互动图和所述至少一个视频的第二人物互动图之间的第二相似度，可以利用文字本身的语法结构以及视频本身的事件结构等信息，进行视频检索，从而能提高根据文本描述检索视频如电影的准确率。
- 30

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，这些附图示出了符合本公开的实施例，

并与说明书一起用于说明本公开的技术方案。

图 1 是根据一示例性实施例示出的检索方法概述框架示意图；

图 2 是根据一示例性实施例示出的一种检索方法的实现流程示意图；

图 3 是根据一示例性实施例示出的一种检索装置的组成结构示意图。

5 具体实施方式

这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本公开实施例相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本公开实施例的一些方面相一致的装置和方法的例子。

10 在本公开实施例使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本公开实施例。在本公开实施例和所附权利要求书中所使用的单数形式的“一种”、“一个”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

应当理解，尽管在本公开实施例可能采用术语第一、第二、第三等来描述各种信息，但这些信
15 息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本公开实施例范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”及“若”可以被解释成为“在...时”或“当...时”或“响应于确定”。

下面结合附图和具体实施例对本公开的检索方法进行详细阐述。

20 图 1 是根据一示例性实施例示出的检索方法概述框架示意图，该框架用于匹配视频和文本，如匹配电影片段和剧情片段。该框架包括两类模块：事件流模块（EFM，Event Flow Module）和人物交互模块（CIM，Character Interaction Module）；事件流模块被配置为探索事件流的事件结构，以段落特征和视频特征为输入，输出视频和段落直接的相似度；人物交互模块被配置为利用人物交互，分别构建段落中的人物互动图和视频中的人物互动图，再通过图匹配算法衡量二图之间的相似度。

25 给定一个查询文本P和一个候选视频Q，上述两个模块分别产生P和Q之间的相似度得分，分别表示为 $S_{efm}(P, Q)$ 和 $S_{cim}(P, Q)$ 。然后将总匹配分数 $S(P, Q)$ 定义为它们的和：

$$S(P, Q) = S_{efm}(P, Q) + S_{cim}(P, Q) \quad \text{式 (1)}$$

具体如何求解将 $S_{efm}(P, Q)$ 和 $S_{cim}(P, Q)$ 在下文中详细描述。

当然，在其他实施例中，总匹配分数也可以是上述两个模块得分的加权和等运算结果。

30 本公开实施例提供一种检索方法，此检索方法可应用于终端设备、服务器或其他电子设备。其中，终端设备可以为用户设备（UE，User Equipment）、移动设备、蜂窝电话、无绳电话、个人数字处理（PDA，Personal Digital Assistant）、手持设备、计算设备、车载设备、可穿戴设备等。在一些可能的实现方式中，该处理方法可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。

如图 2 所示,所述方法主要包括:

步骤 S101、确定文本和至少一个视频之间的第一相似度,所述文本用于表征检索条件。

这里,所述文本是用于表征检索条件的一段文字描述。本公开实施例对获取文本的方式不作限定。例如,电子设备可以接收用户在输入区输入的文字描述,或者,接收用户在语音输入,然后将语音数据转换成文字描述。

这里,所述检索条件包括人名和至少一个表征动作的动词。例如,杰克打了他自己一拳。

这里,所述至少一个视频位于可供检索的本地或第三方视频数据库中。

这里,所述第一相似度是表征视频和文本直接匹配的相似度。

在一个例子中,电子设备将文本的段落特征和视频的视频特征输入到事件流模块,由事件流模块输出视频和文本的相似度,即第一相似度。

在一些可选实现方式中,所述确定文本和至少一个视频之间的第一相似度,包括:

确定所述文本的段落特征,所述段落特征包括句子特征和句子的数量;

确定所述至少一个视频的视频特征,所述视频特征包括镜头特征和镜头的数量;

根据所述文本的段落特征和所述至少一个视频的视频特征,确定所述文本和所述至少一个视频之间的第一相似度。

在一些例子中,确定文本的段落特征,包括:可以利用第一神经网络对文本进行处理,得到文本的段落特征,所述段落特征包括句子特征和句子的数量。例如,每个单词对应一个 300 维的向量,将句子中每个单词的特征加起来就是句子的特征。句子数量是指文本中的句号的数量,将输入的文本用句号将句子分割开,得到句子的数量。

20 在一些例子中,确定视频的视频特征,包括:可以利用第二神经网络对视频进行处理,具体地,先将视频解码成图片流,然后基于图片流得到视频特征;所述视频特征包括镜头特征和镜头的数量。例如,镜头特征是将镜头的 3 张关键帧的图片通过神经网络得到 3 个 2348 维的向量,再取平均。一个镜头是指视频中同一摄像机在同一机位拍摄的连续画面,如果画面切换则是另一个镜头,按照现有的镜头切割算法来得到镜头的数量。

25 如此,通过分析文本的段落特征和视频的视频特征来确定第一相似度,为后续确定出与检索条件相匹配的视频提供依据;利用文字本身的语法结构以及视频本身的事件结构等信息,进行视频检索,从而能提高根据文本描述检索视频的准确率。

上述方案中,可选地,所述第一相似度的计算公式为:

$$S_{efm} = \sum_i \sum_j y_{ij} \phi_j^T \psi_i = \text{tr}(\Phi \Psi^T Y) \quad \text{式 (2)}$$

30 其中,一个段落特征由 M 个句子特征组成,设句子特征为 $\phi_i \in \mathbb{R}^D$,则段落特征表示为 $\Phi = [\phi_1, \dots, \phi_M]^T$; 一个视频特征由 N 个镜头特征组成,设镜头特征为 $\psi_i \in \mathbb{R}^D$,则视频特征表示为 $\Psi = [\psi_1, \dots, \psi_N]^T$; 设布尔分配矩阵 $Y \in \{0, 1\}^{N \times M}$, 用于将每个镜头分配给每个句子, 其中 $y_{ij} = Y(i, j) = 1$ 代表第 i 个镜头被分配给第 j 个句子, $y_{ij} = Y(i, j) = 0$ 代表第 i 个镜头未被分配给第

j 个句子。

上述方案中，可选地，所述第一相似度的计算公式的约束条件包括：

每个镜头最多被分配给 1 个句子；

序号靠前的镜头被分配到的句子，相对于序号在后的镜头被分配到的句子，更靠前。

- 5 因此，可将计算第一相似度转化为求解如下公式（3）的优化目标，将优化目标和约束条件联合起来，可以得到如下优化公式：

$$\max_Y \text{tr}(\Phi \Psi^T Y) \quad \text{式 (3)}$$

$$\text{s.t. } Y1 \leq 1 \quad \text{式 (4)}$$

$$J(y_i) \leq J(y_{i+1}), \forall i \leq N-1 \quad \text{式 (5)}$$

- 10 其中，公式（3）是优化目标；s.t.是 such that 的缩写，引出表示公式（3）约束条件的公式（4）和（5）； y_i 表示 Y 的第 i 行向量， $J(\cdot)$ 表示一个布尔向量的第一个非零值的序号。公式（4）中，Y 是一个矩阵，1 是一个向量（所有元素都是 1 的向量），Y1 是矩阵 Y 和向量 1 的乘积。

进一步地，通过传统的动态规划算法，可以得到该优化问题的解。具体地，通过动态规划算法相关算法，可以解得最优的 Y，从而得到 S_{efm} 的值。

- 15 在其他实施例，也可以对段落特征和视频特征进行其他类型的计算，例如多个段落特征和对应的多个视频特征进行加权或比例运算等，得到所述第一相似度。

步骤 S102、确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图。

这里，人物互动图是用于表征人物之间的角色关系和动作关系的图，包括角色节点和动作节点。

- 20 在一些可选实施方式中，一个文本对应一个第一人物互动图，一个视频对应一个第二人物互动图。

在一些可选实施方式中，所述确定所述文本的第一人物互动图，包括：检测所述文本中包含的人名；在数据库中搜索到所述人名对应的人物的肖像，并提取所述肖像的图像特征，得到所述人物的角色节点；解析确定所述文本的语义树，基于所述语义树得到所述人物的运动特征，得到所述人物的动作节点；将每个所述人物对应的角色节点和动作节点连接。

- 25 其中，数据库是预先存储有大量的人名和肖像的对应关系的库，所述肖像是与该人名对应的人物的肖像。肖像数据可从网络上爬取，如可从 imdb 网站和 tmdb 网站上爬取到肖像数据。其中，所述人物的角色节点用肖像的图像特征表征；所述人物的动作节点采用语义树中的运动特征表征。

- 30 在一些实施例，解析确定所述文本的语义树，包括：通过依存句法算法解析确定文本的语义树。例如，利用依存句法算法将每句话分成一个一个的词，然后根据语言学的一些规则，把词作为节点，建一棵语义树。

先将每个句子得到一个图，然后每一段有多个句子，就是多个图。但是，在数学上，我们可以把这几个图看成一个图（一个非连接图）。也就是说，在数学上图的定义不一定要每个节点到另一个节点都有路径可以达到的，也可以是那种可分割成几个小图的图。

其中，如果多个人名指向同一个动作节点，则将所述多个人名的动作节点两两之间用边连接。

其中，边连接的两个节点特征拼接作为边的特征。

示例性地，可将边连接的两个节点特征分别表示为两个向量，将该两个向量进行拼接（例如维度相加），则得到边的特征。比如一个向量3维，另一个向量4维度，直接拼接成7维的向量。举例来说，若将[1,3,4]和[2,5,3,6]拼接，则拼接的结果是[1,3,4,2,5,3,6]。

- 5 在一些例子中，可以采用 Word2Vec 词向量经神经网络处理后的特征作为动作节点的代表，即作为人物的运动特征。

10 在一些例子中，检测文本中包含的人名时，将文本中的代词替换为所述代词所代表的人名。具体地，通过人名检测工具（如斯坦福人名检测工具包）检测出所有的人名（如“杰克”）。之后通过共指解析工具将代词替换成该词所代表的人名（如“杰克打了他自己一拳”中的“他”提取为“杰克”）。

15 在一些实施例，基于人名在数据库中搜索到所述人名对应的人物的肖像，并通过神经网络提取所述肖像的图像特征；其中，所述图像特征包括人脸和身体特征。通过神经网络确定所述文本中每个句子的语义树以及所述语义树上每个词的词性，如名词、代词、动词等，所述语义树上每个节点是所述句子中的一个词，将句子中的动词作为人物的运动特征，即动作节点，将名词或代词对应的人名作为人物角色节点，将人物的肖像的图像特征附加到人物角色节点；根据所述语义树和所述人名，将每个所述人名对应的角色节点和所述人名的动作节点连接，如果多个人名指向同一个动作节点，则所述多个人名两两之间用边连接。

在一些可选实施方式中，所述确定所述至少一个视频的第二人物互动图，包括：

检测出所述至少一个视频的每个镜头中的人物；

- 20 提取所述人物的人体特征与运动特征；

将所述人物的人体特征附加到所述人物的角色节点上，将所述人物的运动特征附加到所述人物的运动节点上；

将每个人物对应的角色节点和运动节点相连。

- 25 这里，一个镜头是指视频中同一摄像机在同一机位拍摄的连续画面，如果画面切换则是另一个镜头，按照现有的镜头切割算法来得到镜头的数量。

这里，所述人体特征是人物的人脸和身体特征，将镜头对应的图像通过训练好的模型可以得到图像中的人物的人体特征。

这里，所述运动特征是将镜头对应的图像输入训练好的模型得到的图像中的人物的运动特征，例如识别得到的人物在当前图像中的动作（如喝水）。

- 30 进一步地，所述确定所述至少一个视频的第二人物互动图时，还包括：如果一组人物同时出现在一个镜头中，则将同组人物中的人物的角色节点两两相连；将一个镜头中的一位人物和其相邻镜头的每个人物的角色节点都相连。

这里，所述相邻镜头是指当前镜头的前一个镜头和后一个镜头。

其中，如果多个角色节点指向同一个动作节点，则将所述多个角色节点的动作节点两两之间用

边连接。

其中，边连接的两个节点特征拼接作为边的特征。

上述边特征的确定过程可参考第一人物互动图中边特征的确定方法，此处不再赘述。

步骤 S103、确定所述第一人物互动图和所述第二人物互动图之间的第二相似度。

- 5 这里，所述第二相似度是表征第一人物互动图和第二人物互动图二图进行匹配计算得到的相似度。

在一个例子中，电子设备将文本和视频输入到人物互动模块，由人物互动模块构建文本中的第一人物互动图和视频中的第二人物互动图，再通过图匹配算法衡量二图之间的相似度，输出该相似度，即第二相似度。

- 10 在一些可选实施方式中，所述第二相似度的计算公式为：

$$S_{cim}(P, Q) = \sum_{i,a} u_{ia} k_{ia;ia} + \sum_{i \neq j} \sum_{a \neq b} u_{ia} u_{jb} k_{ia;jb} \quad \text{式 (6)}$$

- 其中， u 是二值向量(布尔向量)， $u_{ia} = 1$ 代表 V_p 里第 i 个节点和 V_q 里第 a 个节点能匹配上， $u_{ia} = 0$ 代表 V_p 里第 i 个节点和 V_q 里第 a 个节点不能匹配上。同理， $u_{jb} = 1$ 代表 V_p 里第 j 个节点和 V_q 里第 b 个节点能匹配上， $u_{jb} = 0$ 代表 V_p 里第 j 个节点和 V_q 里第 b 个节点不能匹配上； i, a, j, b 都是索引符号； $k_{ia;ia}$ 代表 V_p 里第 i 个节点和 V_q 里第 a 个节点的相似度， $k_{ia;jb}$ 代表 E_p 里的边 (i, j) 和 E_q 里的边 (a, b) 的相似度。
- 15

设文本中的第一人物互动图为 $G_p = (V_p, E_p)$ ，其中， V_p 是节点的集合， E_p 是边的集合； V_p 由两种节点构成， V_p^a 为第一人物互动图中的动作节点， V_p^c 为第一人物互动图中的角色节点；

- 20 设视频中的第二人物互动图为 $G_q = (V_q, E_q)$ ；其中， V_q 是节点的集合， E_q 是边的集合； V_q 由两种节点构成， V_q^a 为第二人物互动图中的动作节点， V_q^c 为第一人物互动图中的角色节点；

$|V_p| = m = m_a + m_c$ ， m_a 为动作节点数量， m_c 为角色节点数量；

$|V_q| = n = n_a + n_c$ ， n_a 为动作节点数量， n_c 为角色节点数量；

- 给定布尔向量 $u \in \{0,1\}^{nm \times 1}$ ，如果 $u_{ia} = 1$ ，则代表 $i \in V_q$ 被匹配到 $a \in V_p$ ；相似度矩阵 $K \in \mathbb{R}^{nm \times nm}$ ，相似度矩阵 K 对角线元素为节点的相似度 $k_{ia;ia} = K(ia, ia)$ ，衡量 V_q 中第 i 个节点和 V_p 中第 a 个节点的相似度； $k_{ia;jb} = K(ia, jb)$ 衡量边 $(i, j) \in E_q$ 和边 $(a, b) \in E_p$ 的相似度，相似度由节点或边对应的特征，通过点积处理可得。
- 25

在一些可选实施方式中，所述第二相似度的计算公式的约束条件包括：

一个节点只能被匹配到另一个集合的最多一个节点；

不同类型的节点不能被匹配。

- 30 也就是说，匹配必须是一对一匹配，即一个节点之内被匹配到另一个集合的最多一个节点。不同类型的节点不能被匹配，比如角色节点不能被另一集合的动作节点所匹配。

因此，计算上述第二相似度可转化为求解如下优化公式 (7)，最终的优化公式和上述约束条件结合起来，可以得到：

$$\max_u u^T K u, \quad \text{式 (7)}$$

$$\text{s. t.} \quad \sum_i u_{ia} \leq 1 \quad \forall a, \quad \text{式 (8)}$$

$$\sum_a u_{ia} \leq 1 \quad \forall i, \quad \text{式 (9)}$$

$$\sum_{i \in V_q^c} u_{ia} = 0 \quad \forall a \in V_p^a, \quad \text{式 (10)}$$

$$\sum_{i \in V_q^a} u_{ia} = 0 \quad \forall a \in V_p^c, \quad \text{式 (11)}$$

在解优化的过程中，会得到 u ，将 u 带入公式 (7) 就能得到相似度。

在其他实施例中，也可以通过其他运算方式，例如对匹配的节点特征和动作特征进行加权平均等运算，得到所述第二相似度。

步骤 S104、根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频。

在一些可选实施方式中，所述根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频，包括：对每个视频的所述第一相似度和所述第二相似度加权求和，得到每个视频的相似度值；将相似度值最高的视频，确定为与所述检索条件相匹配的视频。

在一些实施例中，权重通过数据库中的验证集确定，在验证集上可以通过调权重方式，根据最终检索结果反馈得到一组最优的权重，进而可直接用到测试集上或直接用到实际检索中。

如此，利用文字本身的语法结构以及视频本身的事件结构等信息，进行视频检索，将相似度值最高的视频，确定为与所述检索条件相匹配的视频，能提高根据文本描述检索视频的准确率。

当然，在其他实施例中，也可以直接将第一相似度和第二相似度相加，得到每个视频对应的相似度。

上述方案中，所述检索方法通过检索网络实现，该检索网络的训练方法，包括：确定文本和训练样本集中的视频之间的第一相似度预测值，所述文本用于表征检索条件；确定所述文本的第一人物互动图和所述训练样本集中的视频的第二人物互动图之间的第二相似度；根据所述第一相似度预测值与所述第一相似度真值确定所述第一相似度的损失；根据所述第二相似度预测值与所述第二相似度真值确定所述第二相似度的损失；根据所述第一相似度的损失以及所述第二相似度的损失，结合损失函数确定总损失值；根据所述总损失值调整所述检索网络的权重参数。

本公开实施例中，所述检索网络对应的检索框架里有不同的组成模块，每个模块里可使用不同类型的神经网络。所述检索框架是事件流模块和人物关系模块共同组成的框架。

在一些可选实施方式中，所述检索网络包括第一子网络以及第二子网络；所述第一子网络用于确定文本与视频的第一相似度，所述第二子网络用于确定文本的第一人物互动图和所述视频的第二人物互动图之间的相似度。

具体地，将文本和视频输入第一子网络，该第一子网络输出文本与视频的第一相似度预测值；将文本和视频输入第二子网络，该第二子网络输出文本的第一人物互动图和所述视频的第二人物互动图之间的相似度预测值；根据标注的真值，能够得到文本与视频的第一相似度真值，以及所述文本的第一人物互动图和所述视频的第二人物互动图之间的相似度真值，根据第一相似度预测值和第

一相似度真值的差异,可得到第一相似度的损失;根据第二相似度预测值和第二相似度真值得差异,可得到第二相似度的损失;根据第一相似度的损失和第二相似度的损失,再结合损失函数调整第一子网络和第二自网络的网络参数。

5 在一个例子中,构建了一个数据集,它包含了 328 部电影的概要,以及概要段落和电影片段之间的注释关联。具体地,该数据集不仅为每部电影提供了高质量的详细概要,而且还通过手动注释将概要的各个段落与电影片段相关联;在这里,每个电影片段可以持续到每个分钟和捕获完整事件。这些电影片段,再加上相关的概要段落,可以让在更大的范围和更高的语义层次上进行分析。在这个数据集的基础上,本公开利用一个包括事件流模块和人物交互模块的框架来执行电影片段和概要段落之间的匹配。与传统的基于特征的匹配方法相比,该框架可显著提高匹配精度,同时也揭示了叙事结构和人物互动在电影理解中的重要性。

在一些可选实施方式中,所述根据所述总损失值调整所述检索网络的权重参数,包括:

基于所述总损失值调整所述第一子网络以及所述第二子网络的权重参数。

在一些可选实施方式中,所述损失函数表示为:

$$\mathcal{L} = \mathcal{L}(Y, \theta_{efm}, u, \theta_{cim}) \quad \text{式 (12)}$$

15 其中, θ_{efm} 表示在事件流模块中嵌入网络的模型参数, θ_{cim} 表示在人物交互模块中嵌入网络的模型参数。

其中, Y 是事件流模块定义的二值矩阵, u 是人物互动模块的二值向量,公式 (12) 表示通过最小化函数 $\mathcal{L}$ 来调整网络的参数,例如下面公式 (13) 所示得到新的网络参数 $\theta_{efm}^*, \theta_{cim}^*$ 。

$$\theta_{efm}^*, \theta_{cim}^* = \arg \min_{\theta_{efm}, \theta_{cim}} \mathcal{L}(Y^*, \theta_{efm}, u^*, \theta_{cim}) = \arg \min_{\theta_{efm}, \theta_{cim}} \mathcal{L}(S^*, \theta_{efm}, \theta_{cim}) \quad \text{式 (13)}$$

20 其中, $\mathcal{L}(S; \theta)$ 表示为:

$$\mathcal{L}(S; \theta) = \sum_i \sum_{j \neq i} \max(0, S(Q_j, P_i) - S(Q_i, P_i) + \alpha) + \sum_i \sum_{j \neq i} \max(0, S(Q_i, P_j) - S(Q_i, P_i) + \alpha) \quad \text{式 (14)}$$

其中, Y^* 是使得公式 (3) 的值最大的 Y , 也称之为最优解。

其中, u^* 是使得公式 (7) 最大的 u 。

25 其中, $S(Q_i, P_j)$ 表示第 i 个视频 Q_i 与第 j 个段落 P_j 的相似度; $S(Q_i, P_i)$ 表示第 i 个视频 Q_i 与第 i 个段落 P_i 的相似度, $S(Q_j, P_i)$ 表示第 j 个视频 Q_j 与第 i 个段落 P_i 的相似度; α 为损失函数的参数,表示最小相似度差值。

本公开所述技术方案可用于各种检索任务中,对检索场景不做限定,比如检测场景包括电影片段检索场景、电视剧片段检索场景、短视频检索场景等。

30 本公开实施例提出的检索方法,确定文本和至少一个视频之间的第一相似度,所述文本用于表征检索条件;确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图;确定所述第一人物互动图和所述第二人物互动图之间的第二相似度;根据所述第一相似度和所述第二相似度,从所述至少一个视频中确定出与所述检索条件相匹配的视频。如此,相对于传统的基于特征的检索

算法,本公开通过确定文本和至少一个视频之间的第一相似度,所述文本的第一人物互动图和所述至少一个视频的第二人物互动图之间的第二相似度,解决了传统的基于特征的检索算法没有利用文字本身的语法结构以及视频本身的事件结构等信息的问题,采用事件流匹配的方法和基于人物互动图匹配的方法进行视频检索,能提高根据文本描述检索视频的准确率。

5 对应上述检索方法,本公开实施例提供了一种检索装置,如图3所示,所述装置包括:第一确定模块10,用于被配置为文本和至少一个视频之间的第一相似度,所述文本用于表征检索条件;第二确定模块20,被配置为确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图;确定所述第一人物互动图和所述第二人物互动图之间的第二相似度;处理模块30,被配置为根据所述第一相似度和所述第二相似度,从所述至少一个视频中确定出与所述检索条件相匹配的视频。

10 在一些实施例中,所述第一确定模块10,被配置为:确定所述文本的段落特征;确定所述至少一个视频的视频特征;根据所述文本的段落特征和所述至少一个视频的视频特征,确定所述文本和所述至少一个视频之间的第一相似度。

在一些实施例中,所述段落特征包括句子特征和句子的数量;所述视频特征包括镜头特征和镜头的数量。

15 在一些实施例中,所述第二确定模块20,被配置为:检测所述文本中包含的人名;在数据库中搜索到所述人名对应的人物的肖像,并提取所述肖像的图像特征,得到所述人物的角色节点;解析确定所述文本的语义树,基于所述语义树得到所述人物的运动特征,得到所述人物的动作节点;将每个所述人物对应的角色节点和动作节点连接;其中,所述人物的角色节点用肖像的图像特征表征;所述人物的动作节点采用语义树中的运动特征表征。

20 在一些实施例中,所述第二确定模块20,还被配置为:将连接同一动作节点的角色节点相互连接。

在一些实施例中,所述第二确定模块20,被配置为:将所述文本中的代词替换为所述代词所代表的所述人名。

25 在一些实施例中,所述第二确定模块20,被配置为:检测出所述至少一个视频的每个镜头中的人物;提取所述人物的人体特征与运动特征;将所述人物的人体特征附加到所述人物的角色节点上,将所述人物的运动特征附加到所述人物的动作节点上;将每个人物对应的角色节点和动作节点相连。

在一些实施例中,所述第二确定模块20,还被配置为:将同时出现在一个镜头中的一组人物作为同组人物,将所述同组人物中的人物的角色节点两两相连。

30 在一些实施例中,所述第二确定模块20,还被配置为:将一个镜头中的一位人物和其相邻镜头的每个人物的角色节点都相连。

在一些实施例中,所述处理模块30,被配置为:对每个视频的所述第一相似度和所述第二相似度加权求和,得到每个视频的相似度值;将相似度值最高的视频,确定为与所述检索条件相匹配的视频。

在一些实施例中,所述检索装置通过检索网络实现,所述装置还包括:训练模块40,被配置为:

确定文本和训练样本集中的视频之间的第一相似度预测值, 所述文本用于表征检索条件; 确定所述文本的第一人物互动图和所述训练样本集中的视频的第二人物互动图之间的第二相似度; 根据所述第一相似度预测值与所述第一相似度真值确定所述第一相似度的损失; 根据所述第二相似度预测值与所述第二相似度真值确定所述第二相似度的损失; 根据所述第一相似度的损失以及所述第二相似度的损失, 结合损失函数确定总损失值; 根据所述总损失值调整所述检索网络的权重参数。

在一些实施例中, 所述检索网络包括第一子网络以及第二子网络; 所述第一子网络用于确定文本与视频的第一相似度, 所述第二子网络用于确定文本的第一人物互动图和所述视频的第二人物互动图之间的相似度; 所述训练模块 40, 被配置为: 基于所述总损失值调整所述第一子网络以及所述第二子网络的权重参数。

10 本领域技术人员应当理解, 图 3 中所示的检索装置中的各处理模块的实现功能可参照前述检索方法的相关描述而理解。本领域技术人员应当理解, 图 3 所示的检索装置中各处理单元的功能可通过运行于处理器上的程序而实现, 也可通过具体的逻辑电路而实现。

实际应用中, 上述第一确定模块 10、第二确定模块 20、处理模块 30 和训练模块 40 的具体结构均可对应于处理器。所述处理器具体的结构可以为中央处理器 (CPU, Central Processing Unit)、微处理器 (MCU, Micro Controller Unit)、数字信号处理器 (DSP, Digital Signal Processing) 或可编程逻辑器件 (PLC, Programmable Logic Controller) 等具有处理功能的电子元器件或电子元器件的集合。其中, 所述处理器包括可执行代码, 所述可执行代码存储在存储介质中, 所述处理器可以通过总线等通信接口与所述存储介质中相连, 在执行具体的各单元的对应功能时, 从所述存储介质中读取并运行所述可执行代码。所述存储介质用于存储所述可执行代码的部分优选为非瞬间存储介质。

20 本公开实施例提供的检索装置, 能提高根据文本检索视频的准确率。

本公开实施例还记载了一种检索装置, 所述装置包括: 存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序, 所述处理器执行所述程序时实现前述任意一个技术方案提供的检索方法。

作为一种实施方式, 所述处理器执行所述程序时实现: 确定文本和至少一个视频之间的第一相似度, 所述文本用于表征检索条件; 确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图; 确定所述第一人物互动图和所述第二人物互动图之间的第二相似度; 根据所述第一相似度和所述第二相似度, 从所述至少一个视频中确定出与所述检索条件相匹配的视频。

作为一种实施方式, 所述处理器执行所述程序时实现: 所述确定文本和至少一个视频之间的第一相似度, 包括: 确定所述文本的段落特征; 确定所述至少一个视频的视频特征; 根据所述文本的段落特征和所述至少一个视频的视频特征, 确定所述文本和所述至少一个视频之间的第一相似度。

作为一种实施方式, 所述处理器执行所述程序时实现: 检测所述文本中包含的人名; 在数据库中搜索到所述人名对应的人物的肖像, 并提取所述肖像的图像特征, 得到所述人物的角色节点; 解析确定所述文本的语义树, 基于所述语义树得到所述人物的运动特征, 得到所述人物的动作节点; 将每个所述人物对应的角色节点和动作节点连接; 其中, 所述人物的角色节点用肖像的图像特征表

征；所述人物的动作节点采用语义树中的运动特征表征。

作为一种实施方式，所述处理器执行所述程序时实现：将连接同一动作节点的角色节点相互连接。

作为一种实施方式，所述处理器执行所述程序时实现：将所述文本中的代词替换为所述代词所代表的所述人名。

作为一种实施方式，所述处理器执行所述程序时实现：检测出所述至少一个视频的每个镜头中的人物；提取所述人物的人体特征与运动特征；将所述人物的人体特征附加到所述人物的角色节点上，将所述人物的运动特征附加到所述人物的动作节点上；将每个人物对应的角色节点和动作节点相连。

10 作为一种实施方式，所述处理器执行所述程序时实现：将同时出现在一个镜头中的一组人物作为同组人物，将所述同组人物中的人物的角色节点两两相连。

作为一种实施方式，所述处理器执行所述程序时实现：将一个镜头中的一位人物和其相邻镜头的每个人物的角色节点都相连。

15 作为一种实施方式，所述处理器执行所述程序时实现：对每个视频的所述第一相似度和所述第二相似度加权求和，得到每个视频的相似度值；将相似度值最高的视频，确定为与所述检索条件相匹配的视频。

20 作为一种实施方式，所述处理器执行所述程序时实现：确定文本和训练样本集中的视频之间的第一相似度预测值，所述文本用于表征检索条件；确定所述文本的第一人物互动图和所述训练样本集中的视频的第二人物互动图之间的第二相似度；根据所述第一相似度预测值与所述第一相似度真值确定所述第一相似度的损失；根据所述第二相似度预测值与所述第二相似度真值确定所述第二相似度的损失；根据所述第一相似度的损失以及所述第二相似度的损失，结合损失函数确定总损失值；根据所述总损失值调整检索网络的权重参数。

25 作为一种实施方式，所述处理器执行所述程序时实现：基于所述总损失值调整所述第一子网络以及所述第二子网络的权重参数。

25 本公开实施例提供的检索装置，能提高根据文本描述检索视频的准确率。

本公开实施例还记载了一种计算机存储介质，所述计算机存储介质中存储有计算机可执行指令，所述计算机可执行指令用于执行前述各个实施例所述的检索方法。也就是说，所述计算机可执行指令被处理器执行之后，能够实现前述任意一个技术方案提供的检索方法。该计算机存储介质可以是易失性计算机可读存储介质或非易失性计算机可读存储介质。

30 本公开实施例还提供了一种计算机程序产品，包括计算机可读代码，当计算机可读代码在设备上运行时，设备中的处理器执行用于实现如上任一实施例提供的检索方法。

该上述计算机程序产品可以具体通过硬件、软件或其结合的方式实现。在一个可选实施例中，所述计算机程序产品具体体现为计算机存储介质，在另一个可选实施例中，计算机程序产品具体体现为软件产品，例如软件开发包（Software Development Kit, SDK）等等。

本领域技术人员应当理解，本实施例的计算机存储介质中各程序的功能，可参照前述各实施例所述的检索方法的相关描述而理解。

在本公开所提供的几个实施例中，应该理解到，所揭露的设备和方法，可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，如：多个单元或组件可以结合，或可以集成到另一个系统，或一些特征可以忽略，或不执行。另外，所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口，设备或单元的间接耦合或通信连接，可以是电性的、机械的或其它形式的。

上述作为分离部件说明的单元可以是、或也可以不是物理上分开的，作为单元显示的部件可以是、或也可以不是物理单元；既可以位于一个地方，也可以分布到多个网络单元上；可以根据实际的需要选择其中的部分或全部单元来实现本实施例方案的目的。

另外，在本公开各实施例中的各功能单元可以全部集成在一个处理单元中，也可以是各单元分别单独作为一个单元，也可以两个或两个以上单元集成在一个单元中；上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储设备、只读存储器（ROM, Read-Only Memory）、随机存取存储器（RAM, Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

或者，本公开上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用，也可以存储在一个计算机可读取存储介质中。基于这样的理解，本公开实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机、服务器、或者网络设备等）执行本公开各个实施例所述方法的全部或部分。而前述的存储介质包括：移动存储设备、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本公开的具体实施方式，但本公开的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本公开揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本公开的保护范围之内。因此，本公开的保护范围应以所述权利要求的保护范围为准。

工业实用性

本公开实施例提供的技术方案，确定文本和至少一个视频之间的第一相似度，所述文本用于表征检索条件；确定所述文本的第一人物互动图和所述至少一个视频的第二人物互动图；确定所述第一人物互动图和所述第二人物互动图之间的第二相似度；根据所述第一相似度和所述第二相似度，从所述至少一个视频中确定出与所述检索条件相匹配的视频。如此，相对于传统的基于特征的检索算法，本公开通过确定文本和至少一个视频之间的第一相似度，所述文本的第一人物互动图和所述

至少一个视频的第二人物互动图之间的第二相似度，可以利用文字本身的语法结构以及视频本身的事件结构等信息，进行视频检索，从而能提高根据文本描述检索视频如电影的准确率。