

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织  
国际局

(43) 国际公布日  
2024 年 1 月 4 日 (04.01.2024)



(10) 国际公布号  
**WO 2024/001057 A1**

(51) 国际专利分类号:  
**G06F 16/73** (2019.01) **G06F 16/783** (2019.01)

(21) 国际申请号: PCT/CN2022/137814

(22) 国际申请日: 2022 年 12 月 9 日 (09.12.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:  
202210768147.X 2022年7月1日 (01.07.2022) CN

(71) 申请人: 深圳先进技术研究院 (SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY) [CN/CN]; 中国广东省深圳市南山区西丽大学城学苑大道1068号, Guangdong 518055 (CN)。

(72) 发明人: 乔宇 (QIAO, Yu); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong

518055 (CN)。陈思然(CHEN, Siran); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。许清林(XU, Qinglin); 中国广东省深圳市南山区西丽大学城学苑大道 1068号, Guangdong 518055 (CN)。王亚立(WANG, Yali); 中国广东省深圳市南山区西丽大学城学苑大道1068号, Guangdong 518055 (CN)。马跃(MA, Yue); 中国广东省深圳市南山区西丽大学城学苑大道1068号, Guangdong 518055 (CN)。

(74) 代理人: 北京中巡通大知识产权代理有限公司 (BEIJING ZHONG XUN TONG DA INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市西城区西直门外大街 1 号院 2 号楼7层7C1, Beijing 100044 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,

(54) Title: VIDEO RETRIEVAL METHOD BASED ON ATTENTION SEGMENT PROMPT

(54) 发明名称: 一种基于注意力片段提示的视频检索方法

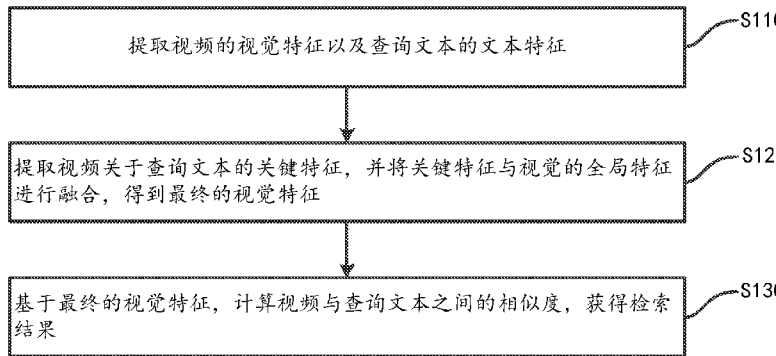


图 1

S110 Extract a visual feature of a video and a text feature of a query text

S120 Extract a key feature of the video relating to the query text, and fuse the key feature with a global visual feature to obtain a final visual feature

S130 Calculate a similarity between the video and the query text on the basis of the final visual feature, so as to obtain a retrieval result

(57) Abstract: Disclosed in the present invention is a video retrieval method based on attention segment prompt. The method comprises: extracting visual information for a video, further calculating a corresponding global feature, and extracting a text feature on the basis of a query text; using a time converter for the visual information, such that the feature of each frame has contextual information; searching for similar video clips in the video on the basis of the text feature, and acquiring video clip information most similar to the query text as a key visual feature; performing weighted sum on the key feature and the global feature to obtain a final visual feature; and calculating a similarity between the query text and the video according to the final visual feature, so as to retrieve a target video meeting a similarity

BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

requirement. According to the present invention, more weight is given to the most relevant screen while global information is reserved, such that the target video can be accurately retrieved.

(57) 摘要: 本发明公开了一种基于注意力片段提示的视频检索方法。该方法包括: 针对视频提取视觉信息, 进而计算对应的全局特征, 并基于查询文本提取文本特征; 将所述视觉信息利用时间转换器, 使得每一帧的特征都具有上下文信息; 基于所述文本特征查找视频中的相似视频片段, 获取与查询文本最相似的视频片段信息, 作为视觉的关键特征; 将所述关键特征和所述全局特征进行加权和, 得到最终的视觉特征; 根据所述最终的视觉特征计算查询文本与视频的相似度, 进而检索出满足相似度要求的目标视频。本发明在保留全局信息的同时赋予最相关画面更多的权重, 有利于准确地检索出目标视频。

## 一种基于注意力片段提示的视频检索方法

### 5 技术领域

本发明涉及视频检索技术领域，更具体地，涉及一种基于注意力片段提示的视频检索方法。

### 背景技术

10 随着多媒体技术的快速发展，互联网视频的数据规模在过去几年呈爆炸式增长，大规模的视觉文本预训练模型推动了视频文本检索的发展。因此对于给定的一个文本查询，如何从庞大的数据库中准确地找到最相关的视频已经成为多媒体领域最关注的任务之一。

在现有技术中，提出了不同方法将 CLIP（对比语言图像预训练模型）  
15 学习到的跨模态信息运用到视频检索中，将视频看作为一组连续的图片，并学习这些图片之间的时间关系，从而将 CLIP 在时间维度上进行扩展。然而，已有方法忽略了视频文本检索中的一个关键问题。由于原始视频可能包含多个场景片段的信息，不同的标注者会为此视频进行不同的文本描述，这些文本往往只描述视频的部分信息，因此已有方法很难将文本信息  
20 与丰富的视频信息匹配起来，无法定位到文本描述的是视频的哪一个片段。

综上，现有的视频文本检索方案忽略了视频文本检索中的一个关键问题（即文本通常只是相应视频中的一个小片段），通常毫无区别地整合整个视频，这不可避免地降低了最终检索的效果，导致只能应用于检索内容单一的视频。

25

### 发明内容

本发明的目的是克服上述现有技术的缺陷，提供一种基于注意力片段提示的视频检索方法，该方法包括以下步骤：

针对视频提取视觉信息，进而计算对应的全局特征，并基于查询文本

提取文本特征；

将所述视觉信息利用时间转换器，使得每一帧的特征都具有上下文信息；

5 基于所述文本特征查找视频中的相似视频片段，获取与查询文本最相似的视频片段信息，作为视觉的关键特征；

将所述关键特征和所述全局特征进行加权和，得到最终的视觉特征；

根据所述最终的视觉特征计算查询文本与视频的相似度，进而检索出满足相似度要求的目标视频。

10 与现有技术相比，本发明的优点在于，基于视频文本检索中内在的特性设计并解决问题，直接使用文本去查询最相关的视频帧，在保留全局信息的同时给可能的最相关画面更多的权重，从而有利于提升检索的准确性和效率。此外，本发明即使在小数据集上，也具有较好的优势，并且能够很好的迁移整合到其他的模型框架中，以进一步提高检索性能。而且，可以做到即插即用，非常简洁方便。

15 通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

## 附图说明

20 被结合在说明书中并构成说明书的一部分的附图示出了本发明的实施例，并且连同其说明一起用于解释本发明的原理。

图1是根据本发明一个实施例的基于注意力片段提示的视频检索方法的流程图；

图2是根据本发明一个实施例的基于注意力片段提示的视频检索框架示意图；

25 图3是根据本发明一个实施例的将注意力片段提示模块迁移现有模型的效果图。

## 具体实施方式

现在将参照附图来详细描述本发明的各种示例性实施例。应注意到：

除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。

5           对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

在这里示出和讨论的所有例子中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它例子可以具有不同的值。

10           应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

本发明提出了一个简洁的注意力片段提示(ASP)框架，可以动态地利用视频中与文本相关的片段来促进检索。简言之，注意力片段提示框架包  
15           括片段提示和视频聚合两个重要部分。对于一组文字和视频，片段提示可以巧妙地构造文本驱动的视觉提示，动态地获取与文本查询相关的视频片段。视频聚合可以获取视频的全局信息。通过片段提示和视频聚合的融合，本发明的模型框架可以有效地学习一种鲁棒的、与文本相关的视觉表示。

具体地，结合图 1 和图 2 所示，所提供的基于注意力片段提示的视频  
20           检索方法包括以下步骤：

### **步骤 S110，提取视频的视觉特征以及查询文本的文本特征。**

构建基于注意力片段提示的视频检索模型框架，参见图 2 所示，以功能模块的形式示意该框架。首先，可以根据查询文本动态查找视频中最相似的视频片段，获取视频关于查询文本的关键特征；然后，对视频信息进行  
25           全局整合，得到视频全局特征。最后，将获取关键特征和全局特征进行加权和，得到最终的视觉特征（或称视觉信息），进而根据最终的视觉特征获取视频检索结果。本发明提供的框架相对于现有技术更有效，因为许多原始视频包含多个场景的信息，而查询文本往往只描述视频的部分片段，这使得文本信息很难和丰富的视频特征进行匹配。而本发明可以根据文本

动态查找到其描述的视频片段，显著的提高视频检索效率。

仍结合图 2，模块 1 用于提取文本特征和视觉特征。例如，对视频进行平均降采样，将视频转化为图片序列，然后使用视觉特征提取器 VIT 提取视觉特征（或称视觉信息），而采用文本特征提取器提取文本特征，文本特征提取器可使用 BERT 或其他的语言模型。在一个实施例中，VIT 和 BERT 使用 CLIP 预训练之后得到的权重进行初始化。CLIP 模型利用图片文本匹配的预训练任务，通过自监督和对比学习方式从大规模的图片文本数据中学习视觉文本的跨模态信息，CLIP 能够 zero-shot 迁移到其他下游任务上，其框架也可以作为其他视觉文本任务的基础，获得更好的性能。

**步骤 S120，提取视频关于查询文本的关键特征，并将关键特征与视觉的全局特征进行融合，得到最终的视觉特征。**

在提取到了视觉特征以及文本特征之后，提出了新颖的基于注意力片段的提示模块（模块 2）。具体地，将视觉信息通过 temporal transformer（时间转换器）模块，这样使得每一帧的特征都具有上下文信息，然后根据查询文本，与视频的每一帧特征进行相似度对比，动态地获取与查询文本最相似的视频片段信息，作为视觉的关键特征。在获取到与文本相关的关键特征之后，对视觉特征提取器 VIT 提取到的视觉信息进行平均计算，得到视觉的全局特征。然后将关键特征与全局特征进行加权求和，得到最终的视觉特征。

本发明通过提取视觉的关键特征，将注意力集中在与查询文本相关度强的重要信息上，从而提高了计算效率高以及特征提取的准确性，并降低了参数量。

在一个实施例中，可采用线性加权或指数加权的形式融合关键特征和全局特征。例如，通过不断尝试，选择最优的超参数来权衡关键特征(片段级别)和全局特征(视频级别)的重要程度，并且，针对不同数据集，根据数据集自身的特点，分别探索各自最优的超参数，最终在多种类型数据集上都实现了较优的性能。

应理解的是，也可以使用不同的视觉特征提取器或者文本的特征编码器，或者新的模型框架，本发明的核心点是基于文本的跨模态交互范式，

在不违背本发明思想的情况下，本领域技术人员也可以选择不同的交互函数，更复杂的交互模块，或者不同的特征融合方式。此外，上述关键特征和全局特征的提取顺序可以并行处理或先后处理，本发明对此不进行限定。

**步骤 S130，基于最终的视觉特征，计算视频与查询文本之间的相似度，获得检索结果。**

在进行上述的视频特征融合，获得最终的视觉特征之后，进一步通过跨模态相似度计算，输出视频检索结果（参见图 2 的模块 3）。

在一个实施例中，使用余弦相似度作为视频和文本相似度的评价准则，计算查询文本与每一个视频的相似度。在模型训练时，根据训练得到的相似度矩阵，计算出损失函数，进行模型的学习。在实际检索时，根据相似度选取最相似的一个或多个视频作为检索到的目标视频进行输出。

本发明的损失函数考虑了查询文本与视频之间的相似度，通过这种设计，在针对性的选择视频帧时，能够直接选取相关性最大的视频帧。在一个实施例中，将损失函数设置为：

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^M (\mathcal{L}_{\text{txt}}^{(i)} + \mathcal{L}_{\text{vdo}}^{(i)}) \quad (1)$$

$$\mathcal{L}_{\text{txt}}^{(i)} = -\log \frac{\exp(\langle \mathbf{Z}^{(i,j)}, \mathbf{T}^{(i)} \rangle)}{\sum_{j=1}^M \exp(\langle \mathbf{Z}^{(i,j)}, \mathbf{T}^{(i)} \rangle)} \quad (2)$$

$$\mathcal{L}_{\text{vdo}}^{(i)} = -\log \frac{\exp(\langle \mathbf{T}^{(i)}, \mathbf{Z}^{(i,j)} \rangle)}{\sum_{j=1}^M \exp(\langle \mathbf{T}^{(i)}, \mathbf{Z}^{(i,j)} \rangle)} \quad (3)$$

其中， $\mathbf{T}^{(i)}$ 表示这个 batch 第 i 个文本的文本特征， $\mathbf{Z}^{(i,j)}$ 表示每个 batch 的第 i 个视频通过第 j 个文本的片段提示生成的视觉特征； $\langle, \rangle$ 表示两个特征之间的余弦相似度， $M$ 表示视频数目。

利用本发明，对于给定的一个文本描述，可以从海量的视频库中有效的找到和文本最相关的一个或多个视频。这是因为，在实际情况中，视频片段可能包括不止一个画面的内容，包括镜头的切换，目标人物进行连续的多个活动等，例如在一段动画片段中，涉及到两个人在餐厅的桌子上喝饮料并交谈，随后他们走出餐厅边走边说话，针对这个视频，有 20 个不同的标注者观看了视频，然后给出了自己的注释，相当一部分人只描述了视频的一部分，例如“两个动画角色在餐桌前”，“两个动画角色边走路边

说话”等。基于这样的考虑，本发明针对性的提出了视频片段推荐的思路，例如，以每个视频帧为中心，通过自我注意力机制，整合前后时间序列的信息，作为推荐片段候选，再用要查询的文本表示分别与这些推荐片段候选计算余弦相似度，直接选择相似度最高的作为局域信息的推荐片段；同时

5 时也兼顾了整体信息，用全部帧表示的平均池化作为全局信息，这样得出的视频表达可以针对特异性的文本输入，动态地计算和视频的相关分数，找到视频中和文本查询最相关的部分，从而满足用户的个性化查询需求。

为进一步验证本发明的效果，进行了实验，从检索准确性、鲁棒性以及可迁移性等不同角度验证了本发明的有效性。

10 1) 准确性。现有技术通过跨模态模型将视频和文本信息整合在一个特征空间中，在视觉特征中，这些方法将视频看作多张连续的图片，直接进行平均求和。而本发明可以根据查询文本动态地查找视频中最相似的视频片段，获取视频关于查询文本的关键特征，从而有利于提升视频文本检索的准确性。参见表 1，其中 our ASP 代表本发明的框架，其他是现有模型。

15 2) 鲁棒性。对于现有技术，在少量训练数据的情况下，鲁棒性较差。而本发明通过查询文本动态地与视频片段对齐，获取与文本最相关的片段，增强了检索的鲁棒性，在少量训练集的情况下也有较好的实验结果，参见表 2，示意了训练数据大小分别为 30、300 和 3000 情况下，与 CLIP4Clip 实验结果对比。

20 3) 可迁移性。现有技术提出的模型复杂并且迁移性差。本申请能够轻便的迁移到现有的方法。如图 3 所示，将注意力片段提示模块迁移到 CLIP2Video 上，分别对比 CLIP2Video、CLIP2Video+注意力片段提示模块和 ASP 在数据集 MSR-VTT 上的实验结果。

表 1：与现有视频检索方法在 MSV-RTT 数据集上的实验结果对比

| MSR-VTT    |          | Text-to-Video |      |      |      |      | Video-to-Text |      |      |     |      |
|------------|----------|---------------|------|------|------|------|---------------|------|------|-----|------|
| Method     | Test Set | R@1           | R@5  | R@10 | MdR  | MaR  | R@1           | R@5  | R@10 | MdR | MaR  |
| CE         | 1k-A     | 20.9          | 48.8 | 62.4 | 6.0  | 28.2 | 20.6          | 50.3 | 64.0 | 5.3 | 25.1 |
| MMT        | 1k-A     | 26.6          | 57.1 | 69.6 | 4.0  | 34.0 | 27.0          | 57.5 | 69.7 | 3.7 | 21.3 |
| MDMMT      | 1k-A     | 38.9          | 69.0 | 79.9 | 2.0  | 16.5 | -             | -    | -    | -   | -    |
| CLIP       | 1k-A     | 31.2          | 53.7 | 64.2 | 4.0  | -    | 27.2          | 51.7 | 62.6 | 5.0 | -    |
| CLIP4Clip  | 1k-A     | 44.5          | 71.4 | 81.6 | 2.0  | 15.3 | 42.7          | 70.9 | 80.6 | 2.0 | 10.2 |
| CLIP2Video | 1k-A     | 45.6          | 72.6 | 81.7 | 2.0  | 14.6 | 43.5          | 72.3 | 82.1 | 2.0 | 10.2 |
| CAMoE      | 1k-A     | 47.3          | 74.2 | 84.5 | 2.0  | 11.9 | 49.1          | 74.3 | 84.3 | 2.0 | 9.9  |
| Our ASP    | 1k-A     | 52.6          | 78.4 | 87.0 | 1.0  | 9.9  | 51.9          | 78.0 | 85.7 | 1.0 | 7.5  |
| CE         | Full     | 10.0          | 29.0 | 41.2 | 16.0 | 86.8 | 15.6          | 40.9 | 55.2 | 8.3 | 58.1 |
| HIT        | Full     | 14.9          | 40.2 | 52.8 | 9.0  | -    | -             | -    | -    | -   | -    |
| CLIP       | Full     | 21.4          | 41.1 | 50.4 | 10.0 | -    | 40.3          | 69.7 | 79.2 | 2.0 | -    |
| UNIVL      | Full     | 21.2          | 49.6 | 63.1 | 6.0  | -    | -             | -    | -    | -   | -    |
| MDMMT      | Full     | 23.1          | 49.8 | 61.8 | 6.0  | 52.8 | -             | -    | -    | -   | -    |
| CLIP2Video | Full     | 29.8          | 55.5 | 66.2 | 4.0  | 45.5 | 54.6          | 82.1 | 90.8 | 1.0 | 5.3  |
| CAMoE      | Full     | 32.9          | 58.3 | 68.4 | 3.0  | 42.6 | 59.8          | 86.2 | 92.8 | 1.0 | 3.8  |
| Our ASP    | Full     | 34.2          | 60.1 | 70.2 | 3.0  | 38.2 | 61.8          | 87.2 | 93.0 | 1.0 | 3.7  |

表 2：在不同训练数据大小与 CLIP4Clip 实验结果对比

| Text-to-Video | R@1       |         | R@5       |         | R@10      |         | MdR       |         | MaR       |         |
|---------------|-----------|---------|-----------|---------|-----------|---------|-----------|---------|-----------|---------|
|               | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP |
| train size    |           |         |           |         |           |         |           |         |           |         |
| 30            | 39.4      | 42.4    | 63.8      | 65.9    | 74.2      | 76.2    | 3.0       | 2.0     | 18.4      | 17.6    |
| 300           | 43.3      | 45.2    | 69.7      | 70.7    | 78.3      | 81.3    | 2.0       | 2.0     | 17.4      | 15.8    |
| 3000          | 47.2      | 48.5    | 73.1      | 74.3    | 81.9      | 84.1    | 2.0       | 2.0     | 13.0      | 11.6    |
| 9000          | 51.2      | 52.6    | 76.6      | 78.4    | 85.2      | 87.0    | 1.0       | 1.0     | 12.0      | 9.9     |
| Video-to-Text | R@1       |         | R@5       |         | R@10      |         | MdR       |         | MaR       |         |
|               | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP | CLIP4Clip | Our ASP |
| train size    |           |         |           |         |           |         |           |         |           |         |
| 30            | 37.8      | 39.9    | 63.4      | 65.3    | 73.2      | 75.4    | 2.0       | 2.0     | 18.7      | 18.5    |
| 300           | 43.1      | 44.5    | 68.3      | 70.8    | 78.0      | 80.6    | 2.0       | 2.0     | 14.4      | 12.6    |
| 3000          | 45.9      | 47.7    | 73.5      | 75.2    | 81.7      | 83.9    | 2.0       | 2.0     | 10.4      | 8.5     |
| 9000          | 51.6      | 51.9    | 77.4      | 78.0    | 85.0      | 85.7    | 1.0       | 1.0     | 9.1       | 7.5     |

通过在多个基准上进行广泛实验，证明了本发明的表现明显优于现有的视频检索方法。例如，在 DiDeMo 数据集上，得到 49.5 R@1 和 13.8 MnR 的结果，与最近提出的 CLIP4Clip、QB-Norm 和 CAMoE 相比，在 R@1 的评估准则上有超过 5%的提升。

综上所述，相对于现有技术，本发明至少具有以下技术效果：

1) 所设计的即插即用的注意力片段提示模块，可以根据查询文本动态地获取相关的视频片段，使得不同的查询文本可以有效地匹配视频中不同的相关内容，增强了检索效果。

2) 给定一段视频和人工注释，由于视频中可能包含了很多个画面，而人工标注仅仅对应了其中的一小部分，所以平均使用全局信息会忽略点

重点。本发明抓住了视频文本检索中内在的特性来设计并解决问题，直接使用文本去查询最相关的视频帧，在保留全局信息的同时给可能的最相关画面更多的权重，从而更好得帮助检索。并且，本发明非常简洁高效，不用耗费大规模的计算资源就能够实现性能的提升。

- 5           3) 众所周知，收集数据往往是深度学习任务中最费时费力的，依赖人工标注的标签，还存在侵犯个人隐私的情况，很多实际情况下，由于具体情况的限制，很难得到非常多的数据用于训练。本发明即使是在小数据集上，也能很好的发挥优势，并且迁移性强，可以将 ASP 模块移植到其他检索模型中，进一步提高检索结果。例如，常规的模型是 CLIP 和针对视频
- 10 做出改进的 CLIP4Clip，本发明在这个基础上添加了自己的贡献，检索性能显著优于 CLIP4Clip 的 baseline（基准）。

本发明可应用于电子设备、服务器或云端，根据查询文本检索出一个或多个目标视频。该电子设备可以是终端设备或者服务器，终端设备包括手机、平板电脑、个人数字助理（PDA）、销售终端（POS）、智能可穿戴设备(智能手表、虚拟现实眼镜、虚拟现实头盔等)等任意终端设备。服务器包括但不限于应用服务器或 Web 服务器，可以为独立服务器、集群服务器或云服务器等。

15

本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

20

计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是但不限于电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：

25 便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质

不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如 Smalltalk、C++、Python 等，以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网（LAN）或广域网（WAN）—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列（FPGA）或可编程逻辑阵列（PLA），该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

这里参照根据本发明实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它

可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、  
5 可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备  
10 上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述  
15 模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。  
20 也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。对于本领域技术人员来说公知的是，通过硬件方式实现、通过软件方式实现以及通过软件和硬件结合的方式实现都是等价的。

25 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术

人员能理解本文披露的各实施例。本发明的范围由所附权利要求来限定。

## 权 利 要 求 书

1. 一种基于注意力片段提示的视频检索方法，包括以下步骤：

针对视频提取视觉信息，进而计算对应的全局特征，并基于查询文本提取文本特征；

5 将所述视觉信息利用时间转换器，使得每一帧的特征都具有上下文信息；

基于所述文本特征查找视频中的相似视频片段，获取与查询文本最相似的视频片段信息，作为视觉的关键特征；

将所述关键特征和所述全局特征进行加权和，得到最终的视觉特征；

10 根据所述最终的视觉特征计算查询文本与视频的相似度，进而检索出满足相似度要求的目标视频。

2. 根据权利要求 1 所述的方法，其特征在于，所述关键特征根据以下步骤获得：

15 将所述文本特征与视频的每一帧特征进行相似度对比，动态地获取与查询文本最相似的视频片段信息，作为视觉的关键特征。

3. 根据权利要求 1 所述的方法，其特征在于，在所述关键特征的获取中，使用余弦相似度作为视频和文本相似度的评价准则，计算得到查询文本与每一个视频片段的相似度。

20 4. 根据权利要求 1 所述的方法，其特征在于，所述全局特征根据以下步骤获得：

对视频进行平均降采样，将视频转化为图片序列；

将所述图片序列输入到视觉特征提取器 VIT，提取出视觉信息；

对所提取的视觉信息进行平均计算，得到视觉的全局特征。

25 5. 根据权利要求 4 所述的方法，其特征在于，利用 BERT 模型提取所述文本特征，所述视觉特征提取器 VIT 和所述 BERT 模型利用 CLIP 预训练的权重进行初始化。

6. 根据权利要求 1 所述的方法，其特征在于，根据所述最终的视觉特征计算查询文本与视频的相似度，进而检索出满足相似度要求的目标视频包括：

以每个视频帧为中心，通过自我注意力机制，整合前后时间序列的信息，作为推荐片段候选；

利用待查询的文本特征分别和这些推荐片段候选计算余弦相似度，直接选择相似度最高的片段，作为局域信息的推荐片段；同时用全部帧表示的平均池化作为全局特征；

动态地计算待查询的文本特征和各视频的相关程度，找到与文本查询最相关的视频。

7. 根据权利要求 1 所述的方法，其特征在于，根据所述最终的视觉特征计算查询文本与视频的相似度利用经训练的跨模态相似度计算模型获得。

8. 根据权利要求 7 所述的方法，其特征在于，训练所述跨模态相似度计算模型的损失函数根据各视频片段与文本特征之间的相似度矩阵计算。

9. 一种计算机可读存储介质，其上存储有计算机程序，其中，该计算机程序被处理器执行时实现根据权利要求 1 至 8 中任一项所述的方法的步骤。

10. 一种计算机设备，包括存储器和处理器，在所述存储器上存储有能够在处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求 1 至 8 中任一项所述的方法的步骤。

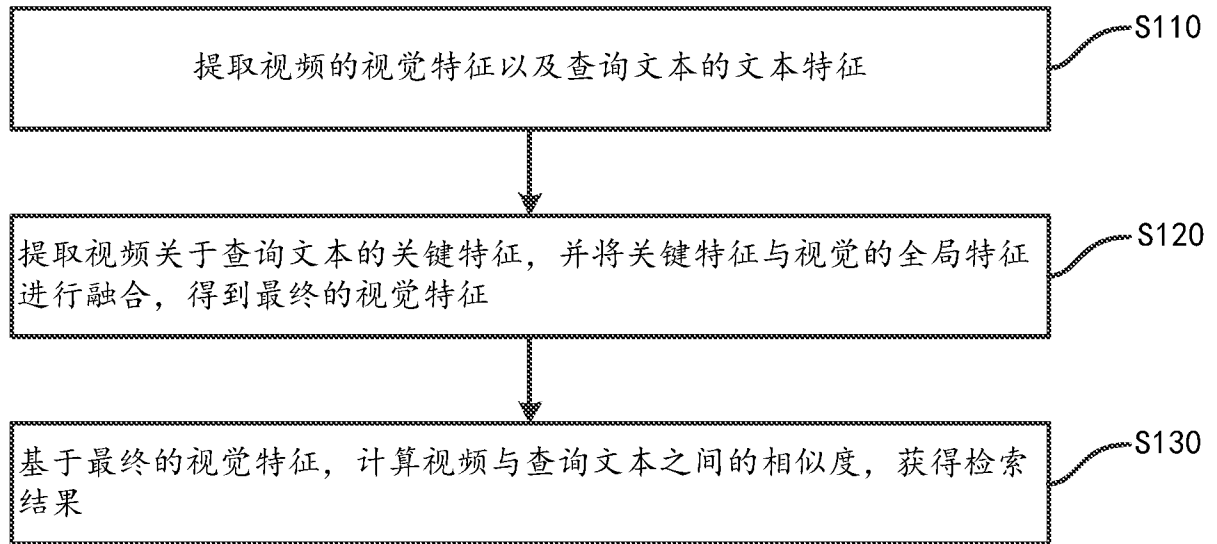


图 1

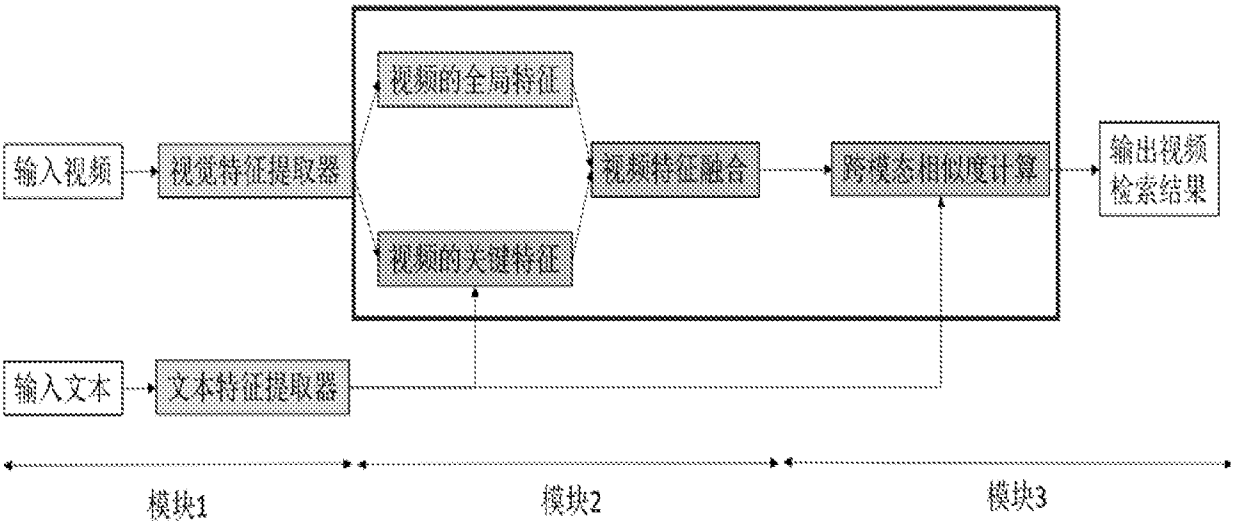


图 2

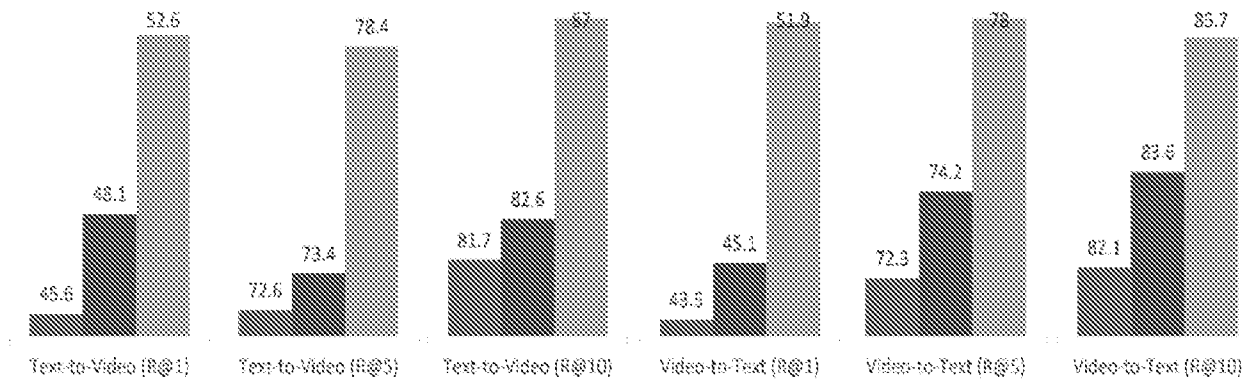


图 3

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/137814

**A. CLASSIFICATION OF SUBJECT MATTER**

G06F 16/73(2019.01)i;G06F 16/783(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 注意力, 视频, 视觉, 全局, 融合, 相似, 局部 or 关键, 帧, 训练, 权重, 跨模态, attention, video, visual, global, fusion, similar, local or key, frame, train, weight, cross-modal

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                               | Relevant to claim No. |
|-----------|----------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| PX        | CN 115269913 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY) 01 November 2022 (2022-11-01)<br>claims 1-10                         | 1-10                  |
| A         | CN 114037945 A (ZHEJIANG GONGSHANG UNIVERSITY) 11 February 2022 (2022-02-11)<br>description, paragraphs 0028-0066                | 1-10                  |
| A         | CN 102750339 A (BEIJING JIAOTONG UNIVERSITY) 24 October 2012 (2012-10-24)<br>entire document                                     | 1-10                  |
| A         | CN 110933518 A (ZHEJIANG UNIVERSITY) 27 March 2020 (2020-03-27)<br>entire document                                               | 1-10                  |
| A         | CN 113590881 A (BEIJING DAJIA INTERNET INFORMATION TECHNOLOGY CO., LTD. et al.) 02 November 2021 (2021-11-02)<br>entire document | 1-10                  |
| A         | US 2021248376 A1 (ADOBE INC.) 12 August 2021 (2021-08-12)<br>entire document                                                     | 1-10                  |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

13 March 2023

Date of mailing of the international search report

22 March 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/  
CN)  
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,  
Beijing 100088

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

**INTERNATIONAL SEARCH REPORT**  
**Information on patent family members**

International application No.

**PCT/CN2022/137814**

| Patent document<br>cited in search report |            |    | Publication date<br>(day/month/year) | Patent family member(s) |            |    | Publication date<br>(day/month/year) |
|-------------------------------------------|------------|----|--------------------------------------|-------------------------|------------|----|--------------------------------------|
| CN                                        | 115269913  | A  | 01 November 2022                     | None                    |            |    |                                      |
| CN                                        | 114037945  | A  | 11 February 2022                     | None                    |            |    |                                      |
| CN                                        | 102750339  | A  | 24 October 2012                      | None                    |            |    |                                      |
| CN                                        | 110933518  | A  | 27 March 2020                        | None                    |            |    |                                      |
| CN                                        | 113590881  | A  | 02 November 2021                     | None                    |            |    |                                      |
| US                                        | 2021248376 | A1 | 12 August 2021                       | US                      | 2022122357 | A1 | 21 April 2022                        |

## 国际检索报告

国际申请号

PCT/CN2022/137814

## A. 主题的分类

G06F 16/73 (2019.01) i; G06F 16/783 (2019.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

## B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

IPC: G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

VEN、CNABS、CNTXT、WOTXT、EPTXT、USTXT、CNKI、IEEE: 注意力, 视频, 视觉, 全局, 融合, 相似, 局部 or 关键, 帧, 训练, 权重, 跨模态, attention, video, visual, global, fusion, similar, local or key, frame, train, weight, cross-modal

## C. 相关文件

| 类 型* | 引用文件, 必要时, 指明相关段落                                                     | 相关的权利要求 |
|------|-----------------------------------------------------------------------|---------|
| PX   | CN 115269913 A (深圳先进技术研究院) 2022年11月1日 (2022 - 11 - 01)<br>权利要求1-10    | 1-10    |
| A    | CN 114037945 A (浙江工商大学) 2022年2月11日 (2022 - 02 - 11)<br>说明书第0028-0066段 | 1-10    |
| A    | CN 102750339 A (北京交通大学) 2012年10月24日 (2012 - 10 - 24)<br>全文            | 1-10    |
| A    | CN 110933518 A (浙江大学) 2020年3月27日 (2020 - 03 - 27)<br>全文               | 1-10    |
| A    | CN 113590881 A (北京达佳互联信息技术有限公司 等) 2021年11月2日 (2021 - 11 - 02)<br>全文   | 1-10    |
| A    | US 2021248376 A1 (ADOBE INC.) 2021年8月12日 (2021 - 08 - 12)<br>全文       | 1-10    |

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

\* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&amp;” 同族专利的文件

国际检索实际完成的日期

2023年3月13日

国际检索报告邮寄日期

2023年3月22日

ISA/CN的名称和邮寄地址

中国国家知识产权局

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

马鑫

电话号码 (+86) 010-53961354

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2022/137814

| 检索报告引用的专利文件 |            |    | 公布日<br>(年/月/日) | 同族专利 |            |    | 公布日<br>(年/月/日) |
|-------------|------------|----|----------------|------|------------|----|----------------|
| CN          | 115269913  | A  | 2022年11月1日     | 无    |            |    |                |
| CN          | 114037945  | A  | 2022年2月11日     | 无    |            |    |                |
| CN          | 102750339  | A  | 2012年10月24日    | 无    |            |    |                |
| CN          | 110933518  | A  | 2020年3月27日     | 无    |            |    |                |
| CN          | 113590881  | A  | 2021年11月2日     | 无    |            |    |                |
| US          | 2021248376 | A1 | 2021年8月12日     | US   | 2022122357 | A1 | 2022年4月21日     |