

(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
2023 年 4 月 6 日 (06.04.2023)



(10) 国际公布号
WO 2023/050295 A1

(51) 国际专利分类号:
G06V 20/40 (2022.01) *G06K 9/62* (2022.01)

(21) 国际申请号: PCT/CN2021/122130

(22) 国际申请日: 2021 年 9 月 30 日 (30.09.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 中远海运科技股份有限公司 (COSCO SHIPPING TECHNOLOGY CO., LTD.) [CN/CN]; 中国上海市浦东新区民生路 600 号, Shanghai 200135 (CN)。上海船舶运输科学研究所有限公司 (SHANGHAI SHIP&SHIPPING RESEARCH INSTITUTE) [CN/CN]; 中国上海市浦东新区民生路 600 号, Shanghai 200135 (CN)。大连海事大学 (DALIAN MARITIME UNIVERSITY) [CN/CN]; 中国辽宁省大连市甘井子区凌海路 1 号, Liaoning 116026 (CN)。

(72) 发明人: 刘晋 (LIU, Jin); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。韩冰 (HAN, Bing); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。李颖 (LI, Ying); 中国辽宁省大连市凌海路 1 号, Liaoning 116026 (CN)。吴中岱 (WU, Zhongdai); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。龚沛朱 (GONG, Peizhu); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。张喜亮 (ZHANG, Xiliang); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。王骏翔 (WANG, Junxiang); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。郭磊 (GUO, Lei); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。胡蓉 (HU, Rong); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。朱晓蓉 (ZHU, Xiaorong); 中国上海市浦东民生路 600 号, Shanghai 200135 (CN)。

(74) 代理人: 北京海虹嘉诚知识产权代理有限公司 (HAIHONG JIACHENG INTELLECTUAL

(54) Title: MULTIMODAL HETEROGENEOUS FEATURE FUSION-BASED COMPACT VIDEO EVENT DESCRIPTION METHOD

(54) 发明名称: 一种基于多模态异质特征融合的视频密集事件描述方法

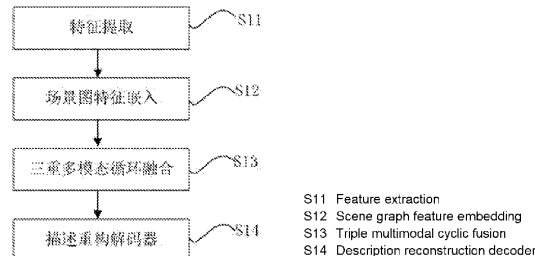


图 1

(57) Abstract: The present invention relates to the field of computer vision, in particular relates to a multimodal heterogeneous feature fusion-based video event description method. According to the present invention, an I3D convolutional network is used for clipping a video and extracting dynamic visual features, and a VGGish model is used for extracting audio prosodic features; semantic representation is performed on scene object information, a scene map is generated to obtain entity coding, attribute coding and relationship coding, and map embedding is performed on a feature vector by means of map convolution; triple multimodal cyclic fusion is performed on three extracted features; adaptive multimodal data balance is carried out to enable dynamic vision and audio prosodic features to match each other so as to ensure that event extraction is reasonable; and a description reconstruction decoder is used for detecting a video event by utilizing a description reconstruction algorithm, and generating the description of a video scene event according to a pre-trained language dictionary. The problem of meaningless descriptions such as co-occurring descriptions being frequently generated in a video description method is effectively solved, and the relationship of scene events is mined by effectively utilizing multimodal information.

[见续页]

PROPERTY & PARTNERS); 中国北京市海淀区海淀南路21号中关村知识产权大厦B座9层, Beijing 100080 (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, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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)。

根据细则4.17的声明:

— 发明人资格(细则4.17(iv))

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本发明涉及计算机视觉领域, 具体涉及一种基于多模态异质特征融合的视频事件描述方法。本发明利用I3D卷积网络剪辑视频并提取动态视觉特征, VGGish模型提取音频韵律特征; 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; 将提取的三个特征进行三重多模态循环融合; 自适应多模态数据平衡, 使动态视觉、音频韵律特征相互契合, 保证事件提取合理; 描述重构解码器, 利用描述重构算法对视频事件进行检测, 并根据预训练语言字典生成视频场景事件的描述。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问题, 并有效利用多模态信息挖掘出场景事件的关系。