Computer Science and Application 计算机科学与应用, 2019, 9(8), 1584-1590
Published Online August 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.98178

Hans汉斯

From Video to Semantic: Video Semantic Analysis Technology Based on Knowledge Graph

Liqiong Deng*, Jixiang Wu, Li Zhang

Air Force Communication NCO Academy, Dalian Liaoning
Email: *tigerss1016@163.com

Received: Aug. 6th, 2019; accepted: Aug. 19th, 2019; published: Aug. 26th, 2019

Abstract

Video understanding has attracted much research attention especially since the recent availability of large-scale video benchmarks. In order to fill up the semantic gap between video features and understanding, this paper puts forward a video semantic analysis process based on knowledge graph, and adopts random walk to quantify semantic consistency between semantic labels. Then video semantic reasoning based-on knowledge graph is studied. The experimental results prove that knowledge graph can improve semantic understanding effectively. Finally, a constructed multilevel video semantic model supports applications in video classifying, video labeling and video abstract, which has some guiding significance for information organization and knowledge management of media semantic.

Keywords

Knowledge Graph, Video, Classify, Semantic Analysis

从视频到语义：基于知识图谱的视频语义分析技术

邓莉琼*，吴吉祥，张　丽

空军通信士官学校，辽宁 大连
Email: *tigerss1016@163.com

收稿日期：2019年8月6日；录用日期：2019年8月19日；发布日期：2019年8月26日

*通讯作者。

文章引用: 邓莉琼, 吴吉祥, 张丽. 从视频到语义: 基于知识图谱的视频语义分析技术[J]. 计算机科学与应用, 2019, 9(8): 1584-1590. DOI: 10.12677/csa.2019.98178

摘　要

随着大规模视频的迅猛发展，视频理解受到了广泛的关注，为了填补视频特征与视频理解之间的语义鸿沟，本文提出了一种基于知识图谱的视频语义分析流程，采用了随机漫步方法对视频语义标签信息进行共生性概率的量化，研究了基于知识图谱的视频语义推理技术，相关的实验结果证明了知识图谱方法能有效提高视频语义分析的准确度，构建后的多层次视频语义模型支持在视频分类、视频标注及视频摘要等方面的应用，对媒体语义中的信息组织和知识管理有一定的指导意义。

关键词

知识图谱，视频，分类，语义分析

1. 引言

在移动互联网、大数据的时代背景下，互联网上的视频数据呈现爆发式增长，由于其内容具有易复制、易分发、难管理、难监控等特性，视频语义内容的有效管理成为了近年来的研究热点。语义鸿沟的存在导致了计算机自动描述视频语义准确率低的问题，针对这一问题，本文提出了基于知识图谱的视频语义分析技术，重点关注视频语义的分析描述研究，知识图谱作为一种智能、高效的知识组织方式，能够帮助用户迅速、准确地查询到自己需要的信息，在增进信息的组织、管理和理解领域具有巨大的应用潜力，是对视频视觉语义理解的一个行之有效的途径。本文将知识图谱技术用于构建视频的语义框架之中，将语义关系融入到特征提取中，有效的弥补语义鸿沟，为视频语义理解提供有效的支撑，该方向的研究具有较高的应用价值和现实意义，可广泛应用于视频检索、人机交互、智能安防等。

2. 相关工作

视频理解在计算机视觉领域是研究热点问题，随着近年来一些大型视频数据集标准(Sports-1M/YFCC-100M/Youtube-8M)的公布以及深度学习和神经网络技术在视频特征提取的运用，视频理解技术得到了巨大的发展。视频分类技术可以分为基于帧层次和基于视频层次两种。在基于帧层次，典型的有 DBoF [1] (deep bag-of-frames)方法，DBoF 借鉴了自然语言处理中的 BOW 的思想，可以理解为将多个帧特征合称为一个视频特征，使用了 deep 的思想，利用 DNN 将帧特征映射到更高维的空间中并进行求和，然后再利用 DNN 将高维特征映射回低维空间中进行分类。LSTM [2] (Long Short-Term Memory)则是不断传入帧数据，用最后的 LSTM 向量来作为视频的表示向量。在基于视频层次，则是对一系列帧进行聚合得到特征向量，然后利用支持向量机等方法进行分类训练，其中 MoE [3] (mixture of experts)在视频层次的特征提取和分类上表现尤其突出。除了以上两种方法外，还有利用视频中文本识别的方法来进行分类[4]。

虽然视频特征提取的准确度有了较大的提高，但语义鸿沟的问题依然存在，面对海量的视频信息，人们期望以更加智能的方式组织图像资源。知识图谱技术的出现使得信息可以在语义层面上进行整合，这种语义层次的关联技术能够为视频的语义分析研判提供强有力的支撑[5]。例如图 1(a)所示，仅仅分析该视频帧的像素特征，由于小孩手中的话筒被挡住了，因而很难得出“她在拿着话筒讲话”这样的结论，但若基于语义知识推理的话则不难得出该结论；图 1(b)所示是一个动物园的例子，但仅仅从像素特征的

分析很难认定是动物园，若基于“有老虎的人为建筑很有可能是动物园”这样的知识，则该视频将很有可能被正确划分为动物园。因此知识图谱的构建能极大填补语义鸿沟的存在。

(a) (b)

Figure 1. (a) Child with microphone, (b) Tiger in the zoo
图 1. (a) 小孩拿话筒，(b) 动物园老虎

知识图谱即为用图对知识和知识间关系进行建模。图节点表示知识的概念或实体，图边表示概念或实体间关系，众多节点和边构成的图即可对知识进行完整而清晰的描述。它们力求通过将知识进行更加有序、有机的组织，对用户提供更加智能的访问接口，使用户可以更加快速、准确地访问自己需要的知识信息，并进行一定的知识挖掘和智能决策。例如将图 1 所示的视频特征建立为图 2 的知识图谱，通过节点之间的关系能够更好的帮助理解视频的语义内容。近年来已经有不少将知识图谱应用于视频等多媒体领域[6]，例如文献[7]使用知识图谱来进行视频的分类，然而该方法的知识图谱是独立于特征模型的，缺少反馈回路，故而准确度不高等。至于知识图谱的具体构建技术不在本文重点研究范围之内。

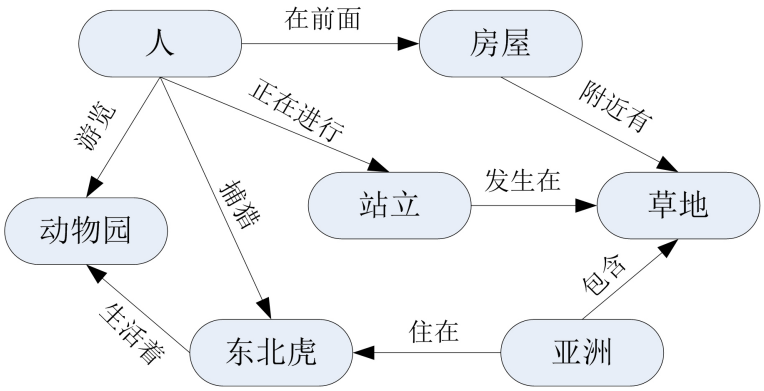

Figure 2. Example of knowledge graph relations for video semantics
图 2. 视频语义的知识图谱关系示例

3. 基于知识图谱的视频语义分析流程

针对视频的语义分析，本文所提出的基于知识图谱的视频语义分析流程图如图 3 所示。

如图 3 所示，输入一个待分析的视频后，首先从关键视频帧中提取出视频特征和音频特征；然后将这些帧向量特征输入到基于帧的建模或基于视频的建模中，生成最终的知识图谱向量，并输入到分类器中。

该分析框架有两个优势，首先，该框架可适用于目前所有的视频分类算法，包括深度学习和浅层学

习等模型，因而具有较高的灵活性；其次，在机器学习的框架中融入了知识图谱的构建，用语义内容之间的关联性填补了视频语义鸿沟，从而提高了准确度。

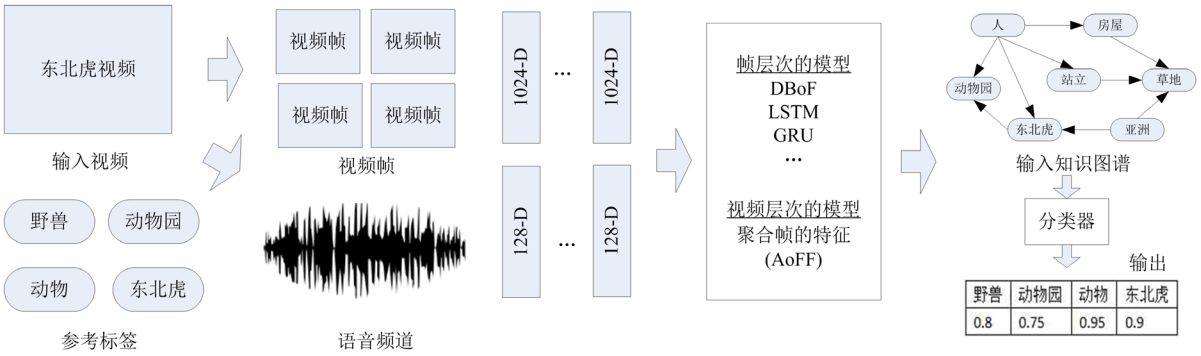

Figure 3. Video semantics analysis framework based on knowledge graph
图 3. 基于知识图谱的视频语义分析框架

3.1. 视频帧序列特征提取

本文通过对输入的视频帧序列提取视频的 3 类特征，包括空间特征(基于 VGG16、AlexNet 的 fc7 层特征)、视频特征(DT 特征)，然后对于可进行融合的特征进行前期融合，再通过一个特征选择器，该特征选择器的作用为选择提取到的及前期融合后得到的特征的组合作为 DBoF、LSTM 等描述模型的输入。

空间特征：本文使用预训练的模型提取视频帧序列图像的空间特征，因为近年来 CNN 在图像分类、目标检测、图像语义分割等领域取得了一系列突破性的研究成果[8]，通过 CNN 提取的特征能够很好地表达图像。因此本文选择在 ImageNet 分类任务数据集中取得很好数据的 CNN 模型 VGG16 和 AlexNet，提取预处理好的视频帧序列中所有图片的 fc7 层的特征，并计算帧序列特征的均值，最终得到一个 4096 维特征向量来表示整个视频。

视频特征：与单独的图片描述问题不同的是，视频帧之间具有时间上的关联性，故而在对视频进行分析时很有必要进行视频的时间上的特征提取。本文使用文献[9]方法提取 DT 特征，在提取 DT 特征时采用不重叠的长方形块覆盖图像上的区域，最后将拼接获取到的各区域的 DT 特征作为整个视频的特征。由于视频特征的提取算法不是本文关注主要问题，因此不在文中详细阐述。

3.2. 视频语义的知识图谱表示

当提取出视频的特征之后，本文利用知识图谱来进一步表示视频里的语义关系[10]，知识图谱用 $G=(V,E)$ 表示，V 表示是途中各节点的组合(即视频中的语义实体或类别标签)，E 表示这些节点之间的连线，即各语义标签之间的关系。本文引用了文献[7]中所描述的语义共生性来量化语义节点之间的语义联系，一般来说两个语义节点之间的语义共生性越高，则表示这两个语义标签出现在同一个视频里的概率越大，例如东北虎和动物园是两个具有较强联系的语义标签，但东北虎和火山则是两个具有弱联系的语义标签。

本文将语义共生性矩阵定义为 S，S 为一个 $L\times L$ 的矩阵，L 表示的是视频内所有语义实体标签的个数，S_{ij} 表示的是语义标签 i 和语义标签 j 之间的语义共生性，需要指出的是，两个语义标签之间既可以有直接的联系(如东北虎和动物园)，也可以有间接的联系(如东北虎－动物园－人－房屋)，两个语义标签之间的联系可能存在多条不同的路径，如果路径越多并且路径距离越短，则表示这两个语义标签之间的语义共生性越强。为了更好的描述和定义该知识图谱中的距离，本文采用随机漫步的方法[11]来进行分析，该方法是利用随机漫步算法对候选的语义标签信息进行重排序，在此过程中不仅定义了共生相似度，还对原始的标签

信息集合进行了基于可信度的排序。根据随机漫步原理，假设有一个醉汉在该图中行走，他行走的起始点为 S_i 的概率为 p_i，每当醉汉要继续往下走时，他都有两个选择，一个是随意选择一条邻接的边行走到另一个节点，另一个选择是跳到另一个节点 S_j，假设最后的稳定状态为 u，则稳定状态矩阵 u 满足如下条件：

$$u=(1-c)Au+cr$$

式中 A 是标准化后的图 G 的邻接矩阵，c 是跳到另外一个节点的概率，本文将其设为 0.15，r 是初始的标注节点权重。

通过计算从一个语义标签 S_i 到达另一个语义标签 S_j 的概率 R_{ij} 来描述这两个语义标签的共生性，概率 R_{ij} 越高则表示这两个语义标签之间的路径越到，他们之间的语义共生度 S_{ij} 则越高：

$$S_{ij}=S_{ji}=\sqrt{R_{ij}R_{ji}}$$

最后，为了提高计算效率，本文利用 KNN (K-nearest neighbor)算法对矩阵 S 进行缩减，即如果 S_{ij} 是第 i 行或者第 j 列最大的前 K 个单元，则标签 i 和 j 被认为是 KNN，这样一样可以在减少计算量的同时保留具有最大共生关系的语义标签。

3.3. 知识合并与演化推理

为了剔除冗余信息，首先需要进行实体对齐与消歧。实体对齐是知识图谱构建以及更新过程中的重要工作之一，通过实体对齐，同一个知识图谱内部的实体得到了精简，可以实现知识图谱之间的链接与合并，从而实现构建一个更大规模，服务范围更广泛的知识图谱系统。

实体对齐是对于物理世界中的同一个对象，要识别出它在不同语言，不同地域，不同数据源或者是同一个数据源下不同的表示形式，然后用一个全局唯一的编号来表征。实体对齐算法设计的主要思路是根据具体的知识图谱的特点和处理方法，利用不同的实体识别技术，具体有使用传统概率模型的方法、以及使用机器学习的方法，来完成实体对齐任务。实体消歧是专门用于解决同名实体产生歧义问题的技术。通过实体消歧，就可以根据当前的语境，准确建立实体链接。同义关系是指在概念层面上相同或相似的实体。同义关系抽取的目标是寻找那些字面不同但是指代同一概念、实体或属性的术语[12]。

知识推理是在知识图谱上进行数据挖掘，使知识图谱不断完善的重要手段，主要包括三个方面：第一，线索挖掘；第二，关系推理；第三，关系预测。线索挖掘是指对于知识图谱中原来并没有关系的实体或概念，挖掘出它们之间的关系或关系模式，英文称为 Storytelling。线索挖掘是对于在知识图谱构建过程中没有关联起来的实体进行相关性推理的过程，涉及到的处理方法主要有对于图的各种操作，比如查找子图、查找连通分支等。

随着知识图谱中实体规模的不断扩大，知识图谱中实体的关联，作为知识图谱补全的重要环节，将变得愈来愈重要。同时，由于对实体关联的高效性要求变得愈来愈高，以及知识图谱建设造成的不一致和噪声的干扰，实体关联的任务也会变得越来越复杂，需要研究出更加高效、更具抗噪声能力的实体关联线索挖掘方法[13]。

关系推理是指根据知识图谱中已有的实体之间的关系推断出实体之间潜在的关系。例如基于规则：“父亲的父亲是爷爷”。然后根据已有的实体之间的关系，这里是康熙对于雍正的关系是父亲和雍正对于乾隆的关系是父亲，推断出康熙对于乾隆的关系是爷爷。基于规则的方法，目前常用的方法是机器学习中的归纳逻辑编程技术，包括基于一阶 Horn 子句的方法或一阶归纳逻辑(FOIL)。

3.4. 实验验证

为了检验基于知识图谱的视频语义分析方法的有效性，本文以视频分类为任务进行实验，实验所用

的视频数据是标准视频库 YouTube-8M，评价标准采用的平均精度均值 MAP 和命中率 HIT，比较对象现有的三种视频特征分类的方法 AoFF，DBoF 和 LSTM，这三个模型的是实现是基于 Google 实现(https://github.com/google/youtube-8m)，实验在这三个模型的基础上比较了没有融入知识图谱的分类结果和本文所提出的融入了知识图谱的分类方法 KGS (Knowledge Graph Semantic)，比较结果如表 1 所示。

从表 1 的结果中能看出，融入了知识图谱视频语义的分类结果有效的提高了分类准确度，MAP 平均提高了 1.7%，HIT 平均提高了 1.6%，这一结果证明了基于知识图谱的方法在填补语义鸿沟的有效性。

Table 1. Comparison of video classification results
表 1. 视频分类结果的比较

	AoFF		DBoF		LSTM	
	MAP	HIT	MAP	HIT	MAP	HIT
—	0.370	0.846	0.287	0.834	0.279	0.838
KGS	0.383	0.849	0.302	0.846	0.295	0.856

4. 视频语义知识图谱的应用

利用知识图谱所表达的视频语义内容信息可以对视频语义内容进行深层次的挖掘，除了视频分类外还可以用于视频摘要、视频标注基于语义的视频检索和视频关联分析等。

4.1. 视频摘要和视频标注

视频摘要，主要目的是对视频在内容上进行压缩，使用户能够在短时间内浏览完一段视频内容而不遗漏重要信息。视频标注是对视频内容标注上有用的文本信息，以帮助用户更好的理解视频内容。基于知识图谱建立的视频语义能够更好体现视频各语义标签之间的关联性，形成结构性语义，对于辅助生成视频摘要内容和对视频进行语义标注具有更强的语义表达作用，利用知识图谱技术在一定程度上克服了自然语言的歧义性，把经过梳理、总结的知识提供给用户，更加清晰、动态的方式展现了各种概念之间的联系。

4.2. 基于语义的视频检索和视频关联分析

基于语义的检索对于克服图像信息中的语义鸿沟具有重要的作用，基于知识图谱生成的图像语义框架可以更好的服务于语义检索领域，这是由于与传统的基于关键字匹配的搜索引擎工作原理不同的是，知识图谱利用概念、实体的匹配度返回给用户与搜索相关的更全面的知识体系。

语义检索是基于之前的语义组织体系，实现知识关联和概念语义检索的智能化检索方式。知识图谱中的语义检索包含两类核心任务：一是利用相关性在知识库中找到相应的实体；二是在此基础上根据实体的类别、关系及相关性等信息找到关联的实体[14]。通过对知识库进行深层次的知识挖掘与提炼后，检索系统为用户反馈出具有重要性排序的准确且完整的知识，并推荐用户感兴趣的相关知识。

语义关联分析的基本任务是根据主题、形式、自然属性、社会属性等，链接具有相似语义信息的图像等视觉媒体，在各种跨媒体关联类型中最关键的是关联数据模型。以知识图谱作为基础构建数据模型，能够更好地实现传统数据模型所不能支持的多种智能分析，时空关联分析、逻辑关联分析、语义相似性搜索、数据世系管理与分析、数据溯源与核查等，提升各种多媒体信息之间的关联分析能力。

5. 结论

本文所提出的基于知识图谱的视频语义分析方法可以增强对视频语义的理解，填补视觉特征与内容之

间的语义鸿沟，具有重大的价值和研究意义。目前利用知识图谱实现对视频等视觉媒体的语义分析研究还处于初级阶段，仍然存在很多的挑战和难题需要解决，例如知识图谱推理规则的学习等。知识图谱在知识组织和展现上体现出来的优势是非常显著的，在未来的多媒体语义分析领域将扮演越来越重要的角色。

参考文献

[1] Abu-El-Haija, S., Kothari, N., *et al*. (2016) Youtube-8M: A Large-Scale Video Classification Benchmark. *Computer Science*, arXiv preprint arXiv 2016:1609.08675.

[2] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., *et al*. (2015) Beyond Short Snippets: Deep Networks for Video Classification. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 4694-4702. https://doi.org/10.1109/CVPR.2015.7299101

[3] Jordan, M.I. and Jacobs, R.A. (1994) Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, **6**, 181-214. https://doi.org/10.1162/neco.1994.6.2.181

[4] Wang, Z., Kuan, K. Ravaut, M., *et al*. (2017) Truly Multi-Modal Youtube-8M Video Classification with Video, Audio, and Text. *Computer Science*, arxiv preprint arxiv 2017:1706.05461.

[5] 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践, 2015, 38(12): 13-18.

[6] 邓莉琼, 张贵新, 郝向宁. 基于知识图谱的图像语义分析技术及应用研究[J]. 计算机科学与应用, 2018, 8(9): 1364-1371.

[7] Fang, Y., Kuan, K., Lin, J., Tan, C. and Chandrasekhar, V. (2017) Object Detection Meets Knowledge Graphs. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 19-25 August 2017, 1661-1667.

[8] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 36(9): 2508-2515.

[9] Wang, H., Klaser, A., Schmid, C. and Liu, C.-L. (2011) Action Recognition by Dense Trajectories. 2011 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 20-25 June 2011, 3169-3176. https://doi.org/10.1109/CVPR.2011.5995407

[10] Paulheim, H. (2017) Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web*, **8**, 489-508. https://doi.org/10.3233/SW-160218

[11] Tong, H., Faloutsos, C. and Pan, J. (2006) Fast Random Walk with Restart and Its Applications. *Sixth International Conference on Data Mining*, Hong Kong, 18-22 December 2006, 613-622. https://doi.org/10.1109/ICDM.2006.70

[12] 孙霞, 董乐红. 基于监督学习的同义关系自动抽取方法[J]. 西北大学学报, 2008, 38(1): 35-39.

[13] 李跃鹏, 金翠, 及俊川. 基于 Word2vec 的关键词提取算法[J]. 科研信息化技术与应用, 2015, 6(4): 54-59.

[14] 杨思洛, 韩瑞珍. 知识图谱研究现状及趋势的可视化分析[J]. 情报资料工作, 2012, 33(4): 22-28.