一种多模态跨媒体检索的融媒影视系统

李春芳* 刘永久 王楷翔 杨睿 张凌飞 李敏 邓智铭 石民勇

(中国传媒大学计算机与网络空间安全学院 北京 100024)

摘 要：视频是最有影响力的传播媒介，然而其非线性检索仍然困难。本文集成创新性工作包括：基于图像识别提取字幕，基于卷积神经网络识别人脸，通过字幕和人脸解决了影视视频的非线性检索问题；从字幕文本提取重要实体，用海量知识库和电子书补全影视关联知识，构建了文本、电子书和视频融合的跨媒体应用；以字幕词云和人物实体词云，实现影视的概览理解和检索导航；以众包实现字幕、电子书、人脸和实体信息的修正。以近代史献礼电影、中国诗词大会和科技纪录片为例系统完整地实现了一个示范性融媒影视系统。

关键词：融媒体；跨媒体检索；字幕识别；人脸识别

中图分类号：TP302 **文献标识码**：A **文章编号**：

DOI：10.11999/JEIT××××××

A Film and Television Media Convergence System Based on Multi-Modal Cross-Media Retrieval

LI Chunfang, LIU Yongjiu, WANG Kaixiang, YANG Rui, ZHANG Lingfei, LI Min, DENG Zhiming, SHI Minyong

(*School of Computer Science and Cybersecurity, Communication University of China, Beijing 100024, China*)

Abstract: Video is the most influential media, but it's difficult to nonlinearly search video content. The integrated creative work of this paper includes: Based on image processing to recognize video subtitle and convolutional neural networks to recognize faces of characters, the problem of film and TV video nonlinear retrieval is solved. Further, we extract important entities from subtitle text and enhance their relevant knowledge with large scale knowledge base and e-books, which constructs a cross-media application system of video, text, and e-book. Word cloud of subtitles and character entities are designed to facilitate video overview understanding and navigating retrieval. Crowdsourcing technology is used to update the amendments of subtitles, e-books, face recognition and entities information. A typical cross-media convergence system are completely implemented including movies in modern history, conference of the Chinese poetry, and information technology documentary video.

Key words: Media convergence; Cross-media retrieval; Subtitle recognition; Face recognition

1 引言

媒体融合发展已上升至国家战略，影像为王的媒介时代，有视频有真相。视频时序播放特点，知识密集型视频，如纪录片、正史影视，受众不能在呈现的几秒内理解视频的全部信息，另一方面用户也常感到线性视频观看信息过少浪费时间。2018 年 11 月，教育部、中宣部印发《关于加强中小学影视教育的指导意见》，然而如何找到与课程内容密切关联的影视作品及视频片段，成为制约应用的瓶颈。

本文着重研究了面向应用场景的视频字幕提取和人脸识别，对重要实体，链接外部知识库和电子课本，对视频做知识增强，支持视频非线性检索，构建一种富信息融媒影视新形式，满足深度知识获取，改善用户收视体验。面向教育文化传播，以近代史电影、中国诗词大会和科技记录片三个场景实现视频融媒应用，尝试应对网络时代的文明恐慌，为新型主流媒体智能化发展赋能。

收稿日期：2020-10-23；改回日期：2020-××-××；网络出版：2021-××-××

*通信作者：李春芳 lcf@cuc.edu.cn

基金项目：国家社科基金艺术学项目资助(18BC034)

Foundation Items：The National Social Science Art Project of China(18BC034)

2 相关研究

2.1 字幕提取

字幕形式的对白或解说词，有场景说明、画面补充、深化内涵的作用，可用于视频非线性检索。字幕提取包括：字幕事件检测、字幕区域定位、字幕分割、基于 OCR(Optical Character Recognition)的文本识别。

字幕识别首先将视频生成尽可能不重复不遗漏的字幕图像序列。从视频提取字幕帧的方法包括三种：逐帧、等帧间隔、帧差法（或字幕事件检测）。从单张图像检测文本区域的方法大致分为四种：基于纹理特征，基于边缘特征，基于连通域和基于深度学习的方法。

2012 年，曹喜信等研究了基于边缘强度的字幕提取 [1]。2017 年，袁闻研究了网络视频字幕关键词提取与检索 [2]。2018 年，石民勇、艾莫尔夫等研究了抽帧和图像分割的字幕提取 [3]，王智慧等提出了先监测字幕帧再锁定区域的字幕提取方法 [4]。

从英文文献看，侧重对字幕和视频的融合应用。2018 年，吕金娜等用识别人脸和字幕实现了一个 StoryRoleNet，自动构建影视剧的人物关系 [5]。2019 年，Tapu 等基于人脸识别、视频分镜、语音识别及字幕识别，把字幕文本标注到说话人附近，实现了帮助聋哑人看视频的 Deep-Hear 系统 [6]。2020 年，旷视科技 Wan Zhaoyi 等提出一种针对泛场景文字识别的深度神经网络方法 TextScanner[7]。

与深度学习方法相比，基于边缘特征定位字幕区几乎无学习代价，轻量简洁。本文基于等帧间隔和帧差法，利用多帧字幕边缘特征的统计特性，提高字幕块定位精度和效率。

字幕块文字识别由 OCR 处理。2020 年百度基于深度学习的 OCR 识别率达 99%，并提供云端 API。此外中文识别还包括汉王 OCR、文通 OCR 和开源 OCR 引擎 Tesseract。本文字幕 OCR 采用了 Tesseract。

2.2 人脸识别

从字幕文本可检索包含关键词的视频时间点，然而存在大量画面人物和字幕人物不一致情况，如字幕包含“毛泽东”的画面，大部分是他人的对白中提到“毛泽东”，为此需要基于人脸识别检索画面。

人脸识别包括：人脸检测，人脸对齐和人脸识别。人脸识别包括 1:1 比较的人脸验证和 1:k 比较的人脸识别，影视人脸识别是一个 1:k 问题。2014 年 Facebook 的研究者提出了 DeepFace，用三维人脸对齐，交叉熵作为损失函数，在人脸库 LFW(Labeled Faces in the Wild)上识别率达到 97.35% [8]。2015 年，Google 的研究者提出了 FaceNet，构建（图像，正例，反例）三元组，人脸图像与正例距离近与反例距离远作为目标函数的训练方法，在 LFW 上识别率达到 99.65%[9]。2016 年，Google 提出了 GoogLeNet 的升级版 Inception-ResNet，PyTorch 实现该算法用于人脸识别 [10]。2016 年，Zhang Kaipeng 等提出构建图像金字塔，将人脸检测与人脸关键点对齐的多任务 MTCNN 模型 [11]。此外，还可采用视频 ReID 技术跟踪识别人脸 [12]。

随着算法到 API 的快速迭代，专家一致认为，AI 创新重点在于应用场景，然而技术远没被应用到主流视频媒体，大量制作精良的视频不能被便利地检索、挖掘和传播，传统媒体内容王者的地位受到严峻挑战。

2.3 跨媒体语义检索

跨媒体检索旨在以任意媒体数据检索其他媒体的相关数据，实现图像、文本等不同媒体的语义互通和交叉检索。2018 年，彭宇新综述了跨媒体检索的概念方法和挑战 [13-14]，认为学习图像和文本间精确的关联关系，提高跨媒体检索准确率。同年，王述和史忠植研究了基于深度典型相关性分析的跨媒体语义检索，从多媒体数据中抽取概念及标签训练，语义映射实现跨媒体检索 [15]。2019 年，卓昀侃等提出跨媒体循环神经网络，挖掘包括图像、视频、文本、音频和 3D 模型的细粒度信息，提升了跨媒体检索的准确率 [16]。

2018 年，许斌团队自动抽取加众包构建了小初高全学科基础教育知识图谱 edukg.cn，用于智慧教育 [17]。与跨媒体理论研究相比，本文工程上实现了一个跨媒体检索系统；与教育知识图谱图文表达相比，本文是以视频为核心的融媒系统。

以下分别论述视频字幕提取、视频的人脸识别、电子书识别，以及集成实现的融媒影视系统。

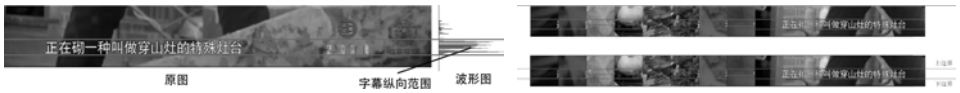
3 基于统计特征的视频字幕提取

本节利用字幕区的边缘统计特征，设计实现了一个高效高识别率的字幕提取算法，并分析了实验结果。

3.1 多帧边缘统计特征用于确定字幕上下边界

图 1(a)是字幕区域 Y 方向的边缘特征构造的二值矩阵的行和，可以明显的分辨出字幕的上下边界。在字

幕帧字数少，且遇到特殊文字，单独取一帧定位不准确。为此，采用多帧字幕统计特征，即取众数（众数，指在统计分布上具有明显集中趋势点的数值，代表数据的一般水平，也是一组数据中出现次数最多的数值），见图 1(b)，多帧字幕大多数的上下边界作为整个视频字幕上下边界，剔除了字形差异的干扰。



(a) 边缘特征二值矩阵的行和　　(b) 不同字幕的边缘特征确定的上下边界

图 1　基于多帧字幕边缘特征定位字幕上下边界

基于以上分析，初始化先确定字幕上下边界。随机选择视频中的 N 帧（N=50），取帧图像的下 1/5 和左 1/2 区域，对该区域做灰度化、中值滤波、用 Sobel 算子提取 Y 方向的边缘特征，进一步二值化（阈值可调，默认 150），构建一个边缘特征存在与否的 one-hot 二值矩阵，计算行和，从行和最大值逐像素向两端滑动检测当前帧的上下边界。对 N 帧样本的边界统计，用大多数帧（众数）上下边界作为视频字幕上下边界。

3.2 基于多帧统计特征确定字幕对齐方式

影视字幕对齐方式分两种，左对齐和居中对齐，即非左即中。随机抽取多帧字幕，灰度化、二值化，用二值化 one-hot 矩阵的列和确定字幕左边界，从列和最大的像素点开始向左按字宽滑动，左侧边界比较集中判断为左对齐，非常分散判断为居中对齐。图 2 所示，从最大的列和开始向左滑动获得左边界。

图 2　基于二值化矩阵列和确定字幕对齐方式

3.3 基于统计特征的视频字幕定位算法

算法 1，输入为影视视频文件，输出字幕格式文件。

算法 1　基于统计特征的视频字幕提取算法

输入: 带有字幕的视频文件（如*.mp4）

输出: 字幕文件.srt

① 初始化：统计多帧经 Sobel 算子生成边缘 one-hot 矩阵确定视频字幕上下边界。

② 初始化：根据多帧字幕统计特征确定左边界，确定对齐方式。

③ 每隔 0.5 秒读取视频的一帧，根据上下边界和对齐方式，确定左右边界，确定是否为字幕帧。非字幕帧则丢弃，继续循环③。

④ 计算当前字幕图像灰度化、二值化矩阵，one-hot 矩阵的中间行与上一帧字幕图像中间行的余弦距离，如果余弦距离>0.7 认为是重复字幕，丢弃，跳转③。

⑤ 根据上下和左右边界分割图像取出当前帧的字幕区域，经灰度化、色阶调整、二值化、黑白翻转、得到白背景黑字的字幕图像。

⑥ 对判定为非重复的字幕帧，经 OCR 识别输出文本。

⑦ 字幕区域图像生成的文本行经正则表达式过滤非中文和数字字符乱码，经莱温斯坦（Levenshtein）字符编辑距离再次去重。

⑧ 计算字幕帧的毫秒时间，按字幕格式写入字幕文件.srt。

⑨ 判断是否超过视频长度，是则结束，否则转③继续提取下一个可能的字幕帧文本。

字幕定位算法的流程参见图 3 所示，对算法 1 的说明如下：

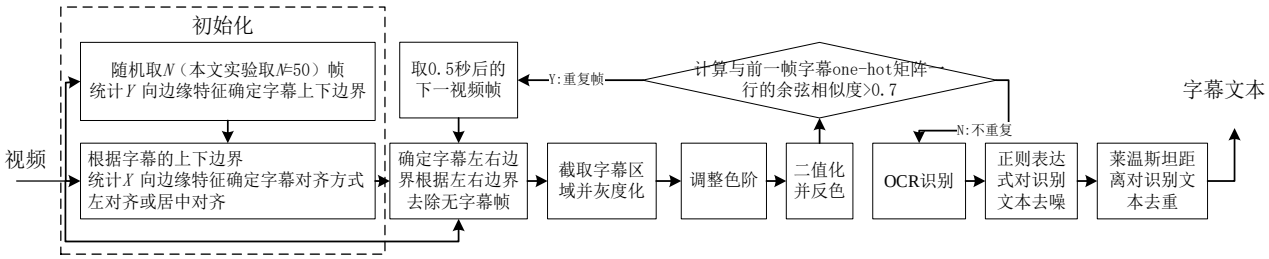


图 3　字幕定位算法流程图

(1)步骤③参数 0.5 秒的选择由实验统计确定。根据统计规律，字幕行停留时间一般在 0.5-7 秒，识别原则是不丢字幕帧并尽可能减少重复字幕帧。

(2) 步骤③会有极少量的无字幕帧被判为有字幕，原因是背景纹理过于复杂造成的干扰，这样的无字幕帧经 OCR 识别为乱码，通过正则表达式滤除。

(3) 步骤④重复字幕帧的判定。依据拥有相同字幕的图像帧，必然有极为相似的字幕边缘特征，对比两帧字幕区域 Y 轴方向边缘 one-hot 矩阵中间行向量的余弦相似度，判断字幕是否重复，本文设定余弦相似度>0.7，为相同字幕帧，重复字幕检测波形参见图 4。此处仍可能产生少量的重复字幕，后续再次去重。

(4) 本算法没有单独处理字幕事件检测，目的是通过抽帧提高识别效率。通过余弦相似度判断抽帧时刻字幕是否改变，图 4 的波形图和字幕序列为《舌尖上的中国》的 600 帧，每 12 帧取一帧，取 50 帧作为样本，共有 11 个波峰，即 11 个对比的抽帧中 one-hot 矩阵中间行的余弦相似度>0.7，每个波峰表示一组相同的字幕，代表了一条不同的字幕，共 12 条不同字幕，波形跳变与字幕一致，是对帧差去重复的直观解释。

图 4　重复字幕检测的波形图

(5) 步骤⑤当判定字幕区域包含字幕且和上一帧不同，对字幕区域灰度化处理。色阶是用直方图描述整张图像的明暗信息。色阶调整使字幕图像与背景色调分离，提高字幕辨识度，如公式(1)所示，含三个参数：像素灰度值 Input，高光值 Highlight 和阴影值 Shadow，该像素输出值 Output。

$$\mathrm{Output}=\begin{cases}0 & \mathrm{Input}-\mathrm{Shadow}<0\\ 255*\dfrac{\mathrm{Input}-\mathrm{Shadow}}{\mathrm{Highlight}-\mathrm{Shadow}} & \mathrm{Input}-\mathrm{Shadow}\geq 0\end{cases} \tag{1}$$

实验表明色阶调整对 OCR 识别率影响较大。图 5 是视频一帧灰度图调整色阶前后对比，并把字幕区域突出显示。可以看出，色阶调整后，图像的对比度下降，但是字幕辨识度改善。以《互联网时代》为例，色阶调整字幕图像可以使得 OCR 识别率由 70%提升到 95%以上。

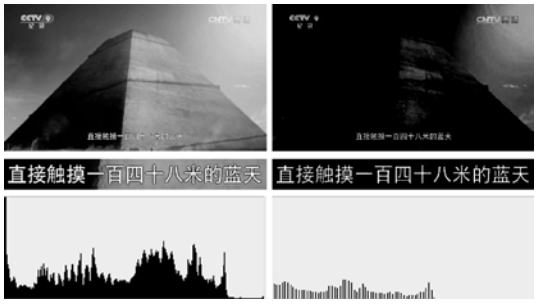


图 5　色阶调整对图像和文字清晰度的影响

(6) 步骤⑤对字幕图像二值化，本文设定灰度>150 映射为 255，否则为 0，得到黑背景的白色文字，再反色处理，得到白背景黑字。定位后处理过程参见图 6，可以看出有效剔除了背景干扰。

图 6　字幕定位后对字幕图像处理示例

(7) 步骤⑦依据正则表达式剔除乱码。使用 OCR 识别文字，仍有部分重复字幕或无字幕的乱码。为提高识别精度，本文针对单一语言字幕识别，在 OCR 识别中文时将出现的标点、符号、英文字符等视为噪声。

大部分中文内容表示范围是[u4e00-u9fa5]，且字幕大都不包括标点，但有数字。因此本文根据 Unicode 的中文编码表，用正则表达式 re.compile(r'[^\u4e00-\u9fa5+0-9]+')匹配，结果只保留中文字符和数字。对于英文字幕使用 re.compile(r'^\w+$')，去除中文和乱码。对于中英文混合双语字幕两次分别识别。

(8) Levenshtein 距离是编辑距离中经典算法，指一个字串转成另一字符串所需的最少编辑次数。编辑操作包括：替换、插入和删除字符。如：将“中央电视台”转化为“中央广播电视总台”，编辑距离为 3，通过插入 3 个字符完成。步骤⑦依据 Levenshtein 编辑距离和字符串相似度过滤 OCR 后的少量重复字幕。

3.4 字幕提取实验

3.4.1 实验环境

实验采用 Python3.7 和 OpenCV，主要函数包括 VideoCapture、cvtColor、medianBlur、Sobel、threshold，分别用于读取视频、灰度化、中值滤波、提取特征边缘和二值化操作。所用 OCR API 为 Tesseract-OCR4.0.0。

3.4.2 字幕块识别率

本文用 5 部中文和 2 部英文视频作为实验数据。对识别字幕块定义：查全率=正确识别字幕条数/字幕总条数，查准率=正确识别字幕条数/识别字幕条数。实验如表 1 所示，以《建军大业》为例，总字幕 1750 条，查全率 99.83%，漏识别 3 条，查准率 98.20%。英文片《The Lion King》的查全率为 99.72%，查准率为 99.81%。

表 1　字幕条数提取实验结果

视频	语言	类型	字幕总条数	识别条数	正确字幕数	查全率	查准率
《中国通史》贞观之治	中文	纪录片	917	926	913	99.56	98.60
《舌尖上的中国 3》-1	中文	纪录片	916	935	912	99.56	97.54
《互联网时代》-4	中文	纪录片	862	899	859	99.65	95.55
《探寻人工智能》1-1	中文	纪录片	539	550	536	99.44	97.45
《建军大业》	中文	电影	1750	1779	1747	99.83	98.20
60 Years of innovation	英文	短视频	62	62	62	100	100
The Lion King	英文	电影	1075	1074	1072	99.72	99.81

3.4.3 字幕文字识别率

开源 OCR 引擎 Tesseract 的中文识别率约为 97%。本文文字识别率实验如表 2 所示，《中国通史》贞观之治的文字查全率 95.81%，查准率 95.43%，《舌尖上的中国》单集文字查全率 95.92%，查准率 94.33%，《互联网时代》文字查全率 96.04%，查准率 94.20%。《建军大业》共 11767 字，查全率 98.6%，查准率 97.73%。

表 2　字幕文字识别率实验结果

视频	时长(min)	分辨率	文字数	共识别	正确识别	查全率	查准率
《中国通史》贞观之治	44:17	1280*720	7644	7675	7324	95.81	95.43
《舌尖上的中国 3》-1	47:58	1280*720	7322	7445	7023	95.92	94.33
《互联网时代》-4	49:58	1280*720	8236	8397	7910	96.04	94.20
《探寻人工智能》	26:46	1920*1080	4729	4851	4492	94.99	92.60
《建军大业》	02:13:04	1280*536	11767	11872	11602	98.60	97.73
60 Years of innovation	03:12	1920*1080	538	538	538	100	100
The Lion King	01:21:51	1280*536	7377	7383	7348	99.61	99.53

本文实验数据规模远高于已有文献，表 3 中与文献[1]和[4]相比，本文中英文字幕块查全率最高，达到

99.65%以上，中文平均查准率达到 97.6%，英文查准率达到 99.8%。

表 3　字幕块提取与已有文献的对比

提取方法	语言	平均查全率	平均查准率
曹喜信等 [1]	中文	98.0	98.5
王智慧等 [4]	中文	99.44	99.56
	英文	99.44	99.56
本文方法	中文	99.65	97;60
	英文	99.70	99.80

4 影视视频人脸识别

为实现影视人脸识别，以《建军大业》为例，在豆瓣爬取主要角色照片，以“编号-演员名-角色名”格式存储，用于人脸识别。主要角色及演员共 57 名，部分数据如表 4 所示。

表 4　《建军大业》人脸识别的演员与角色

序号	照片	演员	角色
1		刘烨	毛泽东
2		朱亚文	周恩来
3		黄志忠	朱德
4		王景春	贺龙

统计角色出场时间流程如图 7 所示。输入原视频，每隔 0.2 秒抽一帧，若当前帧检测到人脸，则用演员照片识别对应角色，记录帧时刻，若没有检测到人脸，继续向后抽帧，识别结果以.srt 字幕格式存储。

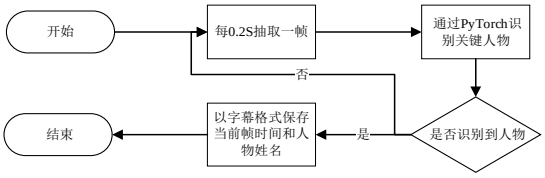


图 7　统计角色出场时间流程图

PyTorch 以高度易用被工程中广泛采用，本文采用其实现的人脸检测和对齐一体的 MTCNN[11]算法和 Inception-ResNet[10]算法实现人脸识别，挂载的预训练参数为 VGGFace2。

视频检索在秒级精度即可，本文忽略毫秒把人脸识别的起始时间和字幕起始时间对应，写入数据表，实现基于字幕和人脸并行的视频非线性检索。表 5 中第一行指在该字幕处，画面出现人物“毛泽东”和“周恩来”。对《建军大业》识别角色人脸，对应到字幕时间，共 540 条字幕附近有角色出现，检索正确率为 98%。

表 5　《建军大业》字幕和人物出场时间表

字幕 ID	时间	字幕	人物
340	00:27:41，440 --> 00:27:43，271	工人阶级是革命的主体	毛泽东 周恩来
341	00:27:43，640 --> 00:27:45，517	这是马克思的基本观点，不对吗？	周恩来
342	00:27:46，280 --> 00:27:48，635	我认为中国不一样	毛泽东

5 电子书 PDF 的数据化

为实现电子书与影视视频的跨媒体关联检索，需要对图像格式的电子书数据化。处理流程见图 8，基于 Python Wand 库和 C++的 ImageMagick 对 PDF 电子书逐页转为图片，图片经过灰度化、边缘提取、二值化、两次膨胀和腐蚀，聚合成一个文字框或者图片区域，获取轮廓后生成切块，并滤掉噪声小块，切割文字或图片区域，记录块的坐标，对切割后的文字区域，经 Baidu-Aip 的 OCR 识别为文字。

表 6 为 3 本电子书数据化的实验结果，以初中历史八年级上册为例，132 页，采用 72×72 分辨率，拆为图片用时 32 秒，用时 5.9 秒划分为 660 个图文块，对其中文字块 OCR 识别共用时 729.2 秒，手工随机抽取 5 页检测，识别率约 98.87%。《人工智能简史》OCR 识别率为 99.18%。

图 8　图像格式电子书 PDF 的数据化

表 6　电子书数据化实验结果

电子书	页数	分辨率	分页时间	分块数	分块时间	OCR 时间	OCR 识别率
小学语文六年级下册	124	72×72	25	314	10	780.3	98.87%
初中历史八年级下册	132	72×72	32	660	5.9	729.2	98.56%
人工智能简史	326	300×300	71.6	1138	10.6	1159.2	99.18%

6 基于多模态跨媒体检索的融媒影视架构

6.1 系统架构

本节设计实现了基于字幕提取、人脸识别、电子书数据化、词频统计的支持视频内容理解、非线性检索和知识增强的融媒影视系统。架构见图 9，演示地址 www.yingshinet.com。

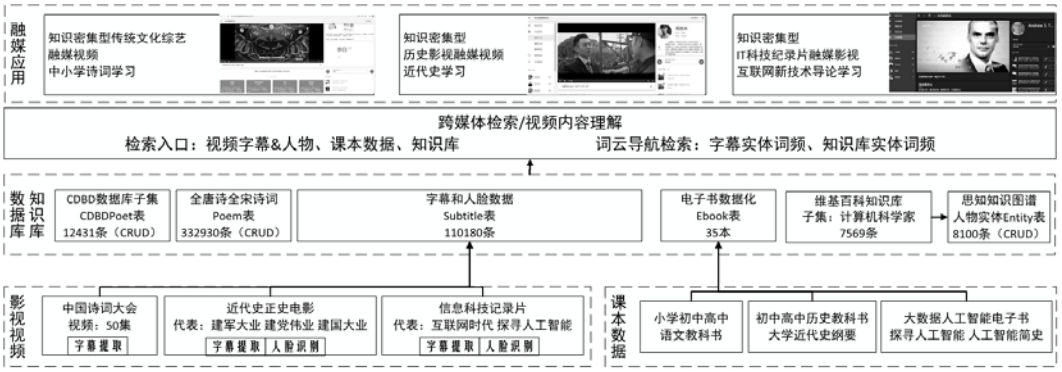


图 9　多模态跨媒体检索的融媒影视系统架构

系统以 3 个应用为例构建了融合知识库的数据库，实现跨媒体检索，以下分别论述实现过程。

6.2 近代史融媒影视的跨媒体语义检索

图 10(a)是《建军大业》视频，左下是主要历史人物的字幕加人脸数，右上是人物信息，抽取自思知(Ownthink)知识图谱，存入本地数据库。视频下方是课本图片和数据化的文字，对人物实体添加链接，点击实现跨媒体检索。右下是字幕检索区，显示了字幕第一帧、时间和文字。

(a)　《建军大业》融媒影视跨媒体检索　　(b)　《中国诗词大会》融媒综艺的跨媒体检索

图 10　支持非线性检索的融媒影视系统

为提高跨媒体检索的准确率，根据对白特点对人物实体添加了检索别名。人物实体名词、别名和人脸识别三者的语义一致，采用“或”关系查询提高了检索准确率。在《建军大业》中“毛泽东”的别名为“润之”，检索字幕查询到 12 条，检索别名返回 4 条，检索人脸返回 139 条，总计 155 条，总数与献礼电影主题一致，角色戏份代表领袖人物的历史地位。在数据化的电子课本中对重要实体添加超链接，实现从电子书文字检索视频 e-book2video，解决了中小学影视教育中与教材关联的视频片段查找难题。

使用人物词云和字幕词云导航检索，实现点击鼠标代替键盘输入，并提供了对视频的概览理解。对单片视频字幕文本分词、统计词频，生成字幕词云导航检索，参见图 10，点击词条返回跨媒体协同检索结果。

6.3 综艺融媒视频《中国诗词大会》

《中国诗词大会》在诗词选择上力求达到“熟悉的陌生题”，强化普及性，增强参与感和代入感，然而摘句寻章的明显不足是放弃了整首诗词的文化意蕴和艺术奥妙。

本节以《中国诗词大会》1-5 季共 50 集视频为例，通过字幕提取（Subtitle 表），融合全唐宋诗词库（Poem 表，33.2 万）、中小学语文课本（Ebook 表，22 本）、哈佛大学的中国历代人物传记资料库（抽取了诗人子集构建 CDBDPoet 表，1.24 万诗人），通过字幕实现视频与知识库的跨库协同，构建了一种富信息融媒综艺视频，参见图 10(b)，视频播放时下面显示与字幕诗句同步的整首诗词和诗人作品，以知识补全解决视频节目中摘句寻章的不足，提供跨媒体关联理解。

6.4 互联网科技融媒纪录片

科技纪录片是典型知识密集型视频。以《互联网时代》为例，汇聚全球 14 个国家互联网领域 200 多位重要人物观点，形成宏观视角、全景式描绘，构建一部互联网史料纪录片，极具重复学习和反复使用价值，然而线性检索限制了传播，查找文字和人物都非常困难。

本节对纪录片提取字幕和对重要人物做人脸识别，实现视频的非线性检索。从维基百科抽取了计算机科学家实体做知识增强，写入 Entity 表，同时用电子书对视频做跨媒体的佐证和补充。

6.5 基于众包的数据校正

尽管字幕文字查全率超过 95%，电子书识别率超过 98.5%，但是错误率需要低于出版标准的 0.01%。本节采用基于众包的人工校对和审核，采用多数人投票原则，即 2 人以上修改相同自动审核通过，不足 2 人的修改等待管理员人工审核。

6.6 知识库管理

对人物实体的增删改查，设计了数据管理模块。依据数据来源的权威性，按照课本、教师用书、思知图谱和 CDBD 的顺序修改，并尽可能提供数据来源说明。对于诗人、诗词设计了增删改查管理，扩充唐朝以前的诗词、明清诗词和毛泽东诗词等。

7 小结

本文融合字幕识别、人脸识别、电子书识别，实现对影视视频的内容理解、非线性检索和知识增强，构建了一个跨媒体协同的视频融媒播放系统。主要工作包括：(1) 提出了一种基于多统计特征的字幕提取方法；(2) 设计了以字幕格式为基准的人物和字幕协同的非线性影视检索方法，解决视频内容检索难题，通过视频溯源课本，通过课本概念定位视频起点；(3) 实现了知识库和电子书协同检索和知识增强的融媒影视播放系统，解决视频信息补全问题，实现视频与多源知识库的跨媒体检索。本文研究可用于影视作品制播后深度开发和传播，提供了一种教育教学中应用影视视频的便利形式，为主流媒体的融媒体全媒体传播赋能。

跨库检索的难题是语义对齐，本文基于隐形的跨媒体公共子空间实现协同检索，后续构建显性的公共子空间、探索语义理解的智能搜索匹配和基于知识图谱三元组的关联检索。

参 考 文 献

[1] CAO Xixin, LIU Jing, YANG Xudong, *et al*. A novel algorithm for the video caption extraction[J], *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2013, 49(02): 197-202.

[2] 袁闻. 网络视频字幕中关键词的提取与检索技术研究[D]. [硕士论文]. 北方工业大学, 2017.
YUAN Wen. The research on the extraction and retrieval of keyword in network video subtitles[D]. [Master’s thesis]. *North China University of Technology*, 2017.

[3] 石民勇, 艾莫尔夫, 等. 一种基于图像分割及动态阈值的字幕提取方法[P]. 中国, 109271988A, 2019-01-25.
SHI Minyong, AIMOERFO, et al. An approach of subtitle extraction based on dynamic threshold and image segmentation[P], China, 109271988A, 2019-01-25.

[4] 王智慧, 李佳桐, 谢斯言, 等. 两阶段的视频字幕检测和提取算法[J]. 计算机科学, 2018, 45(08): 50-53.
WANG Zhihui, LI Jiatong, XIE Siyan, *et al*. Two-stage method for video caption detection and extraction[J]. *Computer Science*, 2018, 45(08): 50-53.

[5] LV Jinna, WU Bin, ZHOU Lili, *et al*. StoryRoleNet: social network construction of role relationship in video[J]. *IEEE Access*, 2018: 25958-25969. doi: 10.1109/ACCESS.2018.2832087.

[6] TAPU Ruxandra, MOCANU Bogdan, ZAHARIA Titus, *et al*. DEEP-HEAR: a multimodal subtitle positioning system dedicated to deaf and hearing-impaired people[J]. *IEEE Access*, 2019: 88150-88162. doi: 10.1109/ACCESS.2019.2925806.

[7] WAN Zhaoyi, HE Minghang, CHEN Haoran, *et al*. TextScanner: reading characters in order for robust scene text recognition[OL]. https://arxiv.org/pdf/1912.12422, 2020.10.

[8] TAIGMAN Yaniv, YANG Ming, RANZATO Marc'Aurelia, *et al*. DeepFace: closing the gap to human-level performance in face verification[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1701-1708. doi: 10.1109/CVPR.2014.220.

[9] SCHROFF Florian, KALENICHENKO Dmitry, and PHILBIN James. FaceNet: A unified embedding for face recognition and clustering[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 815-823. doi: 10.1109/CVPR.2015.7298682.

[10] SZEGEDY Christian, IOFFE Sergey, VANHOUCKE Vincent, *et al*. Inception-v4, Inception-Resnet and the impact of residual connections on learning[OL]. https://arxiv.org/abs/1602.07261, 2020.10.

[11] ZHANG Kaipeng, ZHANG Zhanpeng, Li Zhifeng, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks[J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499-1503. doi: 10.1109/LSP.2016.2603342.

[12] SHENG Hao, ZHENG Yanwei, Liu Yang, *et al*. A heuristic transformation in discriminative dictionary learning for person re-identification[J]. *IEEE Access*, 2019: 40313-40322. doi: 10.1109/ACCESS.2019.2905552

[13] PENG Yuxin, HUANG Xin, ZHAO Yunzhen. An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 28(9): 2372-2385. doi:10.1109/TCSVT.2017.2705068.

[14] 彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望[J]. 计算机研究与发展, 2019, 56(01): 183-208. Doi:10.7544/issn1000-1239.2019.20180770.
PENG Yuxin, QI Jinwei, HUANG Xin. Current research status and prospects on multimedia content understanding[J]. *Journal of Computer Research and Development*, 2019, 56(01): 183-208. doi:10.7544/issn1000-1239.2019.20180770.

[15] 王述, 史忠植. 基于深度典型相关性分析的跨媒体语义检索[J]. 中国科学技术大学学报, 2018, 48(04): 322-330. doi: 10.3969/j.issn.0253-2778.2018.04.008.
WANG Shu, SHI Zhongzhi. Cross-media semantic retrieval with deep canonical correlation analysis[J]. *Journal of University of Science and Technology of China*, 2018, 48(4): 322-330. doi: 10.3969/j.issn.0253-2778.2018.04.008.

[16] 卓昀侃, 綦金玮, 彭宇新. 跨媒体深层细粒度关联学习方法[J]. 软件学报, 2019, 30(04): 884-895. doi: 10.13328/j.cnki.jos.005664
ZHOU Yunkan, QI Jinwei, PENG Yuxin. Cross-media deep fine-grained correlation learning[J]. *Journal of Software,* 2019, 30(4): 884-895. doi: 10.13328/j.cnki.jos.005664.

[17] 杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法. 软件学报, 2018, 29(10): 2931-2947. doi: 10.13328/j.cnki.jos.005552.
YANG Yuji, XU Bin, HU Jiawei, *et al*. Accurate and efficient method for constructing domain knowledge graph. *Journal of Software*, 2018, 29(10): 2931-2947. doi: 10.13328/j.cnki.jos.005552.

李春芳：女，1974 年生，副教授，研究方向为视频内容理解、智能影视大数据、软件工程。
刘永久：男，1991 年生，硕士生，研究方向为视频内容理解、字幕识别.
王楷翔：男，1996 年生，硕士生，研究方向为自然语言处理、情感分析和知识图谱.
杨　睿：男，1997 年生，硕士生，研究方向为数字娱乐与动画技术.
张凌飞：女，1997 年生，硕士生，研究方向为情感分析、视频内容理解.
李　敏：女，1998 年生，硕士生，研究方向为视频图文包装、智能媒体技术.
邓智铭：男，1997 年生，硕士生，研究方向为视频内容理解、文字识别.
石民勇：男，1962 年生，教授，研究方向为智能影视大数据、游戏动画.