

面向语义挖掘的蔬菜病害知识视频场景检测

温皓杰^{1,2} 周 婧³ 傅泽田^{4,5} 张领先^{3,5} 严 谨³ 李鑫星^{3,4}

(1. 中国农业大学重大设施筹建办, 北京 100083; 2. 中国农业大学工学院, 北京 100083;
3. 中国农业大学信息与电气工程学院, 北京 100083; 4. 食品质量与安全北京实验室, 北京 100083;
5. 中国农业大学农业部农业信息获取技术重点实验室, 北京 100083)

摘要: 以蔬菜病害知识视频为对象,研究了视频镜头分割、标注方法和聚类模型,构建了面向语义挖掘的语义场景检测模型。基于多模态融合的视频分割技术建立了蔬菜视频镜头库;结合蔬菜知识中文词典,基于多模态融合的视频内容识别方法对蔬菜视频镜头库进行文本标注;对识别结果进行基于知网的语义相似度度量,构建了蔬菜病害知识视频语义场景模型,通过度量相邻镜头的相似度对语义上相近的镜头进行再聚类,使聚类后的视频场景更适合基于内容的视频检索。实验结果表明,该方法检测蔬菜视频语义场景的查准率达到 96.9%。

关键词: 视频分割; 语义挖掘; 场景检测; 蔬菜病害视频

中图分类号: S625.3 **文献标识码:** A **文章编号:** 1000-1298(2016)S0-0386-06

Semantic Mining Oriented Scene Detection of Vegetable Disease Videos

Wen Haojie^{1,2} Zhou Jing³ Fu Zetian^{4,5} Zhang Lingxian^{3,5} Yan Jin³ Li Xinxing^{3,4}

(1. Large Infrastructure Construction Office, China Agricultural University, Beijing 100083, China

2. College of Engineering, China Agricultural University, Beijing 100083, China

3. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

4. Beijing Laboratory of Food Quality and Safety, Beijing 100083, China 5. Key Laboratory of Agricultural

Information Acquisition Technology, Ministry of Agriculture, China Agricultural University, Beijing 100083, China)

Abstract: In order to improve the current situation of lacking of agricultural video processing technologies, such as agricultural video shot detection, agricultural video annotation, the video scene detection technology, video semantic annotation and disease diagnosis algorithm for agriculture were studied, and the semantic mining oriented semantic scene detection model was constructed. Firstly, a vegetable video shot database was established by using multi-modal fusion video segmentation technology. Secondly, through the study of vegetable knowledge system, a Chinese dictionary of vegetables knowledge was constructed. At last, the multi-modal fusion based semantic annotation model for vegetable scene was presented though analyzing target recognition technology in multiple modals. The recognition results of semantic similarity of three modal were measured by HowNet, and video semantic scene detection model was built. The similarity of adjacent shots was measured, and the similar adjacent shots on semantic would be clustered, so that the clustered video scenes would be more suitable for content-based video retrieval. Experiment results showed that the proposed method reached accuracy rate of 96.9% on vegetable semantic scene detection. This method can solve the ambiguity problem of existing algorithms annotating vegetable videos, and it would help to realize professional and objective semantic annotation of vegetable video scene.

Key words: video segmentation; semantic mining; scene detection; vegetable disease videos

收稿日期: 2016-07-20 修回日期: 2016-08-22

基金项目: 叶类蔬菜产业技术体系北京市创新团队建设专项 (BAIC07-2016)、天津市科技支撑计划研究项目 (15ZCZDNC00120) 和中国博士后科学基金项目 (2014M560139)

作者简介: 温皓杰 (1982—), 女, 工程师, 主要从事农业系统工程与信息化技术研究, E-mail: 76714407@qq.com

通信作者: 李鑫星 (1983—), 男, 副教授, 主要从事农业系统与知识工程研究, E-mail: lxxcau@cau.edu.cn

引言

农民更容易理解视频形式的知识学习模式,而现有的农业知识视频检索粗糙且内容冗长。因此,如何对已有的农业知识视频进行语义挖掘,使其能够针对农民的个性化需求提供专业的农业知识视频检索服务,成为目前急需解决的问题。

长期以来,国内外学者针对视频分析处理问题,包括视频镜头检测方法、镜头聚类模型和语义挖掘算法进行了深入的研究^[1-11]。特别是随着视频处理技术在专业领域内的广泛应用,特定领域的视频分析方法基于特定应用模型,较通用的视频语义挖掘方法具有更高的准确度^[12-16]。目前,视频处理技术在农业视频数据的分析应用方面存在3方面的不足:第一,传统的视频镜头分割方法是根据镜头转换处视频底层特征的变化,从不同的角度抓住视频镜头边缘的特征达到准确分割镜头的目的^[17],忽略了视频中语音信息及文本音频数据对视频分割的重要性,将现有视频分割算法直接应用到农业视频的镜头分割上不能达到镜头语义完整的效果;第二,视频内容的识别方法因其针对性强,在特定领域有不同的模型算法,目前还没有专门针对农业视频内容的识别方法;第三,目前还没有行之有效的视频分析方法可以对农业视频进行完整的语义场景的提取。

本文通过分析蔬菜病害知识视频(以下简称“蔬菜视频”)的内容特点,采用多模态的视频镜头边界检测技术和内容识别技术,提出多模态融合的视频镜头分割与识别方法,以解决现有镜头分割算法对蔬菜视频分割有效性差、语义识别缺乏专业性、容易产生歧义的问题;通过研究蔬菜病害知识体系,构建基于语义相似度的蔬菜视频场景检测模型;通过度量相邻镜头的相似度对语义上相近的镜头进行聚类,使聚类后的视频场景更适合基于内容的蔬菜视频检索。

1 实验材料与方法

1.1 实验视频获取

实验所用视频数据来自中央电视台教育频道提供的蔬菜病害防治节目,包括油菜灰霉病防治、黄瓜霜霉病防治等,共计60个样本视频。视频长度从10 min到60 min不等,视频采用MPEG-2标准压缩,节目伴音采用8 kHz采样频率、8位单声道音频,并伴有字幕。

根据语义理解,对视频样本进行人工检测,得到语义场景1 287个,每一个语义场景都能表达一个完整的语义单元。如黄瓜霜霉病防治节目人工检测

出31个场景,包括5个霜霉病发病条件场景、9个病害感染部位场景、11个症状特征场景和6个防治方法场景。人工场景检测的规则为:将各发病条件、感染部位、发病阶段以及防治方法均作为一个单独的语义场景。这与本研究构建算法所要达到的目的相一致,可以作为算法验证的依据。

1.2 基于多模态的蔬菜视频镜头自动检测

蔬菜视频中语音及字幕是对视频的精确描述,因此在对视频进行镜头边界检测时要同时考虑具有相同字幕信息的镜头相关性和音频镜头与视频镜头的相关性。

采用多模态融合的方法确定蔬菜视频镜头的边界。采用自适应双阈值的镜头边界检测方法^[18]对视频中图像序列进行镜头突变和渐变检测;采用多帧验证技术定位视频中文本行的起始帧和结束帧;采用基于短时平均能量的声音镜头分割算法检测声音镜头的边界;以图像、字幕及声音的时序相关性对分割后的镜头进行复检,构建完整的语义子空间,避免视频镜头分割碎片化。采用时域最大值关键帧法(Temporally maximum occurrence frame, TMOF)^[19]提取镜头序列中出现频率最高的内容作为视频镜头的关键帧。

1.2.1 基于文本的镜头分割复检

蔬菜视频中为了把症状、防治方法等细节表述清楚,同一段文本往往会跨越多个镜头。采用基于文本的镜头聚类方法,依据时间相关性将同一文本对象覆盖的视频镜头聚类为一个新的镜头,并将聚类结果存入向量 $\mathbf{F}_k = (f_k, r_k)$ 。其中, f_k 表示文本聚类后的视频镜头; r_k 是视频镜头标记位,表示视频镜头的类型,含有文本的视频镜头标记为 $r_k = 1$,不包含文本的视频镜头标记为 $r_k = 0$ 。具体步骤如下:

通过自适应双阈值的视频镜头边缘检测算法确定可视镜头边界,用 C_k 表示;采用基于多帧验证技术的字幕快速检测、定位和跟踪方法确定视频中文本起始帧和结束帧,分别用 T_p 和 T_q ($T_p < T_q$)表示,则基于文本的镜头聚类(图1)有3种情况:①当 $T_p \leq C_k$ 且 $C_k \leq T_q$ 时,遍历下一个镜头边界 C_{k+1} ,直到镜头边界 C_{k+n+1} 使 $C_{k+n} < T_q \leq C_{k+n+1}$ 成立,将 C_k 到 C_{k+n+1} 的镜头按时间顺序聚类为新的镜头 f_k ,并标记 $r_k = 1$ 。②当 $T_p < T_q \leq C_k$ 时,同一文本对象没有跨越多个视频镜头,没有聚类发生, $C_k = f_k$,并标记 $r_k = 1$ 。③当 $T_q < C_k$ 且 $C_k < T_{p+1}$ 时,该镜头内没有出现文本且没有聚类发生, $C_k = f_k$,并标记 $r_k = 0$ 。

1.2.2 基于声音的镜头分割复检

通过短时平均能量和特征提取法确定音频镜头,依据声音镜头类型的改变及声音镜头与文本聚

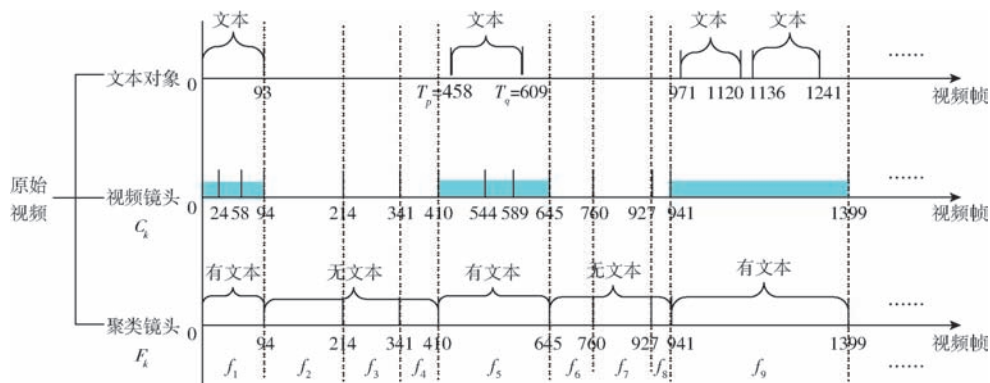


图 1 基于文本的镜头聚类
Fig. 1 Text based shot clustering

类后的视频镜头的分割点是否重合进行基于声音的视频分割复检(图 2),令 f_k 表示 1.2.1 节聚类后的视频镜头, s_k 表示声音镜头, fs_k 表示复检后的视频镜头, r_k 表示镜头是否包含文本对象, t_k 表示镜头中的声音类型(语音类型标记为 1,非语音类型标记为 0), u_k 表示镜头 fs_k 的关键帧,关键帧采用时域最大值关键帧法提取。基于声音的镜头分割复检包括的步骤为:①检测声音镜头 s_j 的类型是否为语音类型,是则继续,否则进入步骤③。②检测声音镜头边界点 s_j 是否有与之重合的视频镜头边界点 f_i ,即当 $f_i = s_j$ 时进入步骤⑤,否则进入步骤⑦。③检测声音镜头 s_j 覆盖的视频镜头 f_i 的类型是否为带字幕型,是则进入步骤⑥,否则继续。④检测两相邻声音镜头类型是否相同,即当 $t_j = t_{j+1}$ 时进入步骤⑦,否则继续。⑤标记镜头边界点 $fs_k = s_j$,得到复检后的视频镜头序列,以向量 $FS_k = (fs_k, t_k, r_k, u_k)$ 表示。⑥标记镜头边界点 $fs_k = f_i$,得到复检后的视频镜头序列,以向量 $FS_k = (fs_k, t_k, r_k, u_k)$ 表示。⑦ $j = j + 1$,转入步骤①检测下一声音镜头。

当 $f_i = V$ 时(V 为原始视频结束帧),检测结束,输出蔬菜病害视频镜头复检结果,以向量 $FS_k = (fs_k, t_k, r_k, u_k)$ 表示。

1.3 多模态融合的蔬菜视频内容自动识别

视频的内容识别是获取视频语义关键词的基本方法,蔬菜视频的内容识别包括对视频帧图像、音频以及对视频中出现的文本进行识别。

采用词袋模型^[20]对 1.2 节提取出的视频镜头关键帧进行目标识别,提取图像视觉特征,构建词袋的统计学模型以获取视频的图像信息,最后对要识别的图像与词袋中的视觉单词进行相似度匹配,最相似的视觉单词组成蔬菜视频关键帧图像的视觉直方图,并将识别结果以关键词的形式存入文本 K_{text} (图 3)。

应用微软的语音合成(TTS)引擎及配套的汉语

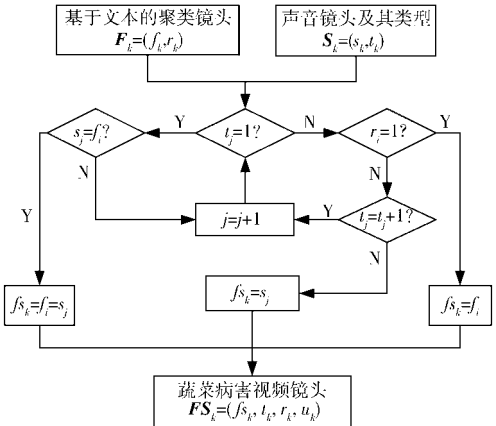


图 2 基于声音的镜头分割复检
Fig. 2 Speech based shot re-inspection

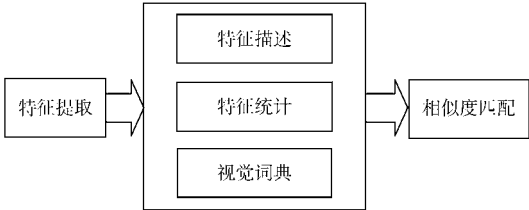


图 3 视频关键帧图像识别流程
Fig. 3 Image recognition flow of key frame

语言模型、汉语声学模型及汉语词表文件,对 1.2.2 节检测到的 $t_k = 1$ (语音类型)的蔬菜病害知识视频镜头进行中文连续语音识别,并将系统输出的识别结果转录为文本,记为 V_{text} 。

蔬菜视频镜头中的语音包含大量专业领域词汇,如:病害名称、防治措施、药品名称等专业术语。这些专业术语在通用语言模型及词表中没有收录或者词频不高,仅用通用词典对视频镜头进行语音识别会出现大量错字,从而导致语音识别率降低。

通过对农业生物学及蔬菜病害知识体系的研究,运用统计学原理,参照大量蔬菜病害知识文献,统计蔬菜病害专有名词及专业术语的词频信息,在农用词典^[21]的基础上设计蔬菜知识中文词典(图 4),用于存放蔬菜病害专用词汇及其词频等信息。该词表包含蔬菜病害专用词汇 895 条,与通用

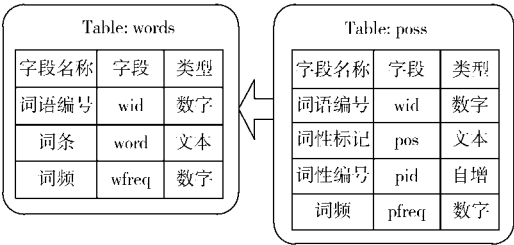


图4 蔬菜病害知识中文词典结构

Fig.4 Data structure of Chinese vegetable disease dictionary

词表相结合,可以有效降低视频语音识别的错误率。

采用基于连通分量的文字识别方法^[22]对1.2.2节检测到的 $r_k=1$ (文本类型)的蔬菜视频镜头进行文字识别(图5):将图像分割成连续的像素区块(即连通分量),根据特征的不同将这些连通分量分为文本连通分量和非文本连通分量,最后将邻近的文本连通分量组合在一起,利用文本的空间连贯性特点形成文字,并将文字输出为文本,记为 W_{text} 。

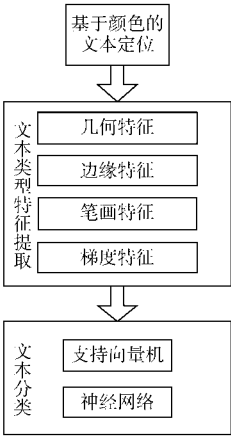


图5 视频文字识别流程

Fig.5 Video text recognition process

1.4 基于语义相似度的蔬菜病害视频语义场景检测

对蔬菜视频镜头的3种模态分别进行识别后,将每个镜头在各模态下的识别结果分别存入文本文件 K_{text} 、 V_{text} 和 W_{text} 中。3种模态的识别结果中,图像目标识别过程决定了 K_{text} 中的文本为词组的集合,且所有词组均为名词词性(如苗、叶子、人),而语音和文本识别输出的结果文本 V_{text} 和 W_{text} 均为连续的汉字串。对 V_{text} 和 W_{text} 进行分析之前首先要分词,将汉字串切分为正确的词组,并对每个词组的词性进行标注。例如:“黄瓜得了白粉病”经过处理后为:黄瓜/名词 得了/谓词 白粉病/名词。这里采用基于隐马尔可夫(HMM)的自动分词及词性标注一体化模型^[23]对文本进行分析(图6)。为了减少歧义字段,将蔬菜中文词典与通用语料库相结合,运用于分词与词性标注模型中,以提高算法的准确率。

语音及文本识别的结果文本经过自动分词及词性标注后,连续的字串被切分为多个标注有词性的

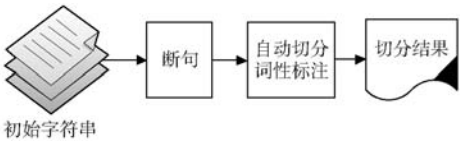


图6 自动分词及词性标注流程

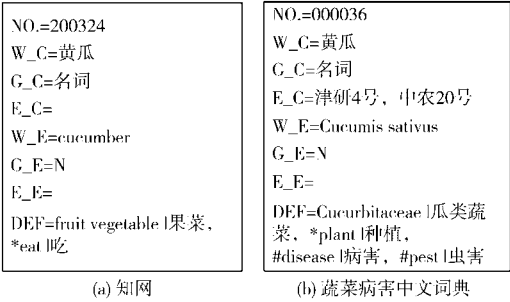
Fig.6 Automatic words segmentation and tagging flow

词组。对蔬菜视频镜头进行语义场景检测,只关注视频中出现的物体、动作及状态。因此,只保留划分结果中的名词、动词和形容词词性的词组,舍弃其他词性的词组,合并具有相同词性的相同词组,避免重复,最后将处理结果存入文本文件 W_{scene} 中。如:初始字符串“在防治黄瓜白粉病时我们可以用海草基因加翠白和鱼蛋白进行叶面喷雾”,经分析处理后得到“防治/动词 黄瓜/名词 白粉病/名词 用/动词 海草基因/名词 翠白/名词 加/动词 鱼蛋白/名词 叶面/名词 喷雾/动词”。这段文本中包含了大量农业专业术语,如果只采用通用词典会产生大量歧义字段导致分词正确率下降,如“白粉/名词 病/名词”。而引入蔬菜中文词典后可以有效改善这种情况。另外,语义镜头文本文件 W_{scene} 中还应包括该场景所在初始视频的标题,并与其他文字串一样做分词和词性标注处理。如标题为“大棚黄瓜主要病害的发生与防治”的语义场景检测文件中均应包括词组“大棚/名词 黄瓜/名词 主要/形容词 病害/名词 发生/名词 防治/名词”,它是界定该场景应用范围的主要依据。

获得语义镜头文本文件 W_{scene} 后,需要测量两相邻镜头的文本文件 $W_{scene,i}$ 和 $W_{scene,i+1}$ 的语义相似度,来判断相邻两镜头是否属于同一语义场景。采用基于知网的语义相似度测量方法^[24]对两文本文件 $W_{scene,i}$ 和 $W_{scene,i+1}$ 中的词组进行相似性度量。由于语义相似度度量的应用有很强的领域性和专业性,因此引入蔬菜中文词典增加农业专有名词的概念描述是十分必要的。如图7所示,“黄瓜”概念在知网中的描述为一种果菜,是用来吃的(图7a);在蔬菜词典中“黄瓜”概念被描述为一种瓜类蔬菜,用来种植,还与病害有关系(图7b)。

将词组语义相似度标准化到 $[0,1]$ 区间,认为当两相邻镜头的语义文本中有词组的语义相似度大于某一阈值 T_{high} ,且该类词组的数量在语义文本中的比例超过阈值 N 时,两相邻镜头处于同一语义环境,将被聚类为一个语义场景,合并语义镜头文本文件 W_{scene} ,并删除文件中的重复词组;否则该语义场景结束,检测接下来相邻的两个镜头。

令 $V_k=(q_1,q_2,\cdots,q_t)$ 表示视频镜头的语义词组集合,其中 V_k 为第 k 个镜头, q 为镜头语义文本



参 考 文 献

- 1 BHAUMIK H, BHATTACHARYYA S, DAS NATH M, et al. Hybrid soft computing approaches to content based video retrieval: a brief review [J]. *Applied Soft Computing*, 2016, 46:1008 – 1029.
- 2 GORI M, LIPPI M, MAGGINI M, et al. Semantic video labeling by developmental visual agents[J]. *Computer Vision and Image Understanding*, 2016, 146:9 – 26.
- 3 RAVI D, BOBER M, FARINELLA G M, et al. Semantic segmentation of images exploiting DCT based features and random forest [J]. *Pattern Recognition*, 2016, 52:260 – 273.
- 4 LUISGARCIA J Y, PEREZAGUIL A R. Automatic representation and segmentation of video sequences via a novel framework based on the nD -EVM and kohonen networks [J]. *Advances in Artificial Intelligence*, 2016, 2016:1 – 34.
- 5 BOULMERKA A, SAID ALLILI M, AIT-AOUDIA S. A generalized multiclass histogram thresholding approach based on mixture modeling [J]. *Pattern Recognition*, 2014, 47(3): 1330 – 1348.
- 6 HU H, UPCROFT B. Automatic object segmentation of unstructured scenes using colour and depth maps[J]. *IET Computer Vision*, 2014, 8(1): 45 – 53.
- 7 LI Y B, MERIALDO B. Multimedia maximal marginal relevance for multi-video summarization [J]. *Multimedia Tools and Applications*, 2016, 75(1):199 – 220.
- 8 NTALIANIS K S, DOULAMIS A D, DOULAMIS N D, et al. Unsupervised segmentation of stereoscopic video objects: constrained segmentation fusion versus greedy active contours[J]. *Journal of Signal Processing Systems for Signal Image and Video Technology*, 2015, 81(2):153 – 181.
- 9 LIU Yizhi, YANG Ying, XIE Hongtao, et al. Fusing audio vocabulary with visual features for pornographic video detection[J]. *Future Generation Computer Systems—The International Journal of Grid Computing And Escience*, 2014, 31: 69 – 76.
- 10 FERNANDEZ-MARTINEZ F, HERNANDEZ GARCIA A, DIAZ DE MARIA F. Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials[J]. *Expert Systems with Applications*, 2015, 42:293 – 305.
- 11 ZHANG B, LIU J, TANG X, et al. Multi-scale video text detection based on corner and stroke width verification[C] //2013 IEEE International Conference on Visual Communications and Image Processing, 2013:1 – 6.
- 12 KHAN A, WINDRIDGE D, KITTLER J. Multilevel Chinese takeaway process and label-based processes for rule induction in the context of automated sports video annotation[C] //IEEE Transactions on Cybernetics, 2014, 44(10):1910 – 1923.
- 13 LEE K-S, ROSLI A N, SUPANDI I A, et al. Dynamic sampling-based interpolation algorithm for representation of clickable moving object in collaborative video annotation[J]. *Neurocomputing*, 2014, 146:291 – 300.
- 14 DUONG Trong Hai, NGUYEN Ngoc Thanh, TRUONG Hai Bang, et al. A collaborative algorithm for semantic video annotation using a consensus-based social network analysis[J]. *Expert Systems with Applications*, 2015, 42(1):246 – 258.
- 15 YAN F, KITTLER J, WINDRIDGE D, et al. Automatic annotation of tennis games: an integration of audio, vision, and learning [J]. *Image and Vision Computing*, 2014, 32(11):896 – 903.
- 16 GONZALEZ-DIAZ I, MARTINEZ-CORTES T, GALLARDO-ANTOLIN A, et al. Temporal segmentation and keyframe selection methods for user-generated video search-based annotation[J]. *Expert Systems with Applications*, 2015, 42(1): 488 – 502.
- 17 马永波. 基于内容的视频分割与检索技术研究[D]. 长春:长春工业大学, 2007.
- 18 MA Yongbo. Video segmentation and content-based video retrieval[D]. Changchun:Changchun University of Technology, 2007. (in Chinese)
- 19 傅泽田, 苏叶, 张领先, 等. 基于自适应双阈值的蔬菜病害知识视频分割方法[J]. *农业工程学报*, 2013, 29(9):148 – 155.
- 20 FU Zetian, SU Ye, ZHANG Lingxian, et al. Method of video segmentation for vegetable disease based on self-adaptive dual thresholds[J]. *Transactions of the CSAE*, 2013, 29(9): 148 – 155. (in Chinese)
- 21 李玉峰. 基于动态帧的视频关键帧提取算法研究[J]. *天津科技大学学报*, 2009, 24(4):69 – 72.
- 22 LI Yufeng. Research on video key frame extraction algorithm based on dynamic frame[J]. *Journal of Tianjin University of Science and Technology*, 2009, 24(4):69 – 72. (in Chinese)
- 23 钱钧, 杨恒, 刘培桢, 等. 一种基于词袋模型的大规模图像层次化分组算法[J]. *应用光学*, 2014, 35(5):799 – 805.
- 24 QIAN Jun, YANG Heng, LIU Peizhen, et al. Large-scale image hierarchical grouping algorithm based on bag-of-words[J]. *Journal of Applied Optics*, 2014, 35(5):799 – 805. (in Chinese)
- 25 李瑶. 面向呼叫中心的农业知识文语转换系统研究[D]. 北京:中国农业大学, 2012.
- 26 LI Yao. Study on the text-to-speech system for agricultural knowledge oriented on call-center[D]. Beijing: China Agricultural University, 2012. (in Chinese)
- 27 YI C, TIAN Y. Text extraction from scene images by character appearance and structure modeling[J]. *Computer Vision and Image Understanding*, 2013, 117(2): 182 – 194.
- 28 代建英. 汉语自动分词系统的研究和实现[D]. 重庆:重庆大学, 2005.
- 29 DAI Jianying. Study on the Chinese automatic word segmentation system[D]. Chongqing: Chongqing University, 2005. (in Chinese)
- 30 BENBIN W, JING Y, LIANG H. Chinese how net-based multi-factor word similarity algorithm integrated of result modification [C] //Neural Information Processing 19th International Conference, ICONIP 2012 Doha Proceedings, 2012, 7667: 256 – 266.