一个知识辅助的视频语义概念探测框架*

白 亮[1],老松杨[1],侯小强[2],刘海涛[1],卜 江[1]
(1.国防科技大学 信息系统工程重点实验室,湖南 长沙 410073;
2.61046部队,北京 100094)

摘 要:视频语义概念探测是视频语义内容分析研究领域的热点和难点问题。语义概念探测方法的性能取决于其是否能够有效地建模和匹配视频语义内容特征。将视频内容抽象为感知概念和语义概念以及概念间的关系,提出了知识辅助的视频语义概念探测框架,利用本体建模概念间关系和上下文知识,从低层特征匹配和上下文匹配两个方面综合考虑语义概念的探测。通过线性融和策略,对匹配结果进行融合得到最终的探测结果。实验结果表明提出的方法探测性能良好。

关键词:本体;视频语义概念探测;上下文信息;低层特征
中图分类号:O23 **文献标志码**:A **文章编号**:1011-2486(2012)04-0090-05

A knowledge-assisted framework for video semantic concept detection

BAI Liang[1], *LAO Songyang*[1], *HOU Xiaoqiang*[2], *LIU Haitao*[1], *BU Jiang*[1]
(1. Science and Technology on Information System Engineering, National University of Defense Technology, Changsha 410073, China;
2. PLA Unit 61046 Beijing 100094, China)

Abstract: Semantic concept detection in video is a challenge for video semantic content analysis. The performance of semantic concept detection methods depends on modeling and matching the video semantic content exactly. In this research, perception concept and semantic concept were defined to abstract and model video semantic content. Furthermore, the knowledge-assisted framework for semantic concept detection was proposed, in which the context knowledge was modeled using ontology, and the semantic concepts were detected by combining with low-level features and context information. Finally, the linear fusion strategy was used to fuse the matching results and detect the semantic concepts. The proposed method was demonstrated in a news video domain and shows promising results.

Key words: knowledge-assisted; ontology; video semantic concept detection; context

视频语义内容分析的目标是抽取视频包含的高层语义内容,为用户提供语义概念的视频浏览、检索服务,语义概念探测是实现这一目标的核心步骤,并成为近期视频语义内容分析领域的重要研究方向[1]。

以往的视频概念探测主要采用基于内容的方法,即通过抽取概念具有的低层特征,学习某种关联模型(基于规则的或是基于统计机器学习的),直接地、独立地建立低层特征与概念之间的关联,探测视频概念。基于规则的方法是在抽取特征的基础上,对特征进行简单或者复杂的阈值判定[2]。这种关联模型的缺点是阈值确定难、算法不鲁棒,并且简单的阈值判断难以有效地表征概念具有的特征多样性。因此,目前采用较多的是基于统计机器学习的关联模型[3-4],即通过某个机器学习模型学习标注的样本数据中低层特征与视频概念之间的统计概率关联模式,然后采用训练好的机器学习模型对新的样本进行识别,探测视频概念。目前的研究表明,支持向量机[5]和最大熵模型(Maximum Entropy Model,简称MEM)[6]是两类较为有效的用于概念探测的机器学习模型。

但是,由于语义鸿沟的存在,低层特征和高层语义的关联并不是一一对应的。不同的视频概念可能具有相似的低层特征,相同的视频概念也可能具有完全不同的低层特征,基于内容的独立概念,探测方法难以克服这个问题。另一方面,视频中的概念并不是独立出现的,不同的概念总是同时出现在视频帧序列中。显然,不同概念的共现性将增加低层特征模式的复杂性,进而影响独立的概念探测性能。但是,从另外一个角度思考,不同概念间的关系信息也为概念探测提供了重要的上下文知识,重要的是如何有效地建模和利用这些知识。

针对语义概念探测存在的困难,本文提出了知识辅助的视频语义概念探测方法。一方面通过

* **收稿日期**:2011-12-20
基金项目:国家自然科学青年基金项目(60902094);"十二五"国家部委项目
作者简介:白亮(1978—),男,陕西清涧人,讲师,博士,E-mail:Bailiang@nudt.edu.cn

定义中层语义以减小语义鸿沟,建立低层特征与高层语义关联的桥梁;另一方面利用概念间的关系和上下文语境,在概念探测中加入语义线索,提高概念探测器的语义识别能力。而本体作为合适的知识建模工具可以有效地描述视频语义内容和建模领域知识,因此利用本体增强概念探测的语义表达和识别能力是必需的也是可行的。

1　视频语义概念探测框架

视频内容跨越了低层感知特征、感知特征模式、简单语义概念、复杂语义概念诸多层次,并不是简单的特征层和语义层就能表示的;更为重要的是,这种层次结构建立了视频内容从低层特征到高层语义的内在关联过程,为跨越语义鸿沟提供了有效途径。另一方面,视频语义内容分析的本质就是各个层次内容的分析抽取和各个层次之间关联的建立。

文献[7]提出并构建了视频领域知识本体和视频概念扩展本体分别对上下文信息和视频低层特征与高层概念的关联关系进行描述。在此基础上,本文认为语义概念的探测应该从两个方面进行考虑。一方面是发现概念具有的低层特征模式,即抽取视频概念关联的感知概念,从感知概念中抽取低层特征训练统计机器学习模型,识别语义概念;另一方面是充分利用上下文信息增强概念探测方法的语义理解和识别能力,即通过视频概念扩展本体的语言层定义的概念之间的关系和概念描述属性对上下文信息进行描述,利用 VOCR 和语音识别技术从视频片段中抽取文本信息,在上下文中建立语义概念与文本信息之间的内在关联,增强概念探测的准确率。

基于上述分析,本文提出知识辅助的视频语义概念探测框架,如图 1 所示。该框架主要分为三个部分:上下文信息匹配、特征匹配和匹配结果融合。在上下文信息匹配中,一方面,通过 VOCR 和语音识别技术从视频片段提取文本信息,另一方面通过视频概念扩展本体的定义获取待探测概念的描述和与其关联的概念,则二者的相关程度暗示了待探测概念出现的可能,上下文信息匹配将定量计算这种相关程度。特征匹配根据低层感知特征相似性计算视频片段中包含视频概念的可能性,首先通过视频概念扩展本体的定义获取待探测概念包含的感知概念,然后抽取视觉对象特征训练概念分类器,计算视觉对象匹配程度,同时统计视频片段包含其他感知概念的情况,计算其他感知概念匹配结果。最后,通过一种融和策略,对匹配结果进行融合,融合结果表示概念探测结果。

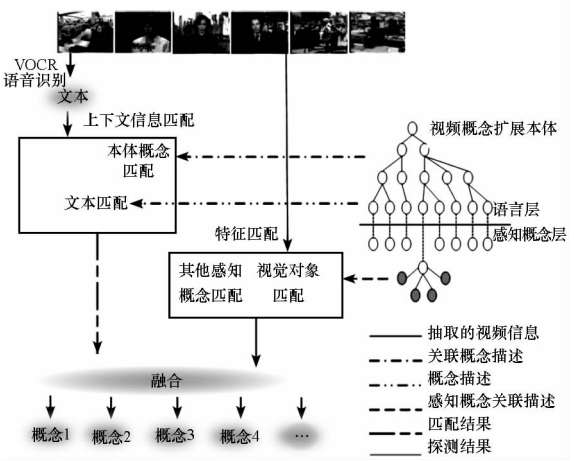

图 1　知识辅助的语义概念探测

Fig. 1　Framework for semantic concept detection using ontology

2　上下文信息匹配

视频语义概念具有的上下文信息包括两个方面,一是概念本身的描述所蕴含的上下文信息,包括概念的内在性质、外在属性的描述和概念具有的同义词集。另一方面的上下文信息表现为概念间的关系。提出的视频概念扩展本体定义两类概念间关系,一类是语义关系,主要包括 Kind-of, Instance-of 和 Part-of 三类父子层次关系;另一类是概念共现关系,定义为不同的概念在视频中同时出现的概率大小。不难理解,具有强共现关系的概念同时出现的概率大,因此一个概念的出现可以作为另一个概念探测的线索;语义关系表征的是概念间的语义相关性,而从自然语言的特点来看,语义相关的概念的出现具有集中性,因此概念语义关系为概念探测提供有用的上下文信息。

根据上下文信息的两个方面,本文分别采用文本匹配和本体匹配两种方法来利用上下文信息进行语义概念探测。

文本匹配通过计算概念描述文本与视频片段包含的文本之间的相似性来判断视频片段包含该概念的可能性大小。

视频中包含的文本信息一方面来自于视频伴随音轨中的语音信息,另一方面来自于视频中字幕、场景文字的识别,即 VOCR。本文通过语音识别和 VOCR 抽取的文本信息记作 vt,进而抽取 vt 中包含的词条集 $nt = \{nt_i\}_{i=1}^{N}$。对于英文分词采用人工标注方法处理。

概念描述文本通过视频概念扩展本体中的定义获得,包括两个部分:一是概念描述 $d, d \in D, D$ 为本体中所有概念描述集合。另一个是概念同义

词集 SynonymsList。对概念描述 d 进行分词,从中抽取词条集,与同义词集合并组成概念描述词条集 $cd = \{cd_j\}_{j=1}^{M}$。

常用且效果较好的文本表示模型是向量空间模型,根据文本向量空间模型的一般定义[8],本文提出计算视频概念 C 描述文本与视频包含的文本之间的匹配程度方法如下:

$$M\_Text(C) = \sum_{i=1}^{N} \frac{tf_{nt_i,vt} \cdot idf_{nt_i}}{norm_{nt}} \cdot \frac{tf_{nt_i,cd} \cdot idf_{nt_i}}{norm_{cd}} \cdot co_{vt,cd} \tag{1}$$

式中,$tf_{t,X}$表示词频,定义为词条 t 在文本集 X 中出现的频繁程度;$norm_{nt}$,$norm_{cd}$是两个归一化因子;idf_{nt_i}是名词术语 nt_i 在元概念描述集中的倒文档频率;$co_{vt,cd}$表示视频包含的文本 vt 与概念描述文本 cd 的相关度。

通过上述计算,可以度量每一个视频概念与待探测视频片段的文本匹配程度。某个概念计算得到的匹配程度越大,说明该视频片段包含该概念的可能性越大。

与文本匹配相同,抽取视频包含的文本集 vt 和 vt 中包含的名词术语集 $nt = \{nt_i\}_{i=1}^{N}$。通过匹配 nt 与视频概念扩展本体中的概念定义,可以获得 nt 对应的一个概念集合 $C = \{C_i\}_{i=1}^{K}$。直观地讲,C 中包含的概念在该段视频中出现的可能性较大。文献[9]提出了一种概念信息内容度量方法,即度量概念与文本内容的相关程度,具有高信息内容的概念具有高的相关程度。本文采用该方法度量概念在视频文本 vt 中的重要程度(即二者的相关程度)。首先,对于概念 c_i,通过视频概念扩展本体定义的关系,抽取与 c_i 相关的本体概念,这里定义"相关"概念为:在本体中与 c_i 语义距离不超过 2 的概念和共现关系集中定义的与 c_i 具有共现关系的概念。语义距离定义为本体关系图中,两个概念间的最短路径包含的边数。不难理解,对于视频概念而言,其相关概念为其父节点概念、二级父节点概念和所有兄弟节点概念。标记与 c_i 相关的所有概念的同义词集的合集为 $RT(c_i)$,则可定义 c_i 在文本集 vt 中的似然度为:

$$p(c_i)_{vt} = \frac{\sum_{t \in RT(c_i)} Count_{vt}(t)}{N} \tag{2}$$

其中,N 为 vt 中所有名词术语出现的次数;$Count_{vt}(t)$为 $RT(c_i)$中的名词术语 t 在 vt 中出现的次数。概念 c_i 的信息内容计算公式为:$I_{c_i} = -\ln p(c_i)$

根据上述定义,如果一个视频概念相对于某个文本集具有较大的信息内容值,则表明该概念对于该文本集描述的内容具有较高的重要度。进而,可以推断该概念在该文本集对应的视频片段中出现的可能性较大。因此,概念 C 的本体匹配利用其信息内容度量,即:

$$M\_Ontolog(C) = I_C \tag{3}$$

3 特征匹配

特征匹配是从低层特征相似性的角度探测元概念,即建立视频低层感知特征与视频语义概念之间的关联。根据视频概念扩展本体定义的语义概念与感知概念之间的包含关系,抽取与概念相关的视觉对象概念。选择标注过视觉对象概念的视频数据作为训练数据集,抽取相同视觉对象概念的颜色、纹理、位置特征,训练视觉对象概念对应的视频概念分类器,选择 SVM 构造分类器。这里需要指出的是,一个视频概念可能包含若干个视觉对象概念,则每一个视觉对象概念都对应一个概念分类器,不同的视觉对象刻画了概念的不同属性特征,通过对多个视觉对象概念对应的概念分类器的探测结果进行融合,得到最终的概念探测结果。

本文分别抽取视觉对象的颜色、纹理和位置特征如下:

(1) 7 维的 HSV 颜色均值和主颜色(dominant color);

(2) 8 维一个尺度,0°,45°,90°,135°四个方向的 Gabor 纹理特征;

(3) 构造视觉对象外接矩形,抽取 2 维的对象矩形中心点位置特征,8 维的对象矩形顶点位置特征。

抽取上述视觉对象特征组成特征向量,训练 SVM 分类器探测概念。概念探测目标是给出视频片段中是否出现特定概念的判断,同时还要给出这种判断的置信度,即后验概率。关于 SVM 后验概率输出的代表性研究是由 Platt 提出来的[10],其主要思想来源于 Wuhba 关于 RKHS 表示定理的研究。本文采用 Platt 提出的方法,获得 SVM 概率化输出。对给定概念对应的不同视觉对象概念,抽取特征训练分类器,每一个分类器的输出表示根据该视觉对象判断给定的概念出现的概率。

对于待探测视频片段,首先抽取其包含的各个视觉对象概念的低层特征。根据视频概念扩展本体定义的语义概念与感知概念间的关联关系,获取每一个视觉对象概念关联的概念。然后选择相应的 SVM 分类器进行概念探测。

容易理解,对于每一个视频概念,根据其相关

的视觉对象概念的不同,可以计算得到若干个该概念出现的概率值,我们通过计算所有概率值的加权和来最终确定视觉特征匹配的程度值。假设概念 C 具有相关视觉对象概念 $\{VO_i\}_{i=1}^{n}$,通过 SVM 分类器获得 C 的探测结果为 $\{p_i\}_{i=1}^{n}$,其中 p_i 为 VO_i 对应的分类器得到的分类结果,则概念 C 的视觉特征匹配结果计算如下:

$$M\_Feature(C) = \sum_i w_i p_i \qquad (4)$$

其中,w_i 为 VO_i 对应的权值。权值反映视觉对象概念对于概念探测的重要程度。直观理解,如果某个概念的出现总是包含着某个视觉对象概念,则该视觉对象一定对这个概念的探测具有重要作用,所以最合适的重要性度量应该能够有效反映视觉对象概念出现与其相关的概念的共现程度。本文采用互信息方法来度量这种共现程度,以确定不同视觉对象概念具有的权值。根据视觉对象概念 VO_i 与概念 C 的共现互信息度量计算权值 w_i 如下:

$$w_i = \ln\left(\frac{p(VO_i \mid C)}{p(VO_i)}\right)$$

其中,$p(VO_i \mid C)$ 表示 VO_i 在概念 C 的训练视频集出现的概率。根据训练数据集的统计,可计算出每个视觉对象概念相对于其相关概念的权值,并进行归一化。进而,可以计算得到视觉特征匹配的结果。

4　匹配结果融合

在分别得到上下文信息匹配和特征匹配的结果之后,我们采用线性融合方法对匹配结果进行融合,得到最终的概念探测结果。有关融合参数的取值问题,通过实验测试多种取值组合,找到最优的参数设置。

通过上述融合计算之后,可以得到给定概念与测试视频子镜头的匹配程度,匹配程度值越大,说明该概念与测试视频相关程度越高,进而推断其出现在视频中的可能性越大。显然,通过匹配计算可以得到一个匹配值列表,越靠前的匹配概念在测试视频子镜头中出现的概率越大。因此,可以根据实际需要,综合考虑探测性能要求,选择前若干个匹配概念作为探测结果。

5　实验结果与分析

为了评估测试本文提出的知识辅助的概念探测方法,我们采集多种来源的电视节目视频,包括 CNN、NBC、CCTV1、CCTV4 和凤凰卫视的新闻节目,共 10h42min,包含 11 215 个子镜头。选择 6258 个子镜头作为训练集,其余的子镜头作为测试集。抽取每个子镜头包含的字幕文本、语音文本和感知概念具有的感知特征向量,并构建了面向视频语义内容分析的“美国外交政策专题”视频概念扩展本体,其中定义了 41 个语义概念。以这 41 个概念作为标注概念集,对实验数据中每个视频子镜头标注其包含的语义概念。

根据本文提出的概念探测方法,计算给定的概念与所有测试子镜头的匹配程度值。从理论上讲,匹配值大于 0 的子镜头都可以看作是给定概念的相关子镜头,即有包含给定概念的可能。显然,选取较小的匹配值作为判断依据,必然会降低探测的准确度;反之,选择较大的匹配值作为判断依据,则会减少可能的相关镜头数。上述情况都会影响对探测方法评估的准确性。因此,需要从探测准确率和探测召回率两方面综合评估。

定义概念 C 的探测准确率(Precision)和召回率(Recall)如下:

$$\text{Precision} = \frac{\{\text{探测到的子镜头}\} \cap \{\text{包含 } C \text{ 的子镜头}\}}{\{\text{探测到的子镜头}\}} \qquad (5)$$

$$\text{Recall} = \frac{\{\text{探测到的子镜头}\} \cap \{\text{包含 } C \text{ 的子镜头}\}}{\{\text{包含 } C \text{ 的子镜头}\}} \qquad (6)$$

根据 Precision 和 Recall,本文采用平均准确率(Average Precision,AP)来评估概念探测方法。平均准确率是对不同 Recall 点上 Precision 的平均,该指标能够综合考虑探测方法的 Precision 和 Recall 表现,更能够反映探测方法的性能。AP 是视频概念探测中常用的评价指标之一。概念 C 探测的 AP 定义如下:

$$AP = \frac{\sum_{r=1}^{N} P(r)}{\text{包含 } C \text{ 的子镜头数}} \qquad (7)$$

其中,N 是探测到的包含 C 的子镜头数,表示探测到 r 个包含 C 的子镜头时的 Precision。训练阶段中,以 0.1 为步长分别测试不同的线性融合参数,经测试,α_1、α_2、α_3 分别取 0.2、0.2、0.6 时获得最好的训练结果。根据上述 AP 的定义,在测试视频集中探测并计算每一个元概念的 AP 值。结果如图 2 所示。

由图 2 可以看到,近 1/2 的概念探测 AP 达到了 0.5 以上,近 2/3 的概念探测 AP 达到了 0.4 以上,所有概念的探测 AP 平均值(Mean Average Precision,MAP)为 0.49。

文献[11]关于概念探测方法的研究具有代表性,提出了一种关联多标签(Correlative Multi-label,

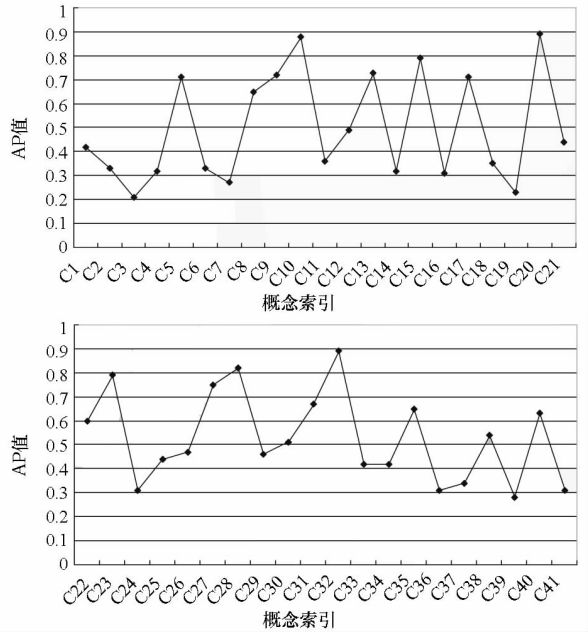

图2 概念探测 AP 结果

Fig. 2 The AP values of the concepts detection

简称 CML)视频概念探测方法,这种方法在训练学习阶段根据概念间的关联关系同时对所有的概念进行分类模型学习和优化,同时有效地利用样本数据同时建模概念分类模式和概念间关系模式,并选取目前应用最为广泛的标准视频概念本体 LSCOM-Lite[12] 中定义的39个概念进行探测实验,其中基于SVM 的 ICA 方法(SVM-ICA)、CBCF 方法和 CML 方法分别得到的 MAP 值为 0.24、0.23、0.29,低于本文提出的概念探测方法取得的平均结果。

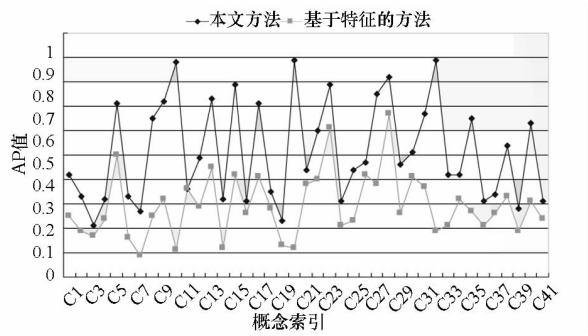

图3 上下文信息匹配对概念探测的影响

Fig. 3 The analysis of using the context information

为了进一步分析验证本文提出的知识辅助的概念探测框架中本体建模的领域知识与上下文信息对探测性能的影响,我们只采用特征匹配方法(Only-feature)对相应的概念进行探测实验,并与本文提出的方法比较,实验结果如图3所示。

图3显示,融合了上下文信息的概念探测对比特征匹配方法大幅提高了探测性能,特征匹配方法的 MAP 值仅为 0.29。这说明,仅仅依赖于低层感知特征来探测视频概念是十分困难的,其根本原因还在于语义鸿沟的存在。通过本体建模视频概念的上下文信息,有效地增加了探测语义识别能力,减小了低层感知特征与视频概念关联的不确定性,提高了探测性能。

6 总结与展望

区别于以往基于内容的视频概念探测方法直接地、独立地建立低层特征与概念之间的关联,本文提出了一个知识辅助的视频语义概念探测框架。从低层特征和上下文语义信息两个方面综合考虑语义概念的探测问题。以感知概念作为低层感知特征和语义概念之间的中间语义,避免了直接建立低层特征和语义概念间的关联,减小了语义鸿沟问题带来的影响。同时,利用本体建模的概念间关系和上下文信息,增强概念探测的语义理解和识别能力。实验结果验证了本文提出方法的有效性。

未来的研究工作主要包括两个方面,一是如何进一步发现和抽取更具表征能力的低层特征;二是如何有效利用本体的自动推理,增强上下文信息的描述和建模能力。同时,复杂概念探测和跨领域概念探测依然是具有挑战性的问题。

参考文献(References)

[1] Chang S F, Ma W Y, Smeulders A. Recent advances and challenges of semantic image/video search[C]// Proceeding of IEEE ICASSP,2007: 1205 - 1208.

[2] Rui Y, Gupta A, Acero A, Automatically extracting highlights for TV baseball programs[C]// Proc. ACM Multimedia, Los Angeles, CA, 2000: 105 - 115.

[3] Snoek C G M, Worring M, Geusebroek J, et al. The semantic pathfinder: using an authoring metaphor for generic multimedia indexing[C]// IEEE Trans. Pattern Anal. Machine Intell., 2006,28:1678 - 1689.

[4] Chang S F, et al. Columbia university trecvid-2006 video search and high-level feature extraction[C]// TREC Video Retrieval Evaluation (TRECVID) Proceedings, 2006.

[5] Vapnik V. The nature of statistical learning theory[M]. New York:Springer-Verlag, 1995.

[6] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification[C]// IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999: 61 - 67.

[7] 白亮. 本体支持的视频情报分析方法与技术研究[D]. 长沙:国防科学技术大学, 2008.
BAI Liang. Research on public video intelligence analysis using ontology [D]. Changsha: National University of Defense Technology,2008. (in Chinese)

[8] Salton G. Development in automatic text retrieval [J]. Science, 1991, 253(5023): 974 - 979.

[9] Resnik P, Using information content to evaluate semantic similarity in a taxonomyin[C]// Int. Joint Conf. Artificial Intelligence, Montréal, QC, Canada, 1995: 448 - 453.

[10] Platt J. Probabilities for SV machines[C]// Advances in Large Margin Classifiers, Cambridge, MA: MIT Press, 2000: 61 - 74.

[11] Qi G J, et al. Correlative multi-label video annotation[C]// Proc. of ACM MM'07, Germany, 2007: 17 - 26.

[12] Naphade M, Smith J, Tesic J, et al. Large-scale concept ontology for multimedia[J]. IEEE Multimedia, 2006, 13(3): 86 - 91.