第37卷 第15期
Vol.37 No.15

计 算 机 工 程
Computer Engineering

2011年8月
August 2011

·多媒体技术及应用·

文章编号：1000—3428(2011)15—0212—03 文献标识码：A 中图分类号：TP391.4

基于Simfusion和本体的视频语义提取

张建明，李 梅，李广翠

(江苏大学计算机科学与通信工程学院，江苏 镇江 212013)

摘 要：目前大多数的视频语义概念提取研究没有考虑到视频多模态之间的关联共生特性，而在样本的标注方面采用自定义的概念进行标注，会影响语义概念提取的准确率。针对上述问题，提出结合Simfusion算法和用本体知识库标注样本的方法提取视频的语义概念，该方法根据镜头内容变化提取关键帧，在提取出镜头内容时，有效地利用镜头多模态之间的时序关联共生特性，同时运用本体知识库中的概念标注样本、训练分类器，弥补传统方法在标注样本时存在的主观、不规范等不足。实验结果表明，该方法在视频语义概念提取的研究中，有较高的准确度、可操作性强。

关键词：Simfusion算法；本体知识库；时序关联共生特性；多模态；视频语义概念

Video Semantic Extraction Based on Simfusion and Ontology

ZHANG Jian-ming, LI Mei, LI Guang-cui

(College of Computer Science and Communications Engineering, Jiangsu University, Zhenjiang 212013, China)

【Abstract】Most of the researches in extracting semantic concepts do not consider the temporal associated co-occurrence characteristic of multimodes, and label the training set using self-define concepts, thus affecting the accuracy of semantic concepts extraction. Aiming at these problems, this paper brings forward a new approach based on Simfusion algorithm and labeling the training set using ontology repository to extract semantic concepts of video. The method extracts key-frame according to the content of the shots, and makes the most of temporal associated co-occurrence characteristic during in multimode. Meanwhile, the method labels the sample set using the ontology repository and training classifier, thus offsetting insufficiency in subjectivity, incorrect. Experimental result shows that the method can get a better accuracy, well operability and universality in the research of semantic concepts of video.

【Key words】Simfusion algorithm; ontology repository; temporal associated co-occurrence characteristic; multimode; video semantic concepts

DOI: 10.3969/j.issn.1000-3428.2011.15.068

1 概述

视频语义概念的提取是目前语义视频研究中的一个热点，为了挖掘视频的语义信息，很多研究者结合机器学习的方法，利用视频的多模态特性提取视频的语义概念，并取得一定的成果。如文献[1]根据新闻视频特点，从台标识别、栏目识别、字幕识别、播音识别等方面对视频语义的提取进行了探索；一些基于语义的标注系统也应运而生，以满足人们对视频进行基于语义概念的检索和管理，如文献[2]利用语音自动识别技术和特定的语义模型实现了对新闻视频的主播场景和镜头的语义标注。

在视频语义研究中，前期融合和后期融合是目前常用的2种主要的融合方式[3]。Snoek C G M等人利用这2种融合方式提取视频语义概念并取得一定成果。但前期融合很难得到一个统一的表达方式；后期融合由于针对每个模态都要进行训练，所以带来了学习训练上的复杂度。同时，前期和后期融合都没有考虑到多模态之间的关联共生特性。

浙江大学的刘亚楠等人利用Simfusion[4]来计算镜头之间的相似度[5]。该算法可以在计算镜头之间相似度时有效地融合异构数据源之间的关系即多模态之间的关联共生关系，使得相似度的计算更加准确和合理，但也存在一些问题。如该方法针对每个镜头只提取一幅关键帧，这样，对内容变化比较大的镜头，一幅关键帧不能很好地表达镜头的语义，从而影响镜头之间相似度的计算，对语义概念的提取等研究存在影响。同时，该方法在训练集的标注时，采用的是自定义的一些概念，这样就存在主观性强、不规范、不具有一般性等不足。

针对上述问题，本文提出一种基于Simfuion和本体相结合的方法，该方法从多模态之间的关联共生特性和样本的标注两方面入手，借鉴文献[6-7]的思想，根据镜头内容变化提取关键帧来反映镜头表达的语义内容。

2 Simfusion算法

该算法首先建立一个表示异构数据对象之间相关关系的统一关系矩阵(Unified Relation Matrix, URM)，URM提供了一个统一的视角来看待数据对象以及它们之间的关系。在URM中，不同类型的数据对象被看成是位于一个统一的数据空间中的元素。

另外定义一个统一相似度矩阵(Unified Similarity Matrix, USM)来表示数据对象之间的相似度，通过在URM上的迭代计算镜头之间的相似度。由于该算法在计算相似度的同时能够有效地利用多模态之间的关联共生特性，提高相似度计算的准确率。

2.1 URM的定义

统一关系矩阵(URM)定义如下：

基金项目：国家自然科学基金资助项目(60673190)；江苏省自然科学基金资助项目(BK2009199)

作者简介：张建明(1964－)，男，教授、博士，主研方向：语音视频，虚拟现实；李 梅、李广翠，硕士研究生

收稿日期：2011-01-12 **E-mail：**zhjm@ujs.edu.cn

$$\begin{vmatrix} \lambda_{11}\boldsymbol{L}_{\text{image}} & \lambda_{12}\boldsymbol{L}_{i\text{-}a} & \lambda_{13}\boldsymbol{L}_{i\text{-}t} & \lambda_{14}\boldsymbol{L}_{i\text{-}s} \\ \lambda_{21}\boldsymbol{L}_{a\text{-}i} & \lambda_{22}\boldsymbol{L}_{\text{audio}} & \lambda_{23}\boldsymbol{L}_{a\text{-}t} & \lambda_{24}\boldsymbol{L}_{a\text{-}s} \\ \lambda_{31}\boldsymbol{L}_{t\text{-}i} & \lambda_{32}\boldsymbol{L}_{t\text{-}a} & \lambda_{33}\boldsymbol{L}_{\text{text}} & \lambda_{34}\boldsymbol{L}_{t\text{-}s} \\ \lambda_{41}\boldsymbol{L}_{s\text{-}i} & \lambda_{42}\boldsymbol{L}_{s\text{-}a} & \lambda_{43}\boldsymbol{L}_{s\text{-}t} & \lambda_{44}\boldsymbol{L}_{\text{shot}} \end{vmatrix} \tag{1}$$

其中，矩阵 $\boldsymbol{L}_{\text{image}}$、$\boldsymbol{L}_{\text{audio}}$、$\boldsymbol{L}_{\text{text}}$ 分别代表同一模态内的相似度矩阵，即它们的每个元素分别表示图像与图像、音频与音频、文本与文本之间的相似性关系。矩阵 $\boldsymbol{L}_{\text{shot}}$ 即为镜头之间的相似度矩阵；而矩阵 $\boldsymbol{L}_{a\text{-}i}$、$\boldsymbol{L}_{t\text{-}i}$、$\boldsymbol{L}_{t\text{-}a}$ 则是指不同模态之间的相关度矩阵，如矩阵 $\boldsymbol{L}_{a\text{-}i}$ 的每个元素表示了图像与音频之间的相关性关系，矩阵 $\boldsymbol{L}_{t\text{-}i}$ 是图像与文本的相关性关系，矩阵 $\boldsymbol{L}_{t\text{-}a}$ 是音频与文本之间的关联度；另外，矩阵 $\boldsymbol{L}_{s\text{-}i}$、$\boldsymbol{L}_{s\text{-}a}$、$\boldsymbol{L}_{s\text{-}t}$ 为图像、音频、文本与镜头的相关性矩阵。每个子矩阵 $\boldsymbol{L}$ 的维数均为 $N\times N$，并且参数 λ 需满足如下关系：

$$\sum_{\forall j}\lambda_{ij}=1 \tag{2}$$

2.2 统一相似度矩阵(USM)定义

统一相似度矩阵定义如下：

$$\begin{vmatrix} S_{11} & S_{12} & \cdots & S_{1T} \\ S_{21} & S_{22} & \cdots & S_{2T} \\ S_{31} & S_{32} & \cdots & S_{3T} \\ \vdots & \vdots & & \vdots \\ S_{1T} & S_{2T} & \cdots & S_{TT} \end{vmatrix} \tag{3}$$

其中，统一相似度矩阵的每个元素 $S_{a,b}$ 表示每个数据对象(在本文中，即为图像、音频、文本和镜头)在一个统一的子空间中的相似度关系。T 是统一子空间中数据对象的总数，即 $T=4\times N$。值得注意的是，每个数据对象在 URM 和 USM 的排列顺序应该是一致的。

3 本体

知识本体(ontology)本来是哲学中的一个概念，它描述的是某一特定领域内的概念以及概念与概念之间的关系；本体是一套得到大多数人认同的、关于概念体系的、明确的、形式化的规范说明，是对概念体系的明确的、形式化、可共享的规范说明[8]。

由于本体中的概念是得到一致认可的，比较规范化，用本体来表达语义是合适的，因此本文用本体中的概念对训练数据集进行标注，更加适合基于语义的视频检索的研究，更具有一般性、规范性。

4 Simfuion 和本体的结合

通过分析 Simfuion 和本体的特点，本文利用 TRECVID[9] 提供的镜头序列，先根据镜头的内容变化提取关键帧，然后提取镜头的图像、音频、文本多模态特征。

4.1 镜头之间相似度的计算

本文借鉴 Simfusion 算法思想，计算镜头之间的相似度，具体过程如下：

Step1 初始化 URM。

(1)同种模态间的相似度矩阵 $\boldsymbol{L}_{\text{image}}$、$\boldsymbol{L}_{\text{audio}}$、$\boldsymbol{L}_{\text{text}}$ 元素值的初始化。$\boldsymbol{L}_{\text{image}}$、$\boldsymbol{L}_{\text{audio}}$ 用欧式距离公式计算，$\boldsymbol{L}_{\text{text}}$ 计算余弦距离。

(2)各模态与镜头之间相关度矩阵 $\boldsymbol{L}_{s\text{-}i}$、$\boldsymbol{L}_{s\text{-}a}$、$\boldsymbol{L}_{s\text{-}t}$ 的初始化。设定图像、音频或文本与它们所属的镜头之间相关度为 1，其他均为 0，即为单位矩阵；同样地，$\boldsymbol{L}_{\text{shot}}$ 的初始也设为单位矩阵。

(3)不同模态间的初始相关度矩阵 $\boldsymbol{L}_{a\text{-}i}$、$\boldsymbol{L}_{t\text{-}i}$、$\boldsymbol{L}_{t\text{-}a}$ 初始化。通过计算异构数据之间相关度的算法——共生数据嵌入(Co-Occurrence Data Embedding, CODE[10]方法来计算得到。

(4)与图像特征相关的矩阵 $\boldsymbol{L}_{\text{image}}$、$\boldsymbol{L}_{a\text{-}i}$、$\boldsymbol{L}_{t\text{-}i}$ 的初始化。先用欧式距离分别计算镜头两两帧图像间相似度，然后取这些相似度值平均值作为 $\boldsymbol{L}_{\text{image}}$ 矩阵元素的初始化值，这样可以更准确地反映镜头间的相似度。对于其他矩阵 $\boldsymbol{L}_{a\text{-}i}$、$\boldsymbol{L}_{t\text{-}i}$ 元素值初始化时，类似计算。

为了便于说明，这里假设有 2 个镜头，根据镜头内容变化提取关键帧，假设每个镜头分别提取出 2 幅关键帧，$\boldsymbol{L}_{\text{image}}$ 值计算如表 1 所示。

表 1 $\boldsymbol{L}_{\text{image}}$ 元素值计算

镜头 2	镜头 1	
	f_{11}	f_{12}
f_{21}	x_1	x_2
f_{22}	x_3	x_4

其中，f_{11}、f_{12} 代表镜头 1 对应的关键帧；f_{21}、f_{22} 代表镜头 2 对应的关键帧。$\boldsymbol{L}_{\text{image}}$ 矩阵元素值 $f=(x_1+x_2+x_3+x_4)/2$。

Step2 记矩阵 $\boldsymbol{S}_{\text{usm}}$、$\boldsymbol{L}_{\text{urm}}$ 分别表示统一相似度矩阵和统一关系矩阵。在 USM 上迭代计算，得到镜头之间的相似度，具体步骤如下：

(1)将矩阵 ***original*** 初始设为单位矩阵，首先采用 Simfusion 算法中最基本的相似度增强公式计算。

$$\boldsymbol{S}_{\text{usm}}^{\text{new}}=\boldsymbol{L}_{\text{urm}}\boldsymbol{S}_{\text{usm}}^{\text{original}}\boldsymbol{L}_{\text{urm}}^{T} \tag{4}$$

(2)采用式(5)继续进行迭代计算，直至收敛或得到比较令人满意的结果 $\boldsymbol{S}_{\text{usm}}^{\text{final}}$。

$$\boldsymbol{S}_{\text{usm}}^{\text{n}}=\boldsymbol{L}_{\text{urm}}\boldsymbol{S}_{\text{usm}}^{\text{n-1}}\boldsymbol{L}_{\text{urm}}^{\text{T}}=\boldsymbol{L}_{\text{urm}}^{\text{n}}\boldsymbol{S}_{\text{usm}}^{0}\left(\boldsymbol{L}_{\text{urm}}\boldsymbol{T}\right)^{n} \tag{5}$$

(3)如同矩阵 $\boldsymbol{L}_{\text{urm}}$ 一样，将 $\boldsymbol{S}_{\text{usm}}^{\text{final}}$ 分割为 4 个 $N\times N$ 相似度矩阵，如式(6)所示：

$$\boldsymbol{S}_{\text{usm}}^{\text{final}}=\begin{bmatrix} \boldsymbol{S}_{\text{image}} & \boldsymbol{S}_{i\text{-}a} & \boldsymbol{S}_{i\text{-}t} & \boldsymbol{S}_{i\text{-}s} \\ \boldsymbol{S}_{a\text{-}t} & \boldsymbol{S}_{\text{audio}} & \boldsymbol{S}_{a\text{-}t} & \boldsymbol{S}_{a\text{-}s} \\ \boldsymbol{S}_{t\text{-}i} & \boldsymbol{S}_{t\text{-}a} & \boldsymbol{S}_{\text{text}} & \boldsymbol{S}_{t\text{-}s} \\ \boldsymbol{S}_{s\text{-}t} & \boldsymbol{S}_{s\text{-}a} & \boldsymbol{S}_{s\text{-}t} & \boldsymbol{S}_{\text{shot}} \end{bmatrix} \tag{6}$$

那么，右下角的子矩阵 $\boldsymbol{S}_{\text{shot}}$ 就表示了镜头之间的相似度关系，并且作为下一步进行降维的输入条件之一。

4.2 标注样本

用本体知识库中的概念标注样本，训练分类器 SVM，具体过程如下：

Step1 对 TRECVID 提供的自动语音识别(ASR)结果进行处理，得到每个镜头对应的关键词集合。

Step2 把每个镜头的关键词集合与本体中的同义词集合进行匹配，得到表达镜头的预概念集合。

Step3 对预概念集合进行歧义消解，得到最终的概念集合。

Step4 用最终的概念集合来标注训练集合。

Step5 训练分类器 SVM。

5 实验结果与分析

本文采用 TRECVID 提供的真实视频数据和标注信息进行测试。选取了爆炸(explosion)、飞机(airplane)、大厦(building)、道路(road)、办公室(office)、卡车(truck)、囚犯(prisoner)、企业领导(corporate leader)、体育(sports)、军事(military)、和天气预报(weather)11 个语义概念来测试，这些概念基本涵盖了新闻视频中人们感兴趣的方面。

文本特征直接采用了 TRECVID 提供的 ASR 结果；图像

特征利用的是颜色直方图、纹理和边界特征；音频特征是用一个镜头对应的音频片段为单位，得到短时音频帧特征，然后，在此基础上计算音频帧特征向量的统计值或方差作为音频特征。采用支持向量机 SVM 分类器模型，用由 LPP[11]降维得到的低维语义空间的坐标和用本体概念对训练集进行标注的信息作为 SVM 的输入，训练分类器。

采用由美国国家标准和技术研究机构(National Institute of Standards and Technology, NIST)提出的 AP(Average Precision)和 MAP(Mean Average Precision)作为评估准则。

实验由 3 个部分组成：

(1)每个镜头提取一个关键帧和根据镜头内容变化提取关键帧的比较，结果如图 1 所示。

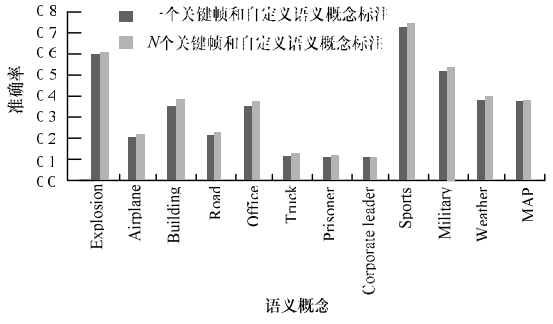

图 1 不同关键帧数的实验结果比较

(2)每个镜头提取一个关键帧时用自定义概念标注和用本体知识库中的概念标注的比较，结果如图 2 所示。

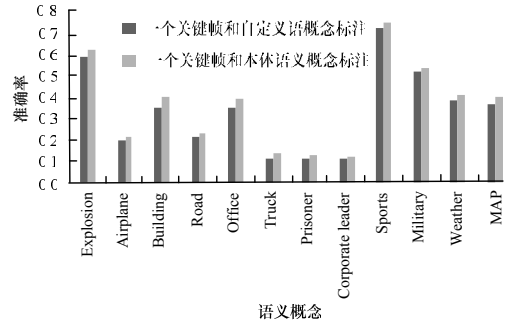

图 2 自定义和本体标注的实验结果比较

(3)根据镜头内容变化提取关键帧时，用自定义概念标注和用本体知识库中的概念标注的比较，结果如图 3 所示。

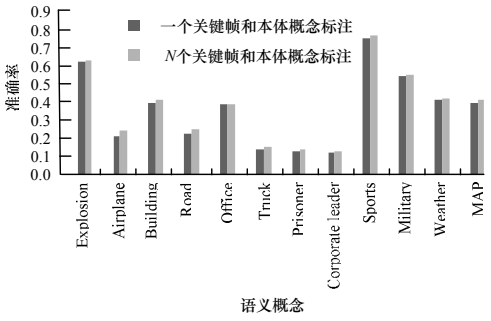

图 3 本体标注时不同关键帧数实验结果比较

图 1 表明，每个镜头只选取一个关键帧与根据镜头内容变化来提取的关键帧相比较，后者更能全面地代表镜头的内容，这样有利于镜头之间相似度的计算，从而提高语义概念提取的准确率。图 2 表明，用本体中的概念标注训练集与传统的用自定义概念标注训练集相比较，本体标注方法更加准确表示镜头的内容，这样也有利于语义概念提取准确率的提高。同时，这样的方法更具有规范性、普遍性。图 3 表明，根据镜头内容变化提取关键帧、有效利用多模态之间的关系与结合本体标注训练集相结合的方法，比传统方法以及其他结合方法在提取语义概念研究方面有更高的准确率。

6 结束语

本文提出的基于 Simfusion 和本体相结合的语义概念提取方法，能够利用镜头内容和多模态之间的关联共生特性，提高镜头之间相似度计算的准确率，更好地反映镜头之间的关系，同时也克服了传统方法中前期和后期融合算法训练过程复杂等缺点。采用 TRECVID 的真实视频数据和标注信息进行测试，结果表明，该方法与传统语义概念提取方法相比具有较高的准确度，且易于实施。

参考文献

[1] 史迎春, 王 韬, 周献中. 基于语义的新闻视频检索研究[J]. 计算机工程, 2004, 30(16): 155-157.

[2] 刘安安, 杨兆选, 李锦涛. 新闻视频结构化浏览与标注系统[J]. 计算机工程, 2009, 35(1): 33-35.

[3] Snoek C G M, Worring M. Multimedia Event-based Video Indexing Using Time Intervals[J]. IEEE Trans. on Multimedia, 2005, 7(4): 638-647.

[4] Xi Wensi, Fox E A. Simfusion: Measuring Similarity Using Unified Relationship Matrix[C]//Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York, USA: ACM Press, 2005: 130-137.

[5] Liu Yanan, Liu Feiwu. Video Semantic Concept Detection Using Multi-modality Subspace Correlation Propagation[C]//Proc. of the 13th Int'l Multimedia Modeling Conference. Berlin, Germany: Springer, 2007: 527-534.

[6] Chau W S, Au O C, Chan T W, et a1. Optimal Key Frame Selection Using Visual Content Metric[C]//Proc. of IEEE ICCCS'05. [S. 1.]: IEEE Press, 2005: 551-555.

[7] Dirfaux F. Key Frame Selection to Represent a Video[C]//Proc. of IEEE ICIP'00. [S. 1.]: IEEE Press, 2000: 275-278.

[8] 李 景. 本体理论在文献检索系统的中的应用研究[M]. 北京: 北京图书馆出版社, 2005.

[9] TRECVID[Z]. (2009-10-20). http://www-nlpir.nist.gov/projects/trecvid/.

[10] Globerson A, Chechik G, Pereira F, et al. Euclidean Embedding of Co-occurrence Data[J]. Journal of Machine Learning Research, 2007, 8: 2265-2295.

[11] He Xiaofei, Niyogi P. Locality Preserving Projections[C]// Advances in Neural Information Processing Systems Conference. Cambridge, USA: MIT Press, 2003: 153-160.

编辑 索书志