

声明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名： 日期：

论文授权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名： 导师签名： 日期：

摘要

随着科学技术的进步和发展,尤其是多媒体技术以及计算机互联网技术的迅猛发展,图像和视频等多媒体数据凭借其生动、形象和直观的特点,逐步成为人们生活、工作和学习过程中不可或缺的重要信息来源。然而由于视频底层特征和高层语义之间语义鸿沟的存在,目前视频内容分析和检索技术的发展远远落后于视频数据的增长速度。因此,对海量视频数据进行快捷有效的分析,从中提取准确的语义信息,缩短乃至跨越语义鸿沟,成为多媒体分析和计算机视觉领域一个重要的研究问题。视频的语义概念分析与提取,需要解决以下两个基本问题:第一,对视频数据内容进行分析和建模,提取有效的视频内容表达;第二,在充分挖掘视频数据内在联系的前提下,设计合理有效的语义概念提取方法。

本文围绕着上述两个基本问题,借助统计分析和机器学习方法,充分挖掘视频数据中内在的时空上下文来辅助目标语义概念的理解,建立底层特征到高层语义概念的合理映射。本文的主要工作和贡献包括:

第一,针对视频内容分析,提出了一种结合空间上下文的弹球模型(Spatial Pachinko Allocation Model, SPAM)。SPAM采用了一种自顶向下的层次隐含主题结构,将视频内容分为三层表示:最底层是视觉关键词层,对应于视频的底层特征及其空间上下文;隐含子主题层位于视觉关键词层之上,对应于视觉关键词在不同子主题下的重要性和相关性;位于最顶层的是建模隐含子主题空域相关性的超主题层。SPAM综合考虑了视觉关键词空间上下文以及隐含主题空域相关性,有效地挖掘了视频内容在隐含主题空间的潜在模式。实验表明,与已有隐含主题分析方法相比,基于SPAM的相关隐含主题分析方法具有更加有效的视频内容描述能力。

第二,通过融合隐含主题分析模型和判别模型,提出了一种基于SPAM和多核函数学习的视频语义概念混合学习方法。混合学习方法吸收了SPAM在视频内容描述方面优势,而且融合了多核学习算法分类精度高的特点,有效地挖掘了不同底层特征的重要性,抑制了噪声对语义概念学习的影响。在TRECVID'05数据集上,混合学习方法取得了与目前国内外前沿的视频语义概念学习和标注方法性能相当的结果。

第三, 针对视频数据的多概念标注问题, 提出了一种序列化多概念标注模式 (Sequence Multi-Labeling, SML)。SML 对视频镜头序列同时标注由多个语义概念组成的标签序列。视频镜头序列中语义概念的空域相关性、单一语义概念的时域一致性以及语义概念之间的时域依赖性与时空上下文线索可以在统一的序列化学习框架中得到充分利用。针对 SML, 提出了一种基于多核函数的判别模型 (Sequence Multi-Label Support Vector Machine, SVM^{SML})。SVM^{SML} 利用混合核函数刻画视频镜头序列中语义概念和底层特征的依赖性、语义概念的时域和空域相关性。在此基础上, 提出了混合核函数学习方法以及基于二值化马尔可夫随机场的标签序列快速预测方法。在 TRECVID'05 和 TRECVID'07 数据集上, SVM^{SML} 取得了显著优于已有的视频语义概念学习和标注方法的结果。

第四, 基于网络视频数据, 设计并实现了视频内容分析和检索系统 VDroid。VDroid 在“Bag-of-words”表示模型基础上添加了空间位置信息, 实现了视觉关键词的倒排索引, 显著提高了视频内容的检索效率。在语义概念检索模块中, VDroid 借助于本文介绍的视频内容分析以及视频语义概念标注方法, 实现了视频类别分类和基于语义概念的检索。在总计约为 400 小时的网络视频数据上, 一系列的实验验证了 VDroid 的有效性。

关键词: 视频内容分析, 时空上下文, 视频语义概念标注, 视频检索

Video Annotation with Spatial and Temporal Context

Yuanning Li (Computer Application)

Supervised by Professor Wen Gao

With the development of capturing, storage, and delivering abilities, large amounts of video data have become available. However, due to the well-known semantic gap between low-level features and high-level concepts, our capabilities of interpret and index such rich corpora have lagged behind. Hence, how to develop effective video annotation techniques is challenging yet important problem for content-based video indexing and retrieval. Video annotation needs to solve two basic problems: the first is how to extract informative video content representation; the second is how to mine the intrinsic correlations of video data and build effective mappings from low-level features to high-level concepts.

Video is by nature informative in spatial and temporal context. In this dissertation, we focus on two basic problems and aim to develop effective video annotation techniques by discovering the intrinsic patterns and their context within video data. The main contributions of this dissertation are summarized as follows:

Firstly, a correlated latent topic model (i.e., Spatial Pachinko Allocation Model, SPAM) is proposed for video content analysis. SPAM models video content in a three level topic model: the first level consists of visual words corresponding to the low-level features and their spatial context; the second level consists of sub-topics to discover the importance and the correlation of visual words in different sub-topics; the third level consists of super-topics which are utilized to model the spatial correlation of sub-topics explicitly. Compared with existing topic models, SPAM discovers the intrinsic patterns of video content in topic space by jointly considering the spatial context of visual words and the spatial correlation of topics. Experiments on video content analysis show the effectiveness of the proposed method.

Secondly, a hybrid learning method of SPAM and Multiple Kernel Learning (MKL) is proposed for semantic concept learning. Hybrid method fuses multiple topic spaces which are inferred by SPAMs in a multiple kernel combination. Via multiple kernel learning, the optimal kernel weights of the topic spaces are learnt together with the MKL-based classifier, effectively depressing the noise information. Accordingly, the advantages of SPAM in video content analysis and MKL in classification accuracy can be integrated within the same supervised learning process. Extensive experiments on TRECVID'05 dataset demonstrate the effectiveness of such hybrid method, which yields comparable performance to the state-of-the-art.

Thirdly, a novel formulation (i.e., Sequence Multi-Labeling, SML) is proposed for video annotation. Different from existing video annotation schemes working on individual shots or adjacent shots, SML predicts a multi-label sequence for a shot sequence. Different types of spatial and temporal context, such as spatial correlation of concepts, temporal consistency of individual concept and temporal dependency of different concepts, can be integrated within the SML framework. For SML, a multiple kernel based discriminative model (i.e. Sequence Multi-Label Support Vector Machine, SVM^{SML}) is proposed. In SVM^{SML} , a joint kernel is employed to measure the dependency of semantic concept over low-level features, spatial and temporal context within the shot sequence jointly. Accordingly, a multiple kernel learning method over shot sequences is proposed to learn the joint kernel as well as the SVM^{SML} score function. To make search more efficient over the large-scale output space of multi-label sequence, an approximate method which maximizes the energy of a Binary Markov Random Field (BMRF) is presented. Extensive experiments on TRECVID'05 and TRECVID'07 datasets show that the proposed SVM^{SML} gains superior performance over the state-of-the-art.

Finally, a video retrieval system (i.e., VDroid) is implemented for website videos. In VDroid, “Bag-of-Words” representation is employed and a reverted index of visual words with spatial context is created for efficient content-based search over large video corpus. For concept-based retrieval, VDroid utilizes video content analysis and video annotation methods presented in this dissertation for semantic concept annotation and video genre classification. Experiments over 400 hour’s website videos demonstrate the effectiveness of VDroid.

Keywords: video content analysis, spatial and temporal context, video annotation, video retrieval