DOI:10.13878/j.cnki.jnuist.2020.03.002

刘波[1,2]

视频摘要研究综述

摘要

近年来,随着计算机技术的发展和终端设备的广泛使用,视频摘要技术得到了广泛的研究.视频摘要是数据摘要的重要研究方向.首先介绍了静态视频摘要的基本概念,然后对研究静态视频摘要的凸松驰方法和行列式点过程法的最新研究进展进行了概述.对于动态视频摘要,主要对分割视频和个性化视频摘要的最新研究进展进行了介绍.最后对视频摘要面临的问题以及将来的研究方向进行了介绍.

关键词

视频分段;动态视频摘要;静态视频摘要;个性化

中图分类号 TP13

文献标志码 A

收稿日期 2020-01-01

作者简介

刘波,男,博士,副教授,主要研究方向为机器学习、视频分析.liubo7971@163.com

1 重庆工商大学 人工智能学院,重庆,400067

2 重庆工商大学 计算机科学与信息工程学院,重庆,400067

0 引言

随着互联网的快速发展以及计算机、终端设备的广泛使用,使得视频数据呈爆炸式增长.据统计,2018 年中国观看网络视频的人数已经达到 6.09 亿,而著名的视频网站 YouTube 每分钟上传的视频有 300 小时左右.视频是人们共享和获取信息的有效载体.

图像、声音和文字是构成视频内容的三个基本要素,这些信息可以构成事件、动作等连续的信息.对于海量的视频,人们若要获取它的主要信息,通常需要观看完整个视频,这将花费大量时间.因此,需要寻找方法来让人们能从视频中迅速获取主要内容.视频摘要是解决该问题的重要方法.所谓视频摘要,就是指从视频中提取包含视频主要内容的视频帧或视频段(Video Segmentation).视频摘要主要涉及的视频类型包括:1)电影、电视节目(比如新闻、体育、娱乐等)的视频.2)视频监控领域.这类视频数量巨大,内容变化较少,视角通常固定.3)Egocentric 视频[1],也称第一人称视频(the First Person Video),通常是指由可穿戴设备(比如 Google 眼镜、微软的 AR 眼镜)所摄像的视频.这类视频的特点是内容会出现较多的遮挡,视角变换频繁、视频内容变化明显、时间长.4)用户视频,通常包含一组有趣的事件,但未经编辑.这类视频通常比较长,存在大量冗余内容[2].

Pfeiffer 等在 1996 年首次提出了视频摘要的概念[3].目前,研究视频摘要主要有两类方法:1)静态视频摘要方法,也称关键帧选择方法.该方法通过提取或选择视频中具有代性的帧(即关键帧(Key Frame))来精简视频的内容.这种方法获得的视频摘要不具有连贯的动态信息和语音信息,所表达的信息有限.2)动态视频摘要,也称为视频剪辑(Video Skimming).该方法通过保留连续的小视频段来实现对视频内容的精简.本文将对这两种视频摘要方法进行详细介绍.

1 静态视频摘要

静态视频摘要的目标是从给定的视频中选择出具有代表性的帧,选择的标注是代表性(representative)和多样性(diversity),有些文献也称代表性为重要性(importance).如果把视频当成一个集合,则每一帧就是集合中的元素,因此选择关键帧的问题可以看成是子集选择(subset selection)问题.子集选择又称为范例选择,在人工智能领域有着广泛的应用,比如从大量的图像中选择具有代表性的图像展示

给不同用户就是一个子集选择问题[4].子集选择是一个 NP 难问题,人们通过各种优化方法来获得它的近似解.

按照求解子集选择方法的不同,静态视频摘要的方法可分为凸松弛(Convex Relaxation)优化、行列式点过程(Determinantal Point Process)等.下面分别对这些方法进行介绍.

1.1 凸松弛的静态视频摘要

通常子集选择问题都得不到全局最优解.为了解决这个问题,人们将子集问题转换为凸规划问题,以便能获取近似解,这种转换也称为凸松弛.2012年,Elhamifar 等[5]在数据集 X 上通过构造样本点的不相似性来选择范例,将行稀疏作为目标函数的正则项,并通过凸优化方法来求解目标函数.该方法在视频摘要上取得了较好的效果.随后他们对原来的方法进一步改进[6],通过在原集合 X 和目标集合 Y 之间构造逐点不相似性(pairwise dissimilarities)来获得具有代表性的样本集,然后通过稀疏恢复的方法来求解目标函数.最近,范例选择被用于动态时序数据中[7],即对于给定的时序数据集 $X=[x_1,x_2,\cdots,x_n]$,$p(x'\mid x_{i_1},\cdots,x_{i_k})$,需要找出 X 中的范例来表示时序数据集 $Y=[y_1,y_2,\cdots,y_T]$ 中的样本.目标函数由 3 个势函数相乘得到,这 3 个势函数分别为:编码势函数(Encoding Potential)、基数势函数(Cardinality Potential)和动态势函数(Dynamic Potential).该问题最终可以转换为一个整数规划问题,并通过最大和消息传递(max-sum message passing)来求解.当多个摄像头对同一位置进行监控时,由于每个摄像头拍摄的视角(View Point)不一样,会呈现多个视图.在对这一位置的监控视频生成摘要时,需要考虑多个视图的相关性(correlation),这种视频摘要称为多视图视频摘要.多视图视频摘要面临两个重要的问题:1)数据量大;2)来自各个摄像头的数据具有一定的相关性.为了有效解决这些问题,Panda 等[8]提出基于子空间嵌入和稀疏表示的多视图视频摘要方法.所提出的方法同时约束一个视频内的相关性和视频之间的相关性,从而提高了关键帧的差异性和稀疏性.

凸松弛方法所找到关键帧通常含有的信息量比较大,具有很好的代表性,但有可能差异性不大.为了提高凸松弛方法所选择的帧的差异性,Wang 等[9]采用结构稀疏作为目标函数的正则项,其中,结构稀疏正则项由行稀疏正则项、局部敏感正则项和差异性正则项组成.差异性正则项主要用于提高关键帧的差异性.具体而言,对于给定的两帧 x_i,x_j,分别找到与这两帧最不相似的帧,并得到它们的不相似值 d_1,d_2,如果 x_i,x_j 的相似度 d_{ij} 比 d_1,d_2 都大,则取 d_{ij} 作为线性组合的系数.最终得到的差异性正则项公式为

$$R_{\mathrm{div}}=\sum_{ij} d_{ij}\|x_i-x_j\|_1 .$$

1.2 行列式点过程的静态视频摘要

行列式点过程是一种概率模型,它最早由 Macchi 于 1975 年提出[10].对于一个给定的整数集 $I=\{1,2,\cdots,N\}$,总共可以得到 2^N 个子集,对于其中的一个子集 $y\subseteq X$ 被选中的概率为

$$P(y;\boldsymbol{L})=\frac{\det(\boldsymbol{L}_y)}{\det(\boldsymbol{L}+\boldsymbol{I})},$$

其中,$\boldsymbol{L}$ 是对称正定矩阵的相似矩阵,$\boldsymbol{I}$ 是单位矩阵,$\boldsymbol{L}_y$ 是子矩阵,它的行和列是根据 y 中的数字从 $\boldsymbol{L}$ 中抽取出来.将行列式点过程用于视频摘要的原理为:将 y 看成是提取的视频帧的编号集合,若提取了完全相同的两帧,$\boldsymbol{L}_y$ 就有完全相同的两列和两行,因此它的行列为 0,从而导致其对应的概率为零.

在使用行列式点过程来选择关键帧时,需要构建矩阵 $\boldsymbol{L}$. Zhang 等[11]通过监督方式来构建矩阵 $\boldsymbol{L}$.首先给出一组标注好的视频摘要,将测试视频中的第 i 帧和第 j 帧取出来与标注好的帧进行逐一比较,选对相似度最大的帧,并计算相似值,将这些相似值加到一起作为矩阵 $\boldsymbol{L}$ 的第 i 行、第 j 列的元素.构造好矩阵 $\boldsymbol{L}$ 后,再通过经典的行列式点过程算法来得到最终要选择的帧.

由于视频的帧具有很强的时序性,因此一个自然的想法就是在基于行列式点过程中加入时序特性,以获得具有更好差异性的视频摘要.目前时序行列式点过程(Sequential Determinantal Point Processes,SeqDPP)是关键帧选择的重要研究方向.时序行列式点过程最早由 Gong 等[12]提出.对于一个给定的视频 V,将其按时间分成 T 个不相交的视频段 $\cup_{i=1}^{T}v_i=V$,设 t 时刻从 V 选择的关键帧子集变量为 Y_t,并设 t 时刻 v_t 对应的某个子集为 y_t,则给定 $t-1$ 时刻的某个子集 y_{i-1} 时,得到 y_t 的条件概率为

$$P(Y_t=y_t\mid Y_{t-1}=y_{t-1})=\frac{\det\boldsymbol{\Omega}_{y_{t-1}\cup y_t}}{\det\boldsymbol{\Omega}_t+I_t},$$

其中,$\boldsymbol{\Omega}_t$ 表示 $y_{t-1}\cup y_t$ 所对应的 $\boldsymbol{L}$ 矩阵.得到条件概率的定义之后,就可以得到所有子集的联合概率

分布：

$$P(Y_1=y_1,Y_2=y_2,\cdots,Y_T=y_T)=P(Y_1=y_1)\prod_{i=2}P(Y_i=y_i\mid Y_{t-1}=y_{t-1}).$$

最后通过后验概率推理来求解联合概率分布：

$$y_1^*=\arg\max_{y\in v_1}P(Y_1=y),$$

$$y_2^*=\arg\max_{y\in v_2}P(Y_2=y\mid Y_1=y_1^*),$$

$$\vdots$$

$$y_t^*=\arg\max_{y\in v_t}P(Y_t=y\mid Y_{t-1}=y_{t-1}^*).$$

Li 等[13]在时序行列式点过程的基础上引入强化学习来解决视频段划分问题.经典的时序行列式点过程虽然考虑了视频的时序特性,但是并没有考虑如何将视频合理地划分成长度不相等的段.所提出的算法能通过隐变量来动态得到视频段的长度,然后划分视频段,整个过程能通过算法自动推理实现.他们还针对所提出的模型给出了一种有效的训练策略.

Sharghi 等[14]根据用户输入的查询信息来对长视频获得关键帧集合.所提出的算法总共分成两部分:1)在序列行列式点过程的基础上,将查询信息作为条件概率的一部分来得到视频段,这是通过 Z 层(Z-Layer)来完成的;2)对得到的视频段,通过时序行列式点过程来去掉不相关信息.提出的模型限制用户输入的查询信息只能是一个或多个名词组合.为了解决时序行列式点过程的偏差问题,Sharghi 等[15]提出了一种基于大间隔的视频摘要算法,该算法能根据用户输入长度来执行视频摘要.

2 动态视频摘要

动态视频摘要主要包括视频分割、视频段重要性评价、选择视频段并形成视频摘要.其中视频分段和选择视频段是动态视频摘要最重要的部分.视频分段将视频分成多个场景或镜头(shots),是动态视频摘要的基础.选择视频段则是根据具体任务选择满足要求的视频段,这通常会很困难,因为不同的人喜欢不同的视频内容,选择出的视频不可能让大家都喜欢,因此,个性化视频摘要是动态视频摘要研究的重要方向.接下来将对视频分割和个性化视频摘要的相关研究进展进行介绍.

2.1 分割视频段

最初人们是通过直方图和图像强度来对结构化视频进行分段并取得了好的效果.2014 年,Gygli 等[2]通过超帧(superframe)来对视频分割,并通过定义的能量函数来评价视频段.为了计算超帧的评分,需对每帧进行评分,然后将这些帧的评分加起来得到超帧的评分.在计算每帧的评分时,会利用帧的低级特征(比如对比度和时空信息显著性等)和高级特征(比如动作和人脸等)的信息.最后利用整数规划来选择视频段.Potapov 等[16]提出了一种变化点(change point)的视频段分割方法.变化点常被用来测信号中的跳跃.他们所提出的算法采用核变化点来检测视频帧的变化情况,在变化较大的地方作为视频帧分段的界线.Ngo 等[17]对结构化视频用谱聚类和时间图分析来进行场景建模,然后通过动作注意建模来进行重要视段段的检测.该算法的具体过程为:1)将视频按时序分成不同的镜头(shots)和子镜头(sub-shots);2)用谱聚类对这些镜头聚类,用注意力模型得到这些镜头的注意力值;3)通过聚类信息和注意力值生成时空图;4)对场景建模和检测;5)生成视步摘要.

Xu 等[18]针对 Egocentric 视频摘要提出了基于凝视(gaze)跟踪信息的视频摘要方法.研究表明图像中内容的相对重要性与人在空间和时间上的注意力分布相关.通常的 Egocentric 视频由可穿戴设备生成,因此凝视产生的视频能够体现佩戴人的意图,从而实现个性化的视频摘要.通过镜头中注视帧(fixation frame)的数量可以得到镜头注意力的评分.所提出的算法通过凝视信息来分段,具体的操作过程为:1)提取每帧视频中的凝视跟踪信息(包括注视、扫视和眨眼);2)去掉有错误的眼部跟踪数据的帧;3)对得到的每段视频选择中心帧作为关键帧,通过深度神经网络 R-CNN 提取这些关键帧的特征,主要提取大小为 100×100 的凝视区域的特征;4)计算关键帧之间的余弦相似度;5)将连续的视频段合并成子镜头,合并的原则是如果相邻视频段的相似距离是在 0.5 及以上,就合并,否则就不合并;6)对于合并后的子镜头,再次选择中心关键帧,并用 R-CNN 计算这些关键帧的特征描述符,若有 k 个子镜头,最后形成的子镜头描述符集合为 $V=\{v_1,v_2,\cdots,v_k\}$.

2.2 个性化的动态视频摘要

随着电子商务的普及,推荐系统成为研究热点,人们想根据每个人的爱好生成相应的视频摘要(即个性化视频摘要),这与个性化推荐相似.个性化视频摘要的研究属于视频摘要的新兴领域.

Xiang 等[19]从情绪基调(emotional tone)、局部主要特性和全局主要特性出发,对视频的个性化推

荐进行了研究.为了得到镜头的情感标记,分别提取相应的音频特征和图像特征,然后再由情感分析模型来对镜头进行标记.这个标记过程也会用到人脸数据.因此视频段对应两种标记:情感标记和人脸标记.通过稀疏情感标记来分析视频的情感状态.

Darabi 等[20]提出了一种根据用户爱好来定制视频摘要的方法.首先由 10 个人根据视频的音频、视觉和文本内容对 6 个不同类别的视频的帧进行评分.然后使用 SIFT 特征描述符按预定义类别来计算每个视频场景的相关性分数,并将这些分类保存在一个矩阵中.接下来以向量的形式得到用户对这些高级视觉概念(类别)的兴趣水平.通过这两组数据来确定用户帧不同视频段的优先级,并根据最终用户生成的配置文件来更新帧的初始平均分数,将得分最高的视频帧作为视频的摘要,并将音频信息和文本内容插入到最终的视频摘要中.

Hant 等[21]通过人工标注关键帧的方式来获得视频段.该方法首先会用图模型的显著性算法来构造显著性映射,该映射由特征映射和激活映射(activation map)组合而成;然后得到帧之间的双向相似性,这种相似性通过帧中图像块之间的余弦距离之和来进行计算.在些基础上,通过 Isomap 算法来完成帧的低维表示.为了计算帧的权重,首先计算帧在低维情形下的时序邻近距离,然后获得观众所选择帧的权重.将这两种权重相加得到帧的最终权重.为了选出关键的视频段,首先用层次聚类算法来找到视频的结构,然后再用整数规划来选择视频段.该论文采用人工方式选择关键帧,并计算这些关键帧的权重,再将这些权重与模型计算的权重融合,从而将个性化引入到视频摘要中.

Yoshitaka 等[22]通过捕获人的动作(比如眼睛移动、播放器操作等)来进行个性化视频摘要.播放操作主要有快进、快退、跳至下一节/上一节,以快速播放、暂停或以慢速播放.在观看视频时,如果不感兴趣,观众通常会进行快进;如果对播放的内容感兴趣,经常会倒回播放或采用慢速播放.所以可以认为倒回播放或慢速重播能表示观众注意力或偏好.作者通过实验说明了播放操作与用户偏好之间的关系.观众在观看视频时,人眼的运动方式也能反映出他对视频内容的偏好.眼睛扫视(saccade)指眼球快速从一个视点转移到另一点视.固视(fixation)是眼睛在注意某个视点(viewpoint)时处于不动状态.基于眼睛所处的状态就能得到观众对视频段的关注或喜爱程度.通过对观众观看足球比赛的视频节目进行测试,可以验证这一观点.基于以上的事实,作者给出了视频摘要的处理流程.

个性化视频摘要还处于起步阶段,仍有很多问题需要解决,比如在根据用户的偏好来选择用户感兴趣的视频段(或关键帧)的研究中,目前的数据集非常有限,仅有的几个数据集都没有给出视频段的评分,更没有将这些视频段与用户的偏好结合起来.

3 总结

视频摘要属于数据摘要的一个分支.由于视频数据是带有时序结构的图像数据,因此很多时序处理的方法(比如长短记忆网络(LSTM)等)都可以用来对视频摘要进行研究,更重要的是很多计算机视觉的方法(比如语义分割、动作识别等)也可以用来对视频摘要进行研究.而视频摘要是一个子集选择问题,可用机器学习、最优化等理论来解决视频摘要的问题.由于视频包含有声音、文字、图像等数据,可将视频摘要看成是一个多模态问题,因此可用多模态方法来对视频摘要建模.

虽然人们对视频摘要进行了广泛研究,但仍有很多问题没有解决,比如:

1)随着网络直播的兴起,在线视频的摘要越来越受到重视,而这方面的研究非常少;

2)随着监控设备的普及,多视图的视频摘要显得越来越重要,而多个摄像机数据融合,多个摄像机数据的相关性等问题都有待解决;

3)目前用于视频摘要研究的数据集很少,著名的视频摘要数据集有 TVSum[23]、SumMe[2].这些数据集都比较小,比如 SumMe 包括 25 个短视频,TVSum 包含 50 个短视频.若需要利用深度学习技术来研究视频摘要,则需要建立更大的数据集.

参考文献

References

[1] Betancourt A,Morerio P,Regazzoni C S,et al.The evolution of first person vision methods:a survey[J].IEEE Transactions on Circuits and Systems for Video Technology,2015,25(5):744-760

[2] Gygli M,Grabner H,Riemenschneider H,et al.Creating summaries from user videos[M]// Computer Vision-ECCV 2014.Cham:Springer International Publishing,2014:505-520

[3] Pfeiffer S,Lienhart R,Fischer S,et al.Abstracting digital movies automatically [J]. Journal of Visual Communication and Image Representation,1996,7(4):

345-353
[4] Simon I, Snavely N, Seitz S M. Scene summarization for online image collections[C]//2007 IEEE 11th International Conference on Computer Vision, 2007: 1-8
[5] Elhamifar E, Sapiro G, Vidal R. See all by looking at a few: sparse modeling for finding representative objects [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2012: 1600-1607
[6] Elhamifar E, Sapiro G, Sastry S S. Dissimilarity-based sparse subset selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(11): 2182-2197
[7] Elhamifar E, Kaluza M C D P. Subset selection and summarization in sequential data[C]//Advances in Neural Information Processing Systems, 2017: 1035-1045
[8] Panda R, Roy-Chowdhury A K. Multi-view surveillance video summarization via joint embedding and sparse optimization[J]. IEEE Transactions on Multimedia, 2017, 19(9): 2010-2021
[9] Wang H X, Kawahara Y, Weng C Q, et al. Representative selection with structured sparsity [J]. Pattern Recognition, 2017, 63: 268-278
[10] Macchi O. The coincidence approach to stochastic point processes[J]. Advances in Applied Probability, 1975, 7(1): 83-122
[11] Zhang K, Chao W L, Sha F, et al. Summary transfer: exemplar-based subset selection for video summarization [J]. arXiv, 2016, arXiv: 1603.03369
[12] Gong B, Chao W L, Grauman K, et al. Diverse sequential subset selection for supervised video summarization[C]//Advances in Neural Information Processing Systems, 2014: 2069-2077
[13] Li Y D, Wang L Q, Yang T B, et al. How local is the local diversity? Reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization [M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 156-174
[14] Sharghi A, Laurel J S, Gong B Q. Query-focused video summarization: dataset, evaluation, and a memory network based approach[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 4788-4797
[15] Sharghi A, Borji A, Li C, et al. Improving sequential determinantal point processes for supervised video summarization[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2018: 517-533
[16] Potapov D, Douze M, Harchaoui Z, et al. Category-specific video summarization [M]// Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 540-555
[17] Ngo C W, Ma Y F, Zhang H J. Video summarization and scene detection by graph modeling[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(2): 296-305
[18] Xu J, Mukherjee L, Li Y, et al. Gaze-enabled egocentric video summarization via constrained submodular maximization[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 2235-2244
[19] Xiang X, Kankanhalli M S. Affect-based adaptive presentation of home videos[C]//Proceedings of the 19th ACM International Conference on Multimedia, 2011: 553-562
[20] Darabi K, Ghinea G. Personalized video summarization using sift[C]// Proceedings of the 30th Annual ACM Symposium on Applied Computing, 2015: 1252-1256
[21] Han B, Hamm J, Sim J. Personalized video summarization with human in the loop[C]//IEEE Workshop on Applications of Computer Vision, 2011: 51-57
[22] Yoshitaka A, Sawada K. Personalized video summarization based on behavior of viewer[C]// Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012: 661-667
[23] Song Y, Vallmitjana J, Stent A, et al. Tvsum: summarizing web videos using titles[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5179-5187

Survey of video summary

LIU Bo[1,2]

1 College of Artificial Intelligence, Chongqing Technology and Business University, Chongqing 400067

2 School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067

Abstract Recently, with the development of computer technology and the widespread use of terminal equipment, video summary technology has been extensively studied, which is an important research direction of data summary. This paper introduces the latest research progress of video summary. Firstly, the basic concepts of static video summary are introduced, and then the latest research progress on the convex relaxation method and determinant point process method of static video summary is surveyed. For dynamic video summaries, the most recent research advances in segmented and personalized video summaries are introduced. Finally, the problems facing the video summary and future research directions are introduced.

Key words video segmentation; dynamic video summary; static video summary; personalization