基于生成式模型的视频基元追踪学习

赵友东[1)] 龚海峰[2)] 贾云得[1)]

[1)](北京理工大学计算机学院 智能信息技术北京市重点实验室 北京 100081)

[2)](莲花山研究院 湖北 鄂州 436000)

摘 要 自然场景视频中含有各种类别的视频基元(video primitives),它们构成了整个高维视频块(video bricks)空间,具有不同的结构维度及复杂度,由空间表观与运动共同描述.视频基元主要有两类:结构视频基元与纹理视频基元.文中使用一个通用生成式模型对两类视频基元进行统一概率建模,每个视频基元的表达能力由其对应的信息增益来度量.利用该度量进行视频基元追踪学习,最终建立一个完整的视频基元集.实验结果显示了文中方法在视频内容表示方面的有效性.

关键词 视频块;视频基元;生成式模型;追踪学习算法;视频内容表示

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.01835

Generative Model Based Atomic Video Primitives Pursuing

ZHAO You-Dong[1)] GONG Hai-Feng[2)] JIA Yun-De[1)]

[1)](*Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing* 100081)

[2)](*Lotus Hill Research Institute (LHI), Ezhou, Hubei* 436000)

Abstract There are a number of video primitives in videos of natural scenes. All video primitives of varying dimensions and rich structures constitute the high dimensional space of video bricks. The structure of a video primitive is characterized by both appearance and motion dynamics. As small video bricks have fewer compositional effects, there is little overlapping between different primitives. Here, the authors categorize video primitives into two types: structural video primitives and textural video primitives. A common generative model is introduced to model them in a unified form. The representation power of a primitive is measured by its information gain, based on which primitives are pursued one by one via a novel pursuit algorithm, and finally a holistic set of video primitives is built up. Promising experimental results on video segmentation and video scene recognition show the potential power of the proposed framework for video representation.

Keywords video brick; video primitives; generative model; pursuit algorithm; video representation

1 引 言

自然场景视频中有各种各样的视觉模式(visual pattern),从简单的天空、运动树枝、行人,到复杂的荡漾水面、摇曳的火苗等,这些视觉模式只有分解到较小的视频块(video bricks),如 15×15×5 像素,才能呈现较纯的结构,从而方便研究.从数学角度来

收稿日期:2010-08-22.本课题得到国家自然科学基金(90920009,60875005)、国家"八六三"高技术研究发展计划项目基金(2009AA01Z323)资助.**赵友东**,男,1983 年生,博士,研究方向为视频内容分析和计算机视觉. E-mail: zyd458@bit.edu.cn. **龚海峰**,男,1978 年生,博士,研究方向为视频内容分析和计算机视觉. **贾云得**(通信作者),男,1962 年生,博士,教授,博士生导师,研究领域为计算机视觉、媒体计算和智能人机交互. E-mail: jiayunde@bit.edu.cn.

看,视频块位于高维(如 15×15×5 维)空间中不同的聚类子空间,构成不同维度与结构的流形,这些聚类子空间称为视频基元(video primitive).在计算机视觉研究领域,图像基元的研究已经有很长一段历史[1-2].近几年,时空视频块获得越来越多的研究[3-5],这些工作主要关注各种判别时空特征或高层的视频事件模型,对视频块空间中各种视频基元的生成式模型研究得很少.文献[6]从稀疏编码的角度对构成自然视频序列的最基本单元进行学习.本文研究视频块空间中视频基元及其对应的数学模型.我们的目标是用一个通用的生成式模型来统一表示这些视频基元模型,研究视频基元追踪学习(pursuing)算法来逐个选择表达能力最强的视频基元,最终建立一个完整的可用于一般视频表示与建模的视频基元集.

视频块的空间分布主要受表观和运动两方面因素的影响.图像表观通常分为两类:结构基元(如物体边界、线条)和纹理.运动也分为两类:可跟踪运动(如运动边缘[7]、运动模板[8]等)和不可跟踪运动(如密集的鸟群飞行、荡漾的水面等).依据表观和运动的分类,将视频基元分为结构视频基元与纹理视频基元.结构视频基元一般由很少几个基向量描述,而纹理视频基元一般由一组统计特性(如滤波响应统计直方图)描述.两种视频基元对应两种不同的子空间,结构视频基元通常是一个低维子空间,而纹理视频基元通常对应高维子空间.一段视频中不同的视频块可能映射到不同子空间(视频基元)上,如图 1 所示.

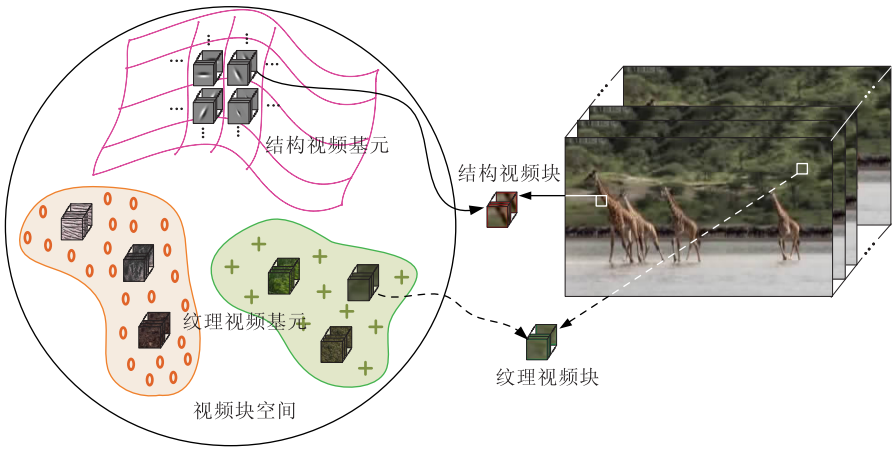


图 1 两个典型视频块分别映射到结构视频基元和纹理视频基元示意图

由于各种视频块子空间的不同维度与复杂结构,不同类型模型的混合比单一模型或单一模型的混合更适合用来对其建模.因此,我们探讨基于信息投影的通用生成式概率模型来统一建模两种类型的子空间.两组具有不同时空尺度的滤波器分别被用作两种视频基元的基本描述符.各个视频基元对整个视频块空间的描述能力由它们的信息增益来度量.依据这个度量,我们将研究视频基元追踪学习算法,它在视频块空间中对基元进行逐个选择,最终,得到按信息增益排序的视频基元集,每个基元描述了视频块空间中的一部分结构.该视频基元集可以应用于多种视频分析任务,比如视频编码、表示与建模等,是中高层视觉任务的基础.动态纹理视频分割与视频场景识别实验显示了本文方法在视频内容表示方面的有效性.

本文工作与 Shi 和 Zhu[9] 的工作密切相关,他们依据图像表观将自然场景图像块映射到两种基本流形上,本文的工作同时考虑时空视频块的表观与运动特征来分析其空间流形分布,对两种视频基元采用了不同于文献[9]中的方法进行有效建模.本文工作也与动态纹理[10-11]、分布式有向能量表达[12]、混合动态纹理[13-14]以及分段混合高斯模型[15]等工作相关.文献[10,12]利用单一模型对时空视频区域进行建模,文献[11,13-15]将视频区域建模为某种模型的混合模型,如 STAR 模型、ARMA 模型或者高斯模型.这些工作在相应的视频分析任务中都取得了很好的结果,但它们仅利用一种模型对各种视频模式进行建模,没有对场景中重要的结构信息进行显式建模.本文工作在生成式模型框架下对两种互补视频基元进行统一建模,可以对视频内容进行更有效的表示建模.

2 视频基元的形式化描述

本文设定视频块空间尺寸为 15×15,时间为

5 帧(约 200ms,与人的早期视觉感知时间间隔相吻合).为了便于描述,将视频块记为$\mathbb{B}$.下面给出两类视频基元的形式化定义.

定义 1. 结构视频基元(Structural Video Primitive,SVP)是由一个或几个基函数张成的子空间

$$\Omega_{\text{svp}}=\left\{\mathbb{B}:\mathbb{B}=\sum_{i=1}^{n}\alpha_i B_i+\varepsilon\right\} \tag{1}$$

其中,基函数 B_i 是从一个时空滤波器池(filter bank)$\Delta_{\text{svp}}=\{B_i\}_{i=1}^{N}$ 中选择得到的,α_i 与 ε 分别是重建系数与重建残差.

时空滤波器池 Δ_{svp} 由各种类型的基函数构成.本文采用 8 个朝向的空间 Gabor 滤波器按照不同速度运动来生成它,并且滤波器运动方向垂直于滤波器朝向.滤波器空间尺度为 13×13 像素,时间尺度为 5 像素,运动速度为 0,±1,±2,±3,±4,±5 和 8 像素.因此,总共有 $N=8\times12=96$ 个时空滤波器.

定义 2. 纹理视频基元(Textural Video Primitive,TVP)是由若干统计特征(如统计直方图)约束的子空间

$$\Omega_{\text{tvp}}=\{\mathbb{B}:H_i(\mathbb{B}*F_i)=H_i^{*}+\varepsilon,\ i=1,2,\cdots,l\} \tag{2}$$

特征 F_i 选自于一个时空滤波器池 $\Delta_{\text{tvp}}=\{F_i\}_{i=1}^{N}$;$(\cdot*\cdot)$表示卷积操作;$H_i(\mathbb{B}*F_i)$是滤波器 F_i 在视频块$\mathbb{B}$中所有像素上响应的统计直方图;H_i^{*} 是特征 F_i 在该聚类上的均值直方图;ε 是视频块统计直方图相对于均值直方图的扰动.实验中使用 8 个朝向的 7×7 Gabor 滤波器,以速度±1,±2,±3 和 5 像素沿法线方向运动或者保持静止.因此,时空滤波器池 Δ_{tvp} 由 $N=8\times8=64$ 个滤波器构成.

对于每一个视频基元 $\Omega^k\in\{\Omega_{\text{svp}},\Omega_{\text{tvp}}\}$,存在一个内在的概率分布 $f_k(\mathbb{B})$ 对其进行描述.建模的目标是学习概率模型 $p_k(\mathbb{B})$ 来近似 $f_k(\mathbb{B})$,该目标由最小化 f_k 与 p_k 之间的 K-L 离散度 $KL(f_k\|p_k)$ 实现.这个模型学习过程是一个特征选择过程.下面给出两类视频基元的统一概率模型:

$$p_k(\mathbb{B}|\Lambda,\Delta)=\frac{1}{Z(\Lambda)}\exp\left\{-\sum_{i=1}^{n}\lambda_i r_i(\mathbb{B})\right\}q_k(\mathbb{B}) \tag{3}$$

其中,$\Delta=\{B_i$ 或 $F_i\}_{i=1}^{n}$ 是选择的特征子集,$\Lambda=(\lambda_1,\cdots,\lambda_n)$是相应的模型参数,$Z(\Lambda)$是划分函数以及$q_k(\mathbb{B})$是参考分布.

如果 Ω^k 是结构视频基元,选择的基函数集 $\Delta=\{B_i\}_{i=1}^{n}\subset\Delta_{\text{svp}}$,一个特征 B_i 在视频块上的响应 $r_i(\mathbb{B})=\max(\text{Sigmoid}(\|\mathbb{B}*B_i\|^2))$(即 $r_i(\mathbb{B})$是特征 B_i 在$\mathbb{B}$上所有点的响应经 Sigmoid 变换调制后的最大值).参考分布 $q_k(\mathbb{B})$通常由自然场景视频块离线估计得到.这个模型与主动基形变模型[16]类似.对于一个视频块$\mathbb{B}$,它属于该视频基元 Ω^k 的程度可由 $\text{Score}^{\text{svp}}=\sum_i r_i(\mathbb{B})$简单近似度量.如果 Ω^k 是纹理视频基元,选择的滤波器集 $\Delta=\{F_i\}_{i=1}^{n}\subset\Delta_{\text{tvp}}$,$r_i(\mathbb{B})=\|H_i(\mathbb{B}*F_i)-H_i^{*}\|^2$,$q_k(\mathbb{B})$通常用均匀分布来描述.一个视频块属于该模型的程度一样可由$\text{Score}^{\text{tvp}}=\sum_i r_i(\mathbb{B})$ 来近似度量.

3　基于生成式模型的视频基元追踪学习

3.1　基于信息投影的视频基元概率建模

给定一组采样于视频基元 Ω^k 的视频块样本集,其概率模型式(3)由如下两步迭代计算得到:

第 1 步.　学习模型参数 λ_+.

对于新引入的特征 B_+ 或 F_+,其对应的参数 λ_+ 由最小化 $KL(f_k\|p_k)$计算得到.为了便于描述,我们把当前模型记为 p,新的增量模型记为 p_+,真实概率模型为 f,如图 2 所示,则最优化增量模型为

$$p_+^{*}=\arg\min_{p_+}KL(f\,\|\,p_+) \tag{4}$$

$$\text{s.t. }E_{p_+}[r_+]=E_f[r_+]=\bar{r}_+=\frac{1}{|\Omega^k|}\sum_{i=1}^{|\Omega^k|}r_+(\mathbb{B}_i) \tag{5}$$

其中,r_+ 是 $r_+(\mathbb{B})$的缩写,表示新特征 B_+ 或 F_+ 在$\mathbb{B}$上的响应.求解上述约束最优化问题,得增量模型:

$$p_+(\mathbb{B})=\frac{1}{Z(\lambda_+)}p(\mathbb{B})\exp\{-\lambda_+ r_+(\mathbb{B})\} \tag{6}$$

其中,λ_+ 取值需使该模型满足式(5)中的约束方程,并且 $Z(\lambda_+)=E_p[\exp\{-\lambda_+ r_+\}]$.

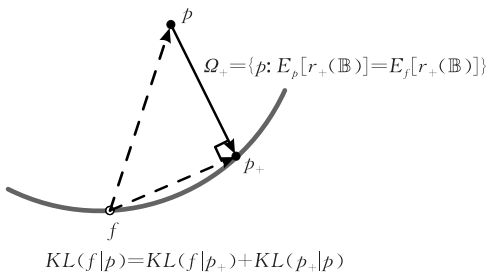


图 2　基于信息投影的特征追踪学习过程示意图

如果保持每次选择的特征 B_+ 或 F_+ 独立于之前选择的所有特征,则有 $E_p[\exp\{-\lambda_+ r_+\}]=$

$E_q[\exp\{-\lambda_+ r_+\}]$，这里 q 表示最初的参考分布，即还没有引入任何特征时的参考分布. 那么，利用式(6)的对数线性性质，可在 λ_+ 与 $E_{p_+}[r_+]$ 之间建立一个查找表，据此计算得到参数 λ_+. 划分函数 $Z(\lambda_+)$ 由相似方式计算.

第 2 步. 选择一个新特征 B_+ 或 F_+.

对于所有候选特征 $\Delta_{\text{svp}}=\{B_i\}$ 或 $\Delta_{\text{tvp}}=\{F_i\}$，它们相应的最佳参数值 λ 由上述步骤计算得到. 如图 2 所示，我们每次选择一个新的特征 B_+ 或 F_+，其取得最大的特征信息增益 $KL(p_+\|p)$：

$$B_+^* \text{ 或 } F_+^* = \arg\max KL(p_+\|p) = \arg\max E_{p_+}\left[\log\frac{p_+}{p}\right] \stackrel{\text{def}}{=} \text{gain}(B_+ \text{ 或 } F_+) \quad (7)$$

由于新特征的响应 r_+ 是个一维约束，并且特征是逐个地选择，新特征的信息增益可以很容易计算：$E_{p_+}\left[\log\frac{p_+}{p}\right]=-\lambda_+ E_{p_+}[r_+]-\log Z(\lambda_+)$. 对于结构视频基元，$\lambda$ 取负值，上述定义的信息增益建议每次选择具有最大响应均值的基函数；对于纹理视频基元，λ 取正值，建议选取具有最小样本方差的特征. 这个学习过程叫做信息投影，每次它将当前模型投影到一个由式(5)定义的新的约束空间来逐步强化模型.

在实验中，由于纹理视频基元通常需要由较多的特征进行约束，为了减少计算代价提高计算效率，需要简化上述纹理视频基元的建模过程. 这里，直接利用滤波器池 Δ_{tvp} 中所有滤波器对视频块进行描述，即形成一个长的类似 SIFT 特征的统计直方图表示(称为隐式表示). 这样每个视频基元可由其包含的所有视频块的特征表示的均值向量描述. 每个视频块到一个视频基元的距离定义为 $r^{\text{tvp}}(\mathbb{B})=\sum_{i=1}^{|\Delta_{\text{tvw}}|}\|H_i(\langle\mathbb{B},F_i\rangle)-H_i^*\|^2=\|H(\mathbb{B})-H^*\|^2$，这里 $H(\mathbb{B})=(H_1^t,\cdots,H_{\Delta_{\text{tvp}}}^t)^t$，$H^*$ 是视频基元的均值向量描述. 那么代入式(3)，纹理视频基元新的概率模型为

$$\begin{aligned} p_k(\mathbb{B}) &= \frac{1}{Z(\lambda)}q(\mathbb{B})\exp\{-\lambda r^{\text{tvp}}(\mathbb{B})\} \\ &= \frac{1}{Z(\lambda)}q(H)\exp\{-\lambda\|-H^*\|^2\} \\ &= p_k(H) \end{aligned} \quad (8)$$

假设 $P_k(H)$ 服从高斯分布，$q(H)$ 服从均匀分布. 一般来说，$P_k(H)$ 是一个多元高斯分布，假设其协方差矩阵的秩很低以至于可将其当作沿最大主轴变化的一元高斯分布，方差为 $\sigma^2=1/|\Omega_{\text{tvp}}^k|\sum_{i=1}^{|\Omega_{\text{tvp}}^k|}\|H(\mathbb{B}_i)-H^*\|^2$. 在此情况下，模型参数 $\lambda=1/2\sigma^2$，$Z(\lambda)=\sqrt{2\pi\sigma^2}$. 实验证明该简化近似是合理且计算有效的.

3.2 视频基元追踪学习

若从大量自然场景视频中提取数以百万计的视频块，则在该视频块空间中存在很多各种类型的结构视频基元和纹理视频基元. 在整个样本集上，采用一个类期望最大化(EM-like)算法分别学习得到一组结构视频基元 $\Omega_{\text{svp}}=\{\Omega_{\text{svp}}^1,\Omega_{\text{svp}}^2,\cdots,\Omega_{\text{svp}}^M\}$ 和纹理视频基元 $\Omega_{\text{tvp}}=\{\Omega_{\text{tvp}}^1,\Omega_{\text{tvp}}^2,\cdots,\Omega_{\text{tvp}}^N\}$. 相应的学习过程分别叫作显式建模和隐式建模，由一个层次 k-means 聚类策略进行初始化. Ω_{tvp} 和 Ω_{svp} 作为视频基元的两个候选集合，两者之间存在严重的空间重叠. 我们希望追踪学习一个视频基元集 $\Omega=\{\Omega_1,\Omega_2,\cdots,\Omega_K\}$，其包含的各种视频基元覆盖整个视频块空间并且基元之间几乎没有重叠，这个学习过程称作混合建模.

给定一组自然场景视频块 $\Omega_{\text{nat}}=\{\{\Omega_{\text{svp}}^k\}_{k=1}^{K_{\text{svp}}},\{\Omega_{\text{tvp}}^k\}_{k=1}^{K_{\text{tvp}}},\Omega_0\}=\{\Omega_0,\Omega_1,\cdots,\Omega_K\}$，其中 $K_{\text{svp}}+K_{\text{tvp}}=K$，$\Omega_0$ 是余下未被解释的均匀分布. Ω_{nat} 的空间分布可由一个混合概率模型近似描述：

$$p(\mathbb{B}|\Lambda,\Delta)=\sum_{k=0}^{K}w_k p_k(\mathbb{B}|\Lambda;\Delta) \quad (9)$$

其中，$p_k(\mathbb{B}|\Lambda,\Delta)$ 是式(3)中定义的第 k 个基元的概率模型，w_k 是相应模型的权重且 $\sum_{k=0}^{K}w_k=1$，Λ 是混合模型参数集合，$\Delta=\{\Delta_{\text{svp}},\Delta_{\text{tvp}}\}$ 是候选滤波器池.

假设 $f(\mathbb{B})$ 是自然视频块 Ω_{nat} 的内在真实分布，希望学习概率模型 $p(\mathbb{B})$ 使其尽可能逼近 $f(\mathbb{B})$，通过最小化 f 与 p 之间的 KL 离散度 $KL(f\|p)=E_f\left[\log\frac{f(\mathbb{B})}{p(\mathbb{B})}\right]=-E_f[\log p(\mathbb{B})]+C$ 来实现，其中 C 表示一个常数. 该最小化问题等价于最大化 $E_f[\log p(\mathbb{B})]$，其可进一步分解为

$$\arg\max_p E_f[\log p(\mathbb{B})]=\arg\max_p E_f\left[\log\sum_{k=0}^{K}w_k p_k(\mathbb{B})\right] \quad (10)$$

$$\geq \arg\max_p\sum_{k=0}^{K}w_k E_f[\log p_k(\mathbb{B})]$$

$$\approx \arg\max_p\sum_{k=0}^{K}w_k E_{\Omega_k}[\log p_k(\mathbb{B})] \quad (11)$$

其中 w_k 由每个子模型的频率近似，不等式关系利用了对数函数的凸性，近似关系成立是由于不同子空

间之间的重叠可以近似忽略不计，$E_{\Omega_k}[\log p_k(\mathbb{B})]=E_{p_k}\left[\log\frac{p_k(\mathbb{B})}{q(\mathbb{B})}+\log q(\mathbb{B})\right]=E_{p_k}\left[\log\frac{p_k(\mathbb{B})}{q(\mathbb{B})}\right]+C$.

为了避免在最小化 $KL(f\|p)$（等价于最大化似然估计）时产生过拟合，在式(11)中引入一个模型复杂度约束项 $-\sum_k \alpha w_k \log w_k$，其中 α 是一个正参数（如 $\alpha=0.1$）. 那么混合模型中任意一个结构视频基元或纹理视频基元的模型信息增益为

$$l_k = w_k E_{p_k}\left[\log\frac{p_k(\mathbb{B})}{q(\mathbb{B})}\right]-\alpha w_k \log w_k \quad (12)$$

其中，$E_{p_k}\left[\log\frac{p_k(\mathbb{B})}{q(\mathbb{B})}\right]=\sum_{i=1}^{n}-\lambda_i E_{p_k}[r_i]-\log Z(\lambda_i)$. 当基元数目很少，且某个基元频率很高时，引入的约束项 $-\alpha w_k \log w_k$ 是一个关于 w_k 的单调减函数，因此鼓励分裂该视频基元为多个. 相反，当基元数很多，单个基元频率很低时，该约束项变成单调增函数而鼓励合并相似的较小基元.

依据式(12)定义的模型信息增益，我们设计了视频基元追踪学习算法（混合建模），见算法 1. 该算法本质上对 Ω_{svp} 与 Ω_{tvp} 中的所有视频基元依据它们的信息增益进行排序，然后选择具有最大增益的模型加入结果模型队列，并移除所有与该模型有重叠的模型中的重叠部分，重新计算有变化的类的模型，最终追踪学习得到几乎没有空间重叠且覆盖整个空间的视频基元集. 从信息增益的定义，该建模过程每次选择具有最大频率且具有最小熵的子模型，同时满足模型复杂度约束. 视频基元集可用于各种视频编码或建模任务. 例如，用它进行视频向量化表示，用于包含复杂表观与运动变化的视频场景的识别任务等. 需要强调的是，本文的视频基元集学习框架是个无监督的学习过程且可用于较大规模的视频数据集.

4　实　验

为了验证本文算法，我们设计了 3 组实验：(1) 动态纹理分割. 该实验显示了本文提出的隐式表示具有较强表达能力，同时说明在 3.1 节对纹理视频基元建模进行的简化与高斯假设是合理的. (2) 一般视频基元集学习. 利用提出的混合建模方法在一组自然场景视频上学习得到一个视频基元集. 通过对基元集的深入分析，我们对自然视频块的空间分布有了一个全局的认识. 对一组测试视频的空间分布降维可视化，发现基于基元集的视频表示很好地捕获了视频的重要本质特征信息. (3) 视频场景识别. 受(2)中有效的视频表示启发，我们设计了一个视频场景识别的实验，定量的比较实验显示了本文方法的优越性.

算法 1.　视频基元追踪学习.

输入：$\Omega_{svp}=\{\Omega^1_{svp},\Omega^2_{svp},\cdots,\Omega^M_{svp}\}$ 与 $\Omega_{tvp}=\{\Omega^1_{tvp},\Omega^2_{tvp},\cdots,\Omega^N_{tvp}\}$

输出：$\Omega=\{\Omega_1,\Omega_2,\cdots,\Omega_K\}$

1. 初始化 $\Omega=\varnothing$，$K\leftarrow 0$；
2. Repeat
3. 选择具有最大信息增益 g_{max} 的 $\Omega^k\in\{\Omega_{svp},\Omega_{tvp}\}$（由式(12)计算）；
4. For 每一个 $\Omega^i\in\{\Omega_{svp},\Omega_{tvp}\}$ do
5. 从 Ω^i 中移除样本 $\mathbb{B}_j$，如果 $\mathbb{B}_j\in\Omega^k$；
6. 如 Ω^i 中样本改变，分别按两种基元定义重新计算其概率模型，选择具有较大信息增益的建模方式；
7. End
8. Until $g_{max}<t$ 或 $\Omega_{svp}\cup\Omega_{tvp}=\varnothing$.

4.1　动态纹理分割

首先验证本文提出的视频隐式表示方法在动态纹理分割任务中的效果. 视频块的隐式表示是一个 64 维 1 范数归一化的滤波器响应统计直方图，每一维对应 Δ_{tvp} 中一个滤波器. 在两段互补的合成视频上与常用的三种视频表示进行比较，即传统的光流、像素滤波响应以及像素灰度值. 像素滤波响应是指直接利用每个像素点的 64 个滤波响应对该像素进行描述. 基于这 4 种表示，采用 Mean-shift 算法[17]进行分割实验，分别有一个带宽参数 h 需要进行调节. 为了定量地比较不同算法的性能，我们采用兰德指数(Rand index)[18]来评价实验分割结果与真实数据的相似度. 兰德指数值 r 取值范围为 $[0,1]$，与两者的像素级匹配程度成正比，值越大表面实验分割得越准确，$r=1$ 说明分割结果与真实数据完全匹配，反之 $r=0$ 表明两者完全不匹配. 如图 3 显示，采用相同分割算法，只有我们的隐式表示成功地对两段视频进行了分割. 图 4 是图 3 中两段测试场景的前景和背景隐式表示直方图：改变动态纹理运动速度和改变动态纹理朝向. 该图说明隐式表示可以区分两种很相似的动态纹理的机理.

Chan 和 Vasconcelos 提出混合动态纹理模型(DTM)[13]与分层动态纹理(LDT)[19]，他们的工作很好地扩展了动态纹理模型应用范围并提高了视频分割的性能. 如图 5 所示，本文基于隐式表示方法与他们的工作在公共数据集[13]上进行了定性与定量比较，其中 GPCA 与 Ising 算法是基准算法[19]，每种方法的分割结果都给出了相应的兰德指数，直观地体现了不同算法特点. DTM 算法需要手工初始

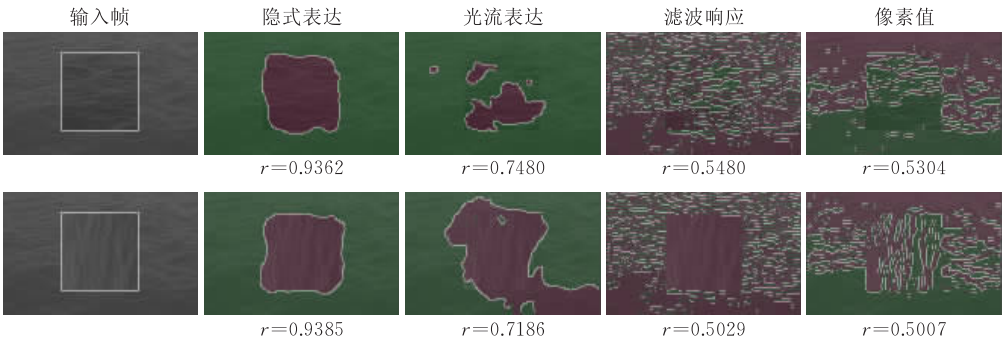
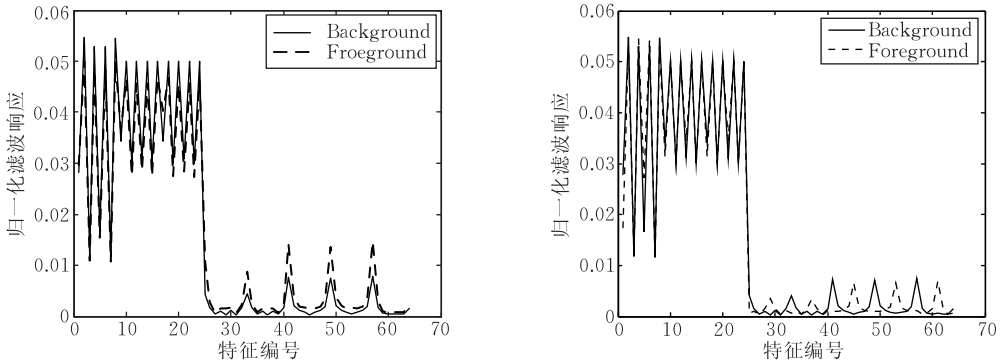


图 3 两段互补的视频场景前景、背景分割(前景表示中间方块区域,背景指余下的部分.第 1 行中前景、背景具有相同的空间表观,但是不同的运动特性(前景区域比背景区域运动加速 2 倍).第 2 行中,前景与背景具有相同的运动特性但是不同的空间朝向(前景区域被旋转 90°).基于隐式表示的方法对两段视频成功分割)



(a) 改变动态纹理运动速度 (b) 改变动态纹理朝向

图 4 图 3 中两段测试场景的前景(虚线)和背景(实线)隐式表示直方图(虽然前景和背景共享很多统计特性,还是被不同的运动速度和运动纹理朝向区分开)

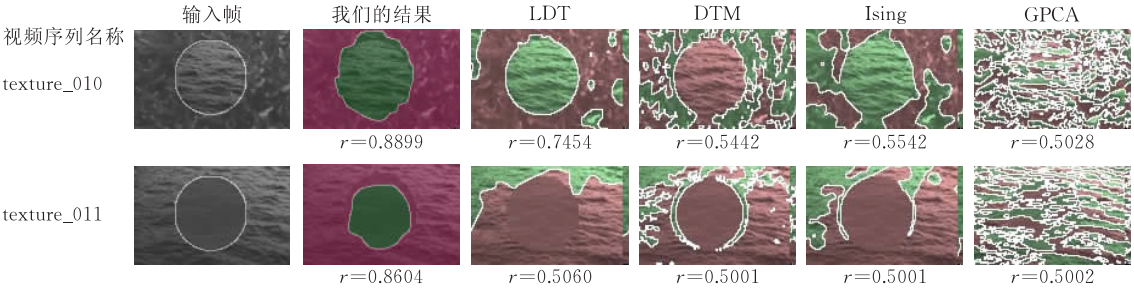


图 5 在公共数据上,基于隐式表示的分割方法与目前主要的动态纹理分割算法比较

化分割轮廓[19],LDT 算法是一个图模型,需要大量的复杂计算,本文分割方法不需要初始化以及复杂的计算.需要说明的是本文方法基于视频块进行分割,很难精确分割区域边界,因此本文方法的分割误差主要出现在前景背景边界区域,但总体上前景背景的分割结果要优于其它方法,这也说明本文的隐式表示更能反应动态纹理的本质特征.

我们也将基于隐式表示的分割应用于两段真实的视频场景.图 6(a)场景包含丰富的具有复杂运动的动态纹理,虽然对其我们很难得到标准的分割结果(ground truth),但使用本文基于隐式表示方法可以得到合理而有意义的实验分割结果,如图 6(b)所示.图 7(a)场景是一个人正通过滑索通过水流湍急的河面,摄像机与人基本保持同步运动.图 7(b)显示基于隐式表示的分割方法成功将人与水面分割开来.作为比较,在该场景上执行本文提出的混合建模方法,结果如图 7(c)所示.整个场景的视频块空间

(a) (b)

图 6 在包含丰富动态纹理的真实视频场景上本文隐式表示方法分割实验结果

被建模为两个纹理视频基元(图 7(d))与 5 个结构视频基元(图 7(e)).混合建模成功捕获了被基于隐式表示方法丢掉的重要结构信息.混合建模的结果从目标分割的角度看是分割过于精细了,但应该更适合于视频内容语义表示,如用于一些高层的视频分析任务(视频场景识别).

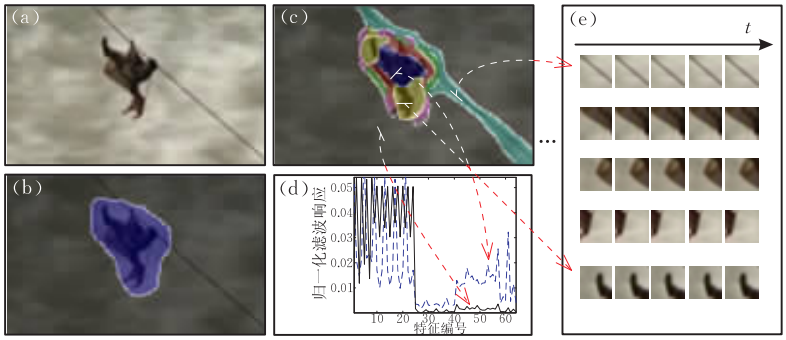


图 7　在一段由动态纹理与结构共同构成的真实视频场景上的建模分割实验结果

4.2　视频基元字典学习

实验中,我们收集 200 段自然视频,视频转换成灰度使用 Xvid 编码器压缩,帧率为 25fps,大小缩放到 310×190,时间至少持续 3s.其中 100 段视频作为训练数据进行视频基元集学习.每段视频截取 60 帧,并从中随机提取 138225 个视频块,总共提取了约 1.4×10^7 个视频块.

通过混合建模,最终得到包含 1983 个视频基元的基元集,如图 8 所示.纹理视频基元主要分布在高信息增益区,而结构视频基元则相反.在中部区域,两种类型基元交替出现.在前 100 个视频基元中,静态平坦区和纹理首先出现,而只有 4 个结构视频基元(平移的地平线以及 3 个平移的简单边),这和我们日常经验一致.对于所有基元,其信息增益逐渐下降,但相应的频率却并不是单调下降,这说明存在聚类与其它类别相比分布更稀疏,即熵更大.如图 9 所示,对 3 段典型的不同复杂度的视频场景进行分析,发现构成它们的主要视频基元类别因场景复杂度的变化而显著变化,复杂视频场景(第 1 列)主要由纹理视频基元构成,而结构视频基元主要出现在简单视频场景(第 3 列).在中间复杂度视频场景中,两种类型视频基元交替出现.不考虑基元的具体细节,如纹理类别、结构类别及运动类型等,前 30 个基元的类别分布即可容易地对三段场景进行区分.

图 8　视频基元集中两种基元的分布(从左至右基元信息增益下降,深色和浅色分别表示纹理和结构视频基元)

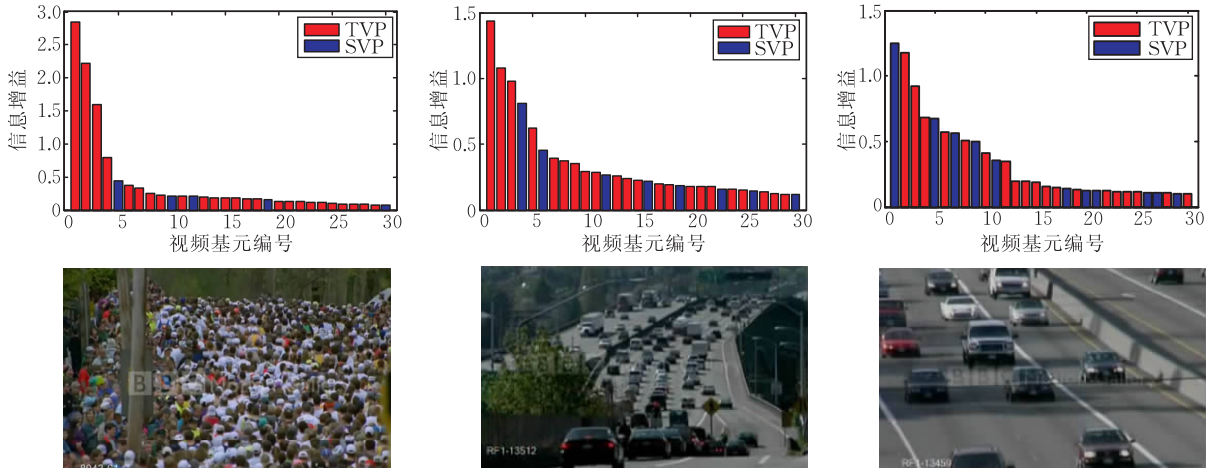


图 9　3 段典型视频场景分析(第 1 行给出 3 段视频中出现的前 30 个视频基元统计信息(纵轴表示信息增益))

为了直观显示所学视频基元集的表示能力,我们设计一个使用 Isomap 算法进行可视化的实验.选择 164 段视频作为测试样本,其中一些视频与训练视频共享一个场景,但取自不同的时间段内.每个测试视频由一个 1983 维的统计向量表达.用 Isomap 算法将 164 段视频投影到一个 2 维空间,如

图 10 所示，每个数据点边上的数值是该场景的编号. 我们发现视频场景的分布与其表观复杂度以及运动类型紧密相关，并呈有规律的分布. 图中标出了一条随场景复杂度变化的分布曲线，并给出相应的 8 个典型场景图示. 第 136 号场景是一个马拉松比赛场景，摄像机镜头固定，大量的运动员呈简单平移运动，这与复杂的随机运动（如场景 52）表现出不同的视觉特性. 上述结果也印证了这样一个事实，即大多数的测试场景是处在中间复杂度，只有少量场景具有极端的较低或较高复杂度.

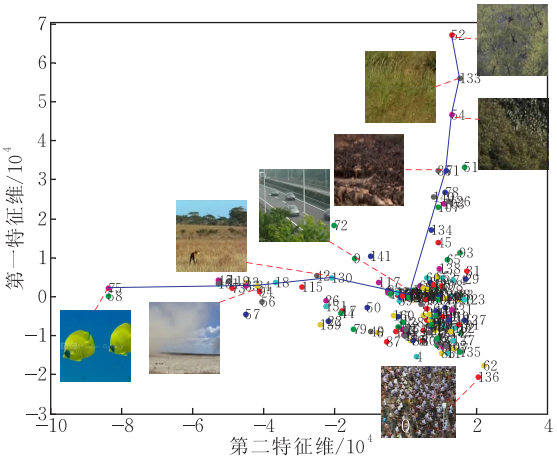


图 10 测试视频空间分布的可视化结果（9 个典型的视频场景展现了视频空间的分布特性，也说明学习得到的视频基元有效地捕获了视频的本质视觉模式）

4.3 视频场景识别

为了定量地研究混合建模得到的视频基元集的表达能力，我们在公共视频场景数据集上执行了视频场景识别任务. Marszałk[20] 等人建立了这个数据库（有 570 段训练视频和 582 段测试视频），并在其上报告算法性能. 他们利用支持向量机对基于视觉词袋（bag-of-words）的视频表示进行分类，使用平均精度（Average Precision，AP），即召回-精度（recall-precision）曲线下的面积，作为分类性能优劣的度量. 他们同时使用静态特征（SIFT）与动态特征（HOG/HOF）对视频进行描述，显示两种特征对视频场景识别都很重要.

利用 Marszałk 等人的评价方法，本文比较了 4 种视频基元学习方法的性能，即 k-means 聚类方法（作为基准方法）、隐式建模方法、显式建模方法和混合建模方法（参见 3.2 节）. k-means 聚类方法使用隐式表示作为视频块的描述. 首先从每段训练视频中随机提取 5000 个视频块，共得约 2.85×10^6 个视频块. K-means 聚类方法采用层次 k-means，得到 2500 个基元. 隐式和显式建模方法分别学习得到 1164 个纹理视频基元和 1813 个结构视频基元. 混合建模共得到 1196 个视频基元，其中有 386 个纹理视频基元和 810 个结构视频基元. 利用不同的视频基元集可以得到不同的视频表示向量. 最后，用 χ^2 核的支持向量机以逐一判别（one-against-rest）的策略进行场景分类识别. 因为 k-means 聚类方法具有随机性，上述基元学习过程我们重复执行 5 次，图 11 与表 1 给出最终的平均识别结果. 由于在 3.1 节对纹理视频基元建模进行简化，隐式建模与 k-means 聚类是很相似的，因此它们总体取得相似的结果. 对于大多数场景，相较显式建模方法，隐式建模方法取得更好的结果，这说明纹理视频基元对视频场景类别识别任务具有更强的判别能力. 而在“INT-Car”与“INT-LivingRoom”两个场景却例外，这可能是由于这两个场景中结构信息更丰富，且提取得到的结构视频基元更具有判别区分能力. 组合了两种不同视频基元的混合建模方法显著提升了识别性能，并且优于 Marszałk 等人的方法. 实验结果进一步证明了结构视频基元与纹理视频基元是两种互补的不同基本视觉模式，对视频表示都具有重要作用. 在我们的方法中两种类型的基元是基于通用的生成式模型进行统一建模，用视频基元追踪学习算法进行组合，而 Marszałk 等人的方法中不同特征的融合是在支持向量机框架下通过多通道高斯核实现.

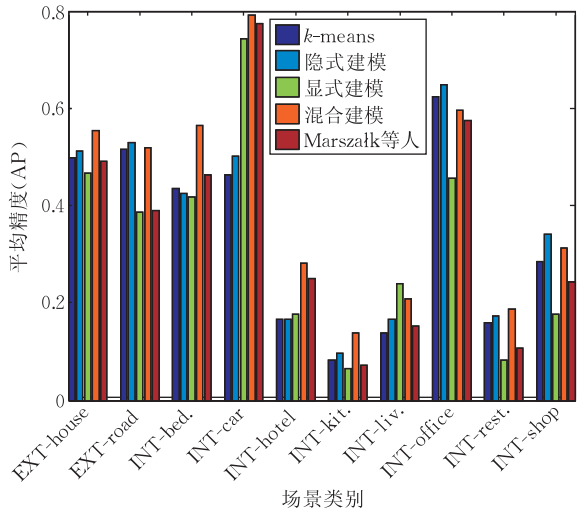


图 11 5 种不同方法在 Hollywood 场景数据集上的视频场景识别结果

表 1 5 种不同方法在所有 10 个场景上的平均识别结果

方法	k-means	隐式建模	显式建模	混合建模	Marszałk 等人的方法
平均 AP	0.336	0.358	0.321	**0.415**	0.351

5 结　论

本文提出了用通用生成式模型对各种不同类型视频块子空间进行统一建模，获得构成视频块空间的各种视频基元的统一概率描述. 这些视频基元的表达能力由它们的信息增益度量. 我们在信息投影框架下对其进行逐个追踪选择，最终建立完整的按信息描述能力排序的视频基元集. 该基元集包含两种互补型的基元，可用于视频的编码、建模或表达. 我们给出了一些基于视频基元表示的应用，与已有方法相比取得了令人鼓舞的结果. 这些视频基元是构成高层视频模式的基本结构，在今后的工作中，我们将研究如何由基本的视频基元复合形成在较大的图像区域和较长的时间序列内更大更复杂的视频运动模式，如人的行为等. 另外，本文在视频基元的分析中没有考虑光照变化影响，没有对光照视频基元进行建模. 这是因为在目前的数据上，光照变化出现频次相对较低，难于单独学习. 我们曾经研究了针对特定光照变化数据类型的视频基元建模[21]，下一步工作将进一步加大数据量，实现一般情况下的光照视频基元建模.

参 考 文 献

[1] Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 1996, 381: 607-609

[2] Zhu S C, Guo C E, Wang Y Z, Xu Z J. What are textons? International Journal of Computer Vision, 2005, 62(1): 121-143

[3] Laptev I. On space-time interest points. International Journal of Computer Vision, 2005, 64(2): 107-123

[4] Dolla'r P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatiotemporal features//Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing, China, 2005, 1: 65-72

[5] Niebles J, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision, 2008, 79(3): 299-318

[6] Olshausen B A. Learning sparse, overcomplete representations of time-varying natural images//Proceedings of the IEEE International Conference on Image Processing. Barcelona, Catalonia, Spain, 2003, 1: 1-4

[7] Black M J, Fleet D J. Probabilistic detection and tracking of motion boundaries. International Journal of Computer Vision, 2000, 38(3): 231-245

[8] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(5): 564-575

[9] Shi K, Zhu S C. Mapping natural image patches by explicit and implicit manifolds//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, Minnesota, USA, 2007, 1: 1-7

[10] Soatto S, Doretto G, Wu Y. Dynamic textures//Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver, Canada, 2001, 2: 439-446

[11] Wang Y, Zhu S C. Modeling textured motion: Particle, wave and sketch//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France, 2003, 1: 213-220

[12] Derpanis K G, Wildes R P. Early spatiotemporal grouping with a distributed oriented energy representation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009, 1: 232-239

[13] Chan A B, Vasconcelos N. Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(5): 909-926

[14] Ravichandran A, Chaudhry R, Vidal R. View-invariant dynamic texture recognition using a bag of dynamical systems//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009, 1: 1651-1657

[15] Greenspan H, Goldberger J, Mayer A. Probabilistic space-time video modeling via piece-wise GMM. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(3): 384-396

[16] Wu Y N, Si Z, Fleming C, Zhu S C. Deformable template as active basis//Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007, 1: 1-8

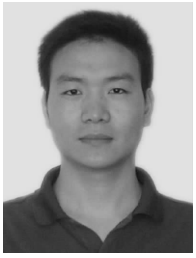
[17] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619

[18] Hubert L, Arabie P. Comparing partitions. Journal of Classification, 1985, 2: 193-218

[19] Chan A B, Vasconcelos N. Layered dynamic textures. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1862-1879

[20] Marszałk M, Laptev I, Schmid C. Actions in context//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009, 1: 2929-2936

[21] Zhao Y D, Gong H, Lin L, Jia Y. Spatio-temporal patches for night background modeling by subspace learning//Proceedings of the 19th International Conference on Pattern Recognition. Tampa, Florida, USA, 2008, 1: 1-4



ZHAO You-Dong, born in 1983, Ph. D.. His research interests include face recognition, background modeling, video content analysis, and computer vision.

GONG Hai-Feng, born in 1978, Ph. D.. His research interests include tracking, video content analysis, and computer vision and pattern recognition.

Jia Yun-De, born in 1962, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision, media computing and human-computer interaction.

Background

There are vast amounts of visual patterns in natural videos ranging from simplicities, such as a static sky or a tree twig swaying in the breeze, to complexities like bubbling water and wild fire. Understanding the structures of the video space is important for video coding, modeling, and even for neural science. Learning a set of video primitives as building blocks for high level models, e. g., action or event is a fundamental problem. In recent literature, video patches have attracted more and more attentions. However, they mainly focus on human action analysis by designing discriminative features or generative events. There lacks a systematic analysis and holistic view for the distribution of these basic motion patterns under a unified modeling framework. In this paper, we study models of small video patches (e. g., $15\times15\times5$), called video bricks, in videos. From a mathematical perspective, these bricks embedded in a high dimensional space, form a variety of clusters of varying dimensions and rich structures. We adopt the notion "atomic video primitives" for these clusters. Our objective is two-fold: (1) Analyzing the space distribution of video bricks and modeling video primitives in two type manners under a common generative model; (2) Studying a primitive pursuit algorithm for selecting video primitives step by step in the brick space and building up a video primitive set for video representation and modeling.

The research of this paper is supported by the Natural Science Foundation of China (No. 90920009): Cognition-based Visual Pattern Representation and Motion Analysis. This project uses complex video intelligent system as application background and target, researches on visual pattern representation based on various levels of cognition and high performance algorithms of motion analysis and recognition or understanding. This project concretely discusses image primitives, video primitives, general image and video pattern representation, and template representation of unified appearance and motion of high level objects. Based on these visual representation models, this project studies practical application problems, such as adaptively tracking, segmentation and recognition of interested objects, in vehicle driving videos. This paper deals with the learning of generic video primitives. As an essential part of the project, both the theory and experimental results will deepen our research and contribute to our projects.