DOI：10.7652/xjtuxb201905019

采用非线性块稀疏字典选择的视频总结

马明阳，梅少辉，万帅
（西北工业大学电子信息学院，710129，西安）

摘要：针对基于稀疏表示的视频总结未充分考虑视频帧之间的非线性关系和关键帧的块稀疏特性，提出了一种采用非线性块稀疏字典选择的视频总结方法。首先考虑视频帧之间的非线性关系，通过核函数把原始视频样本映射到高维空间，使线性不可分的样本变得线性可分，从而实现非线性到线性的转化，建立非线性稀疏字典选择模型；然后考虑关键帧的块稀疏特性，将视频帧分成帧块，每个帧块内的内容具有一定的相似性，进一步建立非线性块稀疏字典选择模型来提取关键帧块；最后设计了一种核化的联合块正交匹配追踪算法对提出的模型进行优化。在基准视频数据集上的实验表明，所提算法能明显提升视频总结的性能指标 F 值，且计算复杂度较低，从而验证了联合使用非线性和块稀疏的有效性。

关键词：视频总结；稀疏表示；非线性；块稀疏；字典选择
中图分类号：TN919.8 **文献标志码**：A **文章编号**：0253-987X(2019)05-0142-07

Nonlinear Block Sparse Dictionary Selection for Video Summarization

MA Mingyang, MEI Shaohui, WAN Shuai
(School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: The current video summarization based on sparse representation does not fully consider the nonlinear relationship between video frames and the block sparsity of key-frames. In this paper, a video summarization scheme with nonlinear block sparse dictionary selection is proposed. The nonlinearity among video frames is considered, and the original video samples are mapped to a very high-dimensional space via a kernel function, which makes the linearly inseparable samples become separable ones to transform the nonlinear cases to linear, thus a nonlinear sparse dictionary selection model is established. The video frames are divided into frame blocks to introduce the block sparsity of key-frames, where the frame contents are similar, and a nonlinear block sparse dictionary selection model is further established to extract key-frame blocks. A kernelized simultaneous block-orthogonal matching pursuit (KSBOMP) method is designed to optimize the proposed model. Experimental analysis for the benchmark video dataset indicates that KSBOMP method can significantly improve the summarization performance in terms of F-score with a low computation complexity, which verifies the effectiveness of simultaneously using nonlinearity and block sparsity.

Keywords: video summarization; sparse representation; nonlinearity; block sparsity; dictionary selection

---

收稿日期：2018-07-28。 作者简介：马明阳(1992—)，男，博士生；梅少辉(通信作者)，男，副教授。 基金项目：国家自然科学基金资助项目(61671383)；中央高校基本科研业务费专项资金资助项目(3102018AX001)。
网络出版时间：2018-12-19 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20181217.1731.003.html

随着互联网、多媒体和智能终端的快速发展，视频的获取和传输变得前所未有的便捷，视频数据呈现爆炸式增长。海量视频数据在给人们带来丰富信息的同时，也对视频管理带来了巨大的挑战，比如检索、浏览和存储等。因此，快速有效地访问视频的主要内容和重要信息已经成为与日俱增的需求。视频总结是目前解决此需求的一种有效途径，它是通过对整个视频结构和内容进行分析来完成，既涵盖了视频的最主要的信息，也极大地减少了数据量[1-2]。

早期的视频总结研究主要有基于聚类的方法和基于镜头边界检测的方法。基于聚类的方法首先采用聚类技术将内容相似的视频帧聚为一类，然后将聚类中心或离聚类中心最近的帧选为关键帧[3-5]。基于镜头边界检测的方法首先对视频进行镜头边界检测，在把视频分割为镜头之后，取镜头中的首帧或末帧作为视频的关键帧。此类算法的最关键步骤是镜头边界检测技术，基本思想是识别视觉内容的不连续性[6-8]。但是这两类方法将视频总结描述为根据距离或者相似度进行数据分类的问题，使得视频的场景语义信息提取仅仅依赖于有限的特征表示[9]。

近年来，稀疏表示理论在模式识别以及计算机视觉领域取得了显著的成果，研究人员也逐渐将稀疏表示理论应用于视频总结，例如：Kumar 等提出了一种基于稀疏表示的方法来提取关键帧，使用随机投影矩阵将视频帧投影到一个低维的随机特征空间，并且在随机特征空间中利用稀疏表示来分析视频数据并生成关键帧[10]；Cong 等将视频总结表述为稀疏字典选择问题，并采用宽松约束条件 $L_{2,1}$ 范数以确保稀疏性，使用 Nesterov 最优梯度算法来提取关键帧[11]；Mei 等将视频总结问题表述为一个最小化稀疏重建问题，使用 L_0 范数对稀疏度约束，采用贪婪算法使关键帧直接被选择为稀疏字典[12]；Cong 等针对网络视频总结提出了一种自适应的稀疏字典选择模型，该算法包含用于选择关键帧的前向步骤和用于移除已经选择的较差的关键帧的后向步骤[13]。

然而，以上算法都是基于线性稀疏表示，即假设视频帧之间是线性关系。这一假设并不总是准确，因为在对视频帧进行特征表示时，特征通常存在内在的非线性属性。另外，由于视频中的关键帧成块出现，这些算法都只考虑了关键帧的稀疏特性，未充分考虑块稀疏特性。Ma 等对视频中的非线性进行了探究，提出了非线性的稀疏表示的视频总结方法[14]。本文在文献[14]的基础上首先建立非线性稀疏字典选择模型，然后利用块稀疏特性进一步扩展建立非线性块稀疏字典选择模型。针对模型的优化，设计了一种核化的联合块正交匹配追踪（kernelized simultaneous block-orthogonal matching pursuit，KSBOMP）算法。基准视频数据集上的实验结果验证了 KSBOMP 算法能明显提高视频总结的客观评价结果。

1　非线性块稀疏字典选择模型

假设一个视频有 n 帧，第 i 帧可以通过特征提取技术表示为在实数集 $\mathbf{R}$ 上的 d 维的特征向量 $\boldsymbol{f}_i$，$\boldsymbol{f}_i\in\mathbf{R}^{\mathrm{d}}$，$1\leqslant i\leqslant n$，该视频可以表示为 $\boldsymbol{F}$，$\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]\in\mathbf{R}^{d\times n}$。视频总结的目标是从原始视频中提取一个包含 k 个关键帧的子集 $\boldsymbol{F}_{\mathrm{key}}$，$\boldsymbol{F}_{\mathrm{key}}=[\boldsymbol{f}_{i_1},\boldsymbol{f}_{i_2},\cdots,\boldsymbol{f}_{i_k}]\in\mathbf{R}^{d\times k}$，其中 $i_1,i_2,\cdots,i_k\in\{1,2,\cdots,n\}$ 为 k 个关键帧的帧序号，要求此关键帧集合包含视频的主要内容，同时关键帧的数量越少越好。

1.1　非线性稀疏字典选择

基于稀疏表示的视频总结方法首先建立一个假设，即视频中的每一帧都可以表示为关键帧集合的线性组合。由于关键帧是视频的子集，所以此假设可以用公式表述为

$$\boldsymbol{f}_i=\boldsymbol{F}\boldsymbol{x}_i,\quad i=1,2,\cdots,n \tag{1}$$

式中：$\boldsymbol{x}_i$ 是稀疏表示系数，$\boldsymbol{x}_i\in\mathbf{R}^n$，$\boldsymbol{x}_i$ 最多有 k 个非零值，每个非零值对应一个关键帧。

然而，在对视频帧进行特征表示时，特征通常拥有内在的非线性属性，所以假设视频帧之间是线性关系并不总是准确的，会降低视频总结的性能。将原始特征通过核函数映射到高维特征空间，再进行稀疏表示，这样在原始空间中不能线性可分的样本在高维空间变得线性可分，而在原始空间中线性可分的样本在高维空间中能够更加准确地线性可分[15]。

假设存在一个映射 $\boldsymbol{\psi}(\cdot)$ 可将原始样本映射到高维空间，即 $\boldsymbol{f}_i\in\mathbf{R}^{\mathrm{d}}\rightarrow\boldsymbol{\psi}(\boldsymbol{f}_i)\in\mathbf{R}^D$，其中 $D\gg d$，可能是无穷大。那么，原始视频通过映射后在高维特征空间下就可以表示为 $\boldsymbol{\Phi}=\boldsymbol{\psi}(\boldsymbol{F})=[\boldsymbol{\psi}(\boldsymbol{f}_1),\cdots,\boldsymbol{\psi}(\boldsymbol{f}_n)]=[\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_n]\in\mathbf{R}^{D\times n}$。在高维空间下，式(1)中的假设可以用公式表述为

$$\boldsymbol{\phi}_i=\boldsymbol{\Phi}\bar{\boldsymbol{x}}_i,\quad i=1,2,\cdots,n \tag{2}$$

式中 $\bar{\boldsymbol{x}}_i$ 是高维空间下的稀疏表示系数，$\bar{\boldsymbol{x}}_i\in\mathbf{R}^n$。

通过式(2)中的假设，视频总结问题可以被表述为一个非线性稀疏字典选择问题，其公式为

$$\boldsymbol{\Phi}=\boldsymbol{\Phi}\widetilde{\boldsymbol{X}},\quad \text{s.t.}\quad \|\widetilde{\boldsymbol{X}}\|_{\mathrm{row},0}\leqslant k \tag{3}$$

式中：$\widetilde{\boldsymbol{X}}$ 是所有视频帧的稀疏表示系数，$\widetilde{\boldsymbol{X}}=[\bar{\boldsymbol{x}}_1,$

$\bar{\boldsymbol{x}}_2,\cdots,\bar{\boldsymbol{x}}_n]\in\mathbf{R}^{n\times n}$；$\|\tilde{\boldsymbol{X}}\|_{\text{row},0}$是$\tilde{\boldsymbol{X}}$的非零行的数量。通过求解式(3)得到稀疏表示系数$\tilde{\boldsymbol{X}}$，其中$\tilde{\boldsymbol{X}}$的非零行所对应的视频帧即为提取的关键帧集合。

1.2 非线性块稀疏字典选择

目前，基于稀疏表示的方法基本只考虑关键帧的稀疏特性，而没有考虑块稀疏特性。块稀疏是指视频的关键帧或非关键帧是成块出现的，这是由小邻域内视频帧的内容相似性所决定的，即如果某一帧可以被看作关键帧，那么其短时邻域的任一帧都可以作为关键帧。基于块稀疏特性，视频可以被表示为帧块的结构形式，公式为

$$\bar{\boldsymbol{\Phi}}=[\underbrace{[\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_{d_1}]}_{\bar{\boldsymbol{\Phi}}[1]},\underbrace{[\boldsymbol{\phi}_{d_1+1},\cdots,\boldsymbol{\phi}_{d_1+d_2}]}_{\bar{\boldsymbol{\Phi}}[2]},\cdots,\underbrace{[\boldsymbol{\phi}_{n-d_m+1},\cdots,\boldsymbol{\phi}_n]}_{\bar{\boldsymbol{\Phi}}[m]}] \tag{4}$$

式中：m是视频中帧块的数量；$\bar{\boldsymbol{\Phi}}[i]$是第$i$个帧块，$1\leqslant i\leqslant m$；$d_i$为第$i$个帧块内包含的视频帧的数量，$\sum_{i=1}^{m}d_i=n$。

相应地，式(2)中的假设以块结构的形式可以表示为

$$\boldsymbol{\phi}_i=\bar{\boldsymbol{\Phi}}\bar{\boldsymbol{x}}_i,\quad i=1,2,\cdots,n \tag{5}$$

式中$\bar{\boldsymbol{x}}_i=[\bar{\boldsymbol{x}}_i^{\mathrm{T}}[1],\bar{\boldsymbol{x}}_i^{\mathrm{T}}[2],\cdots,\bar{\boldsymbol{x}}_i^{\mathrm{T}}[m]]^{\mathrm{T}}$也具有块结构的形式，而且最多含有$k$个非零块，分别对应$k$个关键帧块。

将式(5)写成联合表示的形式，可以得到

$$\bar{\boldsymbol{\Phi}}=\bar{\boldsymbol{\Phi}}\bar{\boldsymbol{X}} \tag{6}$$

式中：$\bar{\boldsymbol{X}}=[\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_2,\cdots,\bar{\boldsymbol{x}}_n]$是所有帧的联合表示系数，可以写成以行为块的形式，即

$$\bar{\boldsymbol{X}}=[\underbrace{[\bar{\boldsymbol{X}}_1^{\mathrm{T}},\cdots,\bar{\boldsymbol{X}}_{d_1}^{\mathrm{T}}]}_{\bar{\boldsymbol{x}}^{\mathrm{T}}[1]},\cdots,\underbrace{[\bar{\boldsymbol{X}}_{n-d_m+1}^{\mathrm{T}},\cdots,\bar{\boldsymbol{X}}_n^{\mathrm{T}}]}_{\bar{\boldsymbol{x}}^{\mathrm{T}}[m]}]^{\mathrm{T}} \tag{7}$$

其中，$\bar{\boldsymbol{X}}_j$是$\bar{\boldsymbol{X}}$的第j行，$\bar{\boldsymbol{X}}[i]$对应$\bar{\boldsymbol{\Phi}}$中第i个帧块，$\bar{\boldsymbol{X}}[i]\in\mathbf{R}^{d_i\times n}$，$1\leqslant i\leqslant m$。

$\bar{\boldsymbol{X}}$的块稀疏度$\|\bar{\boldsymbol{X}}\|_{\mathrm{F},0}$表示含有非零元素的块的数目，其定义公式为

$$\|\bar{\boldsymbol{X}}\|_{\mathrm{F},0}=\sum_{i=1}^{m}I_i \tag{8}$$

式中I_i是关于$\|\bar{\boldsymbol{X}}[i]\|_{\mathrm{F}}$的指示函数，若$\|\bar{\boldsymbol{X}}[i]\|_{\mathrm{F}}>0$，则$I_i=1$，否则$I_i=0$，$\|\cdot\|_{\mathrm{F}}$是Frobenius范数。

通常，在视频总结时，关键帧重建的信息与原始视频的信息允许有一定的误差，而式(6)中的表示是理想情况，因此只需要最小化表示误差，同时保持关键帧块的数量在要求的最大数量之内即可。所以，基于非线性块稀疏字典选择的视频总结模型可以用公式表述为

$$\min\|\bar{\boldsymbol{\Phi}}-\bar{\boldsymbol{\Phi}}\bar{\boldsymbol{X}}\|_{\mathrm{F}}^2,\quad \text{s.t.}\quad \|\bar{\boldsymbol{X}}\|_{\mathrm{F},0}<k \tag{9}$$

式中：第一项Frobenius范数可以看作是测量信息损失的函数；第二项对系数$\bar{\boldsymbol{X}}$的约束用来控制块稀疏度。通过求解式(9)，得到稀疏表示系数$\bar{\boldsymbol{X}}$，其中非零块所对应的视频帧块就是选择的关键帧块。

2 优化算法

2.1 核函数

$\Psi(\cdot)$实现了原始样本空间到高维空间的非线性映射，核空间中的两点$\Psi(\boldsymbol{x})$和$\Psi(\boldsymbol{y})$之间的内积定义为$\langle\Psi(\boldsymbol{x}),\Psi(\boldsymbol{y})\rangle=k_{\mathrm{f}}(\boldsymbol{x},\boldsymbol{y})$，其中$k_{\mathrm{f}}(\boldsymbol{x},\boldsymbol{y})$称为该核空间所对应的核函数。虽然$k_{\mathrm{f}}(\boldsymbol{x},\boldsymbol{y})$定义为内积的形式，但通常不直接在高维空间计算两个映射样本的内积，高维空间中的两个映射样本的内积可以通过低维空间样本的核函数来计算，这就避免了将数据映射到高维特征时的计算花销。

常见的核函数包括线性核、多项式核以及高斯核，其中应用最为广泛的是高斯核函数，若用σ表示高斯核函数的的带宽参数，则其表达式为

$$k_{\mathrm{f}}(\boldsymbol{x},\boldsymbol{y})=\exp(-\|\boldsymbol{x}-\boldsymbol{y}\|_2^2/2\sigma^2) \tag{10}$$

本文实验选用高斯核函数，带宽参数σ需要人工设定，用于控制高维核空间中样本对内积的具体取值，σ设为0.24。通过核函数可以得到核矩阵$\boldsymbol{K}$，$\boldsymbol{K}\in\mathbf{R}^{n\times n}$，其中$K_{i,j}=k_{\mathrm{f}}(\boldsymbol{f}_i,\boldsymbol{f}_j)$。为了利用块特性，将$\boldsymbol{K}$写成块形式，即

$$\boldsymbol{K}=\begin{bmatrix}\boldsymbol{K}_{[1,1]}&\boldsymbol{K}_{[1,2]}&\cdots&\boldsymbol{K}_{[1,m]}\\\boldsymbol{K}_{[2,1]}&\boldsymbol{K}_{[2,2]}&\cdots&\boldsymbol{K}_{[2,m]}\\\vdots&\vdots&&\vdots\\\boldsymbol{K}_{[m,1]}&\boldsymbol{K}_{[m,2]}&\cdots&\boldsymbol{K}_{[m,m]}\end{bmatrix} \tag{11}$$

式中$\boldsymbol{K}_{[p,q]}=\boldsymbol{K}(\sum_{i=1}^{p-1}d_i+1:\sum_{i=1}^{p}d_i,\sum_{i=1}^{q-1}d_i+1:\sum_{i=1}^{q}d_i)$，$1\leqslant p\leqslant m$，$1\leqslant q\leqslant m$，$\boldsymbol{K}(a_1:a_2,a_3:a_4)$表示$\boldsymbol{K}$的$a_1\sim a_2$行、$a_3\sim a_4$列。

2.2 核化的联合块正交匹配追踪算法

为了求解式(9)中的非线性块稀疏字典选择模型，基于正交匹配追踪(OMP)算法，设计了KSBOMP算法。假设在第t次迭代中，关键帧块的索引集为Λ_t，稀疏重建系数是$\bar{\boldsymbol{X}}_t$，重建误差是$\boldsymbol{R}_t$。

在OMP算法的每次迭代中，字典中与重建误差相关性最大的原子被选中。类似地，在KSBOMP算法中，与所有帧的重建误差同时产生最大相关性的帧块被选择为关键帧块。另外，帧块的大小d_i可能不同，帧块越大，帧块和重建误差之间的相关性越

大，所以采用平均相关性来消除帧块尺寸大小的影响，用公式表示为

$$\lambda_t = \arg\max_{1\leqslant i\leqslant m} \frac{\| \overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i]\boldsymbol{R}_{t-1} \|_{\mathrm{F}}^2}{d_i} \tag{12}$$

式中：λ_t 是第 t 次迭代选中的关键帧块的序号；$\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i]\boldsymbol{R}_{t-1}$ 是帧块 $\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i]$ 和重建误差 $\boldsymbol{R}_{t-1}$ 之间的相关性矩阵。

一旦确定了关键帧块序号 λ_t，就可以得到关键帧块索引集合 $\Lambda_t=[\lambda_1,\cdots,\lambda_t]$。$\Lambda_t$ 确定了参与稀疏重建的关键帧块，还需要求解这些关键帧块在重建时的系数 $\overline{\boldsymbol{X}}_t$。重建系数的目的就是要使关键帧块对 $\overline{\boldsymbol{\Phi}}$ 的重建误差最小，因此可以用公式表示为

$$\overline{\boldsymbol{X}}_t = \arg\min_{\boldsymbol{X}_t} \| \overline{\boldsymbol{\Phi}} - \overline{\boldsymbol{\Phi}}[\Lambda_t]\boldsymbol{X}_t \|_{\mathrm{F}}^2 \tag{13}$$

式中 $\overline{\boldsymbol{\Phi}}[\Lambda_t]$ 是索引集 Λ_t 所对应帧块。

为了获得最小误差，可用最小二乘法求解重建系数，求解公式为

$$\overline{\boldsymbol{X}}_t = (\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[\Lambda_t]\overline{\boldsymbol{\Phi}}[\Lambda_t])^{-1}\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[\Lambda_t]\overline{\boldsymbol{\Phi}} = \boldsymbol{K}^{-1}_{[\Lambda_t,\Lambda_t]}\boldsymbol{K}_{[\Lambda_t,:]} \tag{14}$$

式中：$\boldsymbol{K}_{[\Lambda_t,\Lambda_t]}$ 是以 Λ_t 为行和列索引的 $\boldsymbol{K}$ 的子矩阵块；$\boldsymbol{K}_{[\Lambda_t,:]}$ 是以 Λ_t 为行索引的 $\boldsymbol{K}$ 的子矩阵块。

在得到关键帧块和重建系数后，重建误差更新

$$\boldsymbol{R}_t = \overline{\boldsymbol{\Phi}} - \overline{\boldsymbol{\Phi}}[\Lambda_t]\overline{\boldsymbol{X}}_t = \overline{\boldsymbol{\Phi}} - \overline{\boldsymbol{\Phi}}[\Lambda_t]\boldsymbol{K}^{-1}_{[\Lambda_t,\Lambda_t]}\boldsymbol{K}_{[\Lambda_t,:]} \tag{15}$$

值得注意的是重建误差 $\boldsymbol{R}_t$ 并不能直接计算，因为映射函数 $\Psi(\cdot)$ 和映射后的特征 $\overline{\boldsymbol{\Phi}}$ 都是未知的。另外，算法的目的也不是求解重建误差，而是通过迭代得到关键帧块，关键帧块是通过式(12)进行选择的。观察式(12)，帧块 $\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i]$ 和重建误差 $\boldsymbol{R}_{t-1}$ 之间的相关性矩阵 $\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i]\boldsymbol{R}_{t-1}$ 是可计算的，公式为

$$\overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i]\boldsymbol{R}_{t-1} = \overline{\boldsymbol{\Phi}}^{\mathrm{T}}[i](\overline{\boldsymbol{\Phi}} - \overline{\boldsymbol{\Phi}}[\Lambda_t]\boldsymbol{K}^{-1}_{[\Lambda_t,\Lambda_t]}\boldsymbol{K}_{[\Lambda_t,:]}) = \boldsymbol{K}_{[i,:]} - \boldsymbol{K}_{[i,\Lambda_{t-1}]}\boldsymbol{K}^{-1}_{[\Lambda_t,\Lambda_t]}\boldsymbol{K}_{[\Lambda_t,:]} \tag{16}$$

通过上述迭代步骤，直到算法满足停止准则，就可以得到关键帧块。在获取关键帧块后，通过一定的策略 f 从每个块中选取一帧就可以得到关键帧集合。

综上所述，设计的 KSBOMP 算法步骤如下。

输入：视频 $\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]$，关键帧的最大数目 k，视频帧块的大小 $d_i(1\leqslant i\leqslant m)$。

输出：关键帧索引集 Γ。

初始化：核矩阵 $\boldsymbol{K}$，重建误差 $\boldsymbol{R}_0=\overline{\boldsymbol{\Phi}}$，关键帧块索引集 $\Lambda_0=\varnothing$，迭代次数 $t=1$。

步骤：

1. WHILE ($t\leqslant k$) DO；

2. 根据式(12)和(16)，选择与当前重建误差最相关的帧块作为关键帧块，记录索引值 λ_t；

3. 更新关键帧块索引集，$\Lambda_t=\Lambda_{t-1}\cup\lambda_t$；

4. 根据式(14)，计算稀疏重建系数 $\overline{\boldsymbol{X}}_t$；

5. 增加迭代次数 $t=t+1$；

6. END WHILE；

7. 从关键帧块中选择关键帧，即 $\Gamma=f(\Lambda)$，选择策略 f 可以具体设计，本文实验简单地采用关键帧块的最中间帧作为关键帧。

3　实验与讨论

3.1　实验设置

3.1.1　*实验数据集*　采用 VSUMM 数据集[5]，该数据集包含 50 个从 Open Video Project 收集的视频，视频的格式为 MPEG-1，长度在 1～4 min 之间，帧率为 30 帧/s，视频的内容包括纪录片、教育、演讲、历史、风景等多种类别。该数据集还为每个视频提供了 5 个用户手工提取的关键帧集，可以用于与视频总结的关键帧进行比较。在实验中，对原始视频帧进行 5 倍下采样。

3.1.2　*视频帧特征表示*　实验中对视频帧的表示采用一种 360 维的混合特征[11]，包括 252 维的 CENTRIST 特征[16]和 108 维的颜色特征。CENTRIST 特征提取视频帧的结构信息，对每帧图像采用空间金字塔提取 2 层共 6 块图像，每块图像分别提取 CENTRIST 特征，然后采用 PCA 降为 42 维，所以每帧图像的 CENTRIST 特征为 $42\times6=252$ 维。在提取颜色特征时，将图像分成 3×4 个不重叠的小块，并对每块图像的 RGB 三个颜色通道采用颜色矩(均值、方差和斜度)提取颜色特征，所以颜色特征为 $(3\times3)\times(3\times4)=108$ 维。

3.1.3　*评价方式*　VSUMM 数据的评价方式采用客观评价，每个视频都由 1 个自动总结(automatic summaries，AS)算法结果与 5 个用户总结(user summaries，US)结果进行比较，采用 3 个评价指标对比较结果进行评价，具体包括精度 P、召回率 R 和 F 值 F_{score}，具体公式为

$$P = \frac{N_{\mathrm{m}}}{N_{\mathrm{AS}}} \tag{17}$$

$$R = \frac{N_{\mathrm{m}}}{N_{\mathrm{US}}} \tag{18}$$

$$F_{\mathrm{score}} = \frac{2PR}{P+R} \tag{19}$$

式中：N_{m} 是自动总结与用户总结相匹配的关键帧个数；N_{AS} 和 N_{US} 分别是自动总结和用户总结的关

键帧的数量。根据定义可知，精度反映了自动总结选中匹配关键帧的能力，召回率反映了匹配关键帧击中用户总结的能力，F 值是对精度和召回率的平衡，是对视频总结效果进行评价的最重要的整体指标。

在实验中，每个视频的自动总结分别与 5 个用户总结单独地进行比较和评价，然后用 5 个用户总结的评价均值作为该视频的最终评价结果，最后用 50 个视频的评价结果的平均值作为此数据集的最终评价结果。

3.2 实验结果

3.2.1 *实验参数分析* 算法的参数对算法的性能存在影响，因此需要根据领域知识预先设置参数或探究参数的影响。本文所提的 KSBOMP 算法的主要参数包括视频帧块的大小 d_i 和关键帧的数量 k。

(1)视频帧块的大小 d_i。帧块大小统一设置为 13 帧。由于对原始视频进行了 5 倍的下采样，所以在原始视频中的帧块大小是 65 帧，对应的时长大约为 2 s。通常，视频的内容在 2 s 内不会发生明显的变化，块内的帧具有相似性。

(2)关键帧的数量 k。根据视频总结的目的，不难理解选择过少或者过多的关键帧都是不理想的。关键帧过少会导致遗漏部分重要的信息，相反，选择过多的关键帧不仅会提取到无关紧要的信息，还有可能会造成冗余。

随着关键帧数目的变化，KSBOMP 算法性能随关键帧的变化规律如图 1 所示，分析得出：随着关键帧数量的增加，精度先缓慢上升后下降，召回率一直增加，但增加的速度呈下降趋势，F 值先增加后下降；当关键帧数量较少时，随着关键帧的增加，更多的关键帧将会和用户总结中的关键帧相匹配，性能会逐渐提升；随着更多的关键帧被选择，会选择到不能与用户总结相匹配的关键帧，而且多个关键帧可能会匹配到同一个用户总结的关键帧，即出现关键帧冗余情况，性能下降。

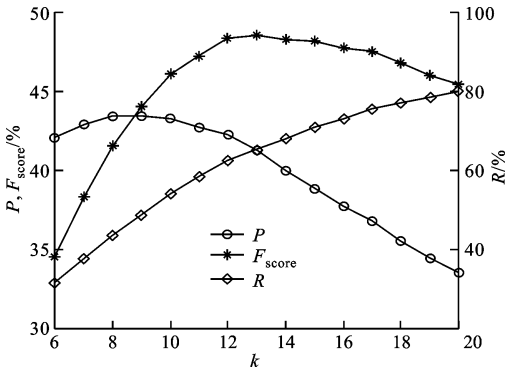

图 1 KSBOMP 算法性能随关键帧数量的变化规律

3.2.2 *性能比较* 将 KSBOMP 算法分别与 DT[3]、STIMO[4]、VSUMM[5]、MSR[12]、SMRS[17]、SOMP[18]、AGDS[13]、NSDS[14] 算法进行比较。DT、STIMO、VSUMM 算法是基于聚类的视频总结算法，它们的结果可以从 VSUMM 算法的官方网站下载，其他的 5 种算法都是基于稀疏表示的算法。表 1 是不同算法的实验结果数据对比，表中加粗数据代表最优数据。

表 1 不同算法的实验结果数据对比

算法	指标			
	P/%	R/%	F_{score}/%	平均关键帧数
DT[3]	35.51	26.71	29.43	6.2
STIMO[4]	34.73	40.03	35.75	10.0
VSUMM[5]	**47.26**	42.34	43.52	7.7
MSR[12]	36.94	57.61	43.39	13.4
SMRS[17]	38.55	56.44	44.32	12.3
SOMP[18]	39.59	57.66	44.93	12.8
AGDS[13]	37.69	64.76	45.65	14.0
NSDS[14]	39.94	63.97	47.16	13.0
KSBOMP	41.29	**65.58**	**48.58**	13.0

从表 1 可以看出，KSBOMP 算法的 F 值在所有对比算法中最高，说明 KSBOMP 算法的总体性能最好。一般来说，选择的关键帧的数目越多，与用户总结相匹配的关键帧也就越多，召回率也就会越高，因此 MSR、SMRS、SOMP、AGDS、NSDS、KSBOMP 算法的召回率高于其他 3 种算法。VSUMM 拥有最高的精度，但是选择的关键帧的数目较少，所以会遗漏部分关键帧，导致召回率和 F 值不高。具体来看，KSBOMP 算法的精度仅低于 VSUMM 算法，召回率在所有算法中最高，说明用户总结中的大部分关键帧被 KSBOMP 算法选中，且选择的不与用户总结相匹配的关键帧也不是很多。综合来看，KSBOMP 算法在精度、召回率和 F 值三个评测标准中都具有优秀的表现。

为了对算法的时间复杂度进行比较，将 KSBOMP 算法和基于稀疏表示的对比算法在相同的计算平台(CPU 型号 Core i7-6700，3.4 GHz，RAM 容量 12 GB)上实验，以 50 个视频的平均运行时间对算法的时间复杂度进行表示，实验结果如表 2 所示，可以看出，KSBOMP 算法的运行时间仅多于

MSR 和 NSDS 算法，但是远远少于其他的算法。结合表 1，说明 KSBOMP 算法以增加少量时间复杂度为代价，获得了性能的显著提升。

表 2　不同算法的平均运行时间

算法	MSR[12]	SMRS[17]	SOMP[18]
时间/s	0.45	5.51	28.30
算法	AGDS[13]	NSDS[14]	KSBOMP
时间/s	77.76	0.97	1.13

3.2.3　*核函数的影响*　为探究不同类型的核函数对 KSBOMP 算法的影响，采用高斯核函数和线性核函数分别进行实验。两种核函数的性能比较如图 2 所示，可以看出：随着关键帧数量的增加，采用高斯核函数的性能明显优于采用线性核函数，即使采用线性核函数，KSBOMP 算法的性能也优于表 1 中的对比算法。

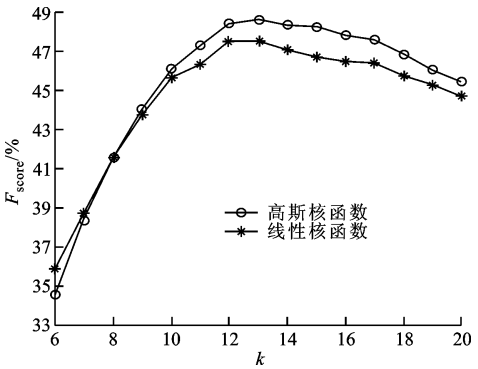

图 2　高斯核函数和线性核函数性能比较

3.2.4　*非线性和块稀疏的有效性*　将 KSBOMP 算法与只考虑非线性和只考虑块稀疏的两种算法进行对比，证明联合考虑两种属性的有效性。具体地，KSBOMP 算法和以下两种算法进行比较：(1)只考虑帧间的非线性属性，未引入视频的块稀疏特性的非线性稀疏字典选择算法[14]，简称为非线性算法；(2)只考虑视频的块稀疏特性，未引入帧间的非线性属性的算法，简称为块稀疏算法。KSBOMP 算法同时考虑非线性和块稀疏特性，三种算法的性能比较结果如图 3 所示，分析得出：不论关键帧的数量是多少，同时考虑非线性和块稀疏特性的 KSBOMP 算法的性能总是优于非线性算法，此结果证明了考虑块稀疏特性的有效性；当关键帧数量较少时，同时考虑非线性和块稀疏特性的 KSBOMP 算法的性能和只考虑块稀疏算法的性能基本一致，但是随着关键帧数量的增加，前者的性能明显优于后者，此结果证明了考虑非线性的有效性。

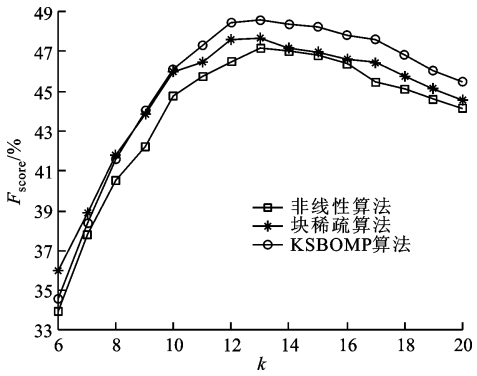

图 3　三种算法的性能比较

4　结　论

本文主要研究基于稀疏表示的视频总结方法，同时考虑视频帧之间的非线性关系和关键帧的块稀疏特性。通过核函数将视频帧映射到高维特征空间，使视频帧之间的关系由非线性转化线性，建立了非线性稀疏字典选择模型用于提取关键帧；在非线性模型的基础上，通过视频帧邻域内的内容相似性，将视频分为帧块，将关键帧的块稀疏特性纳入模型，建立了非线性块稀疏字典选择模型来提取关键帧块。为了优化所提出的模型，本文还设计了 KSBOMP 算法，并在基准视频数据集上与文献[3-5，12-14，17-18]的算法进行了对比实验，结果表明，KSBOMP 算法在性能和效率上都有一定的优势。

目前，KSBOMP 算法还有待改进之处。首先，每个帧块包含相同数量的帧，但实际中固定长度的帧块中的视频帧可能会存在不相似的内容，因此在下一步的工作中，可以考虑结合镜头边界检测技术，根据视频的内容变化来划分帧块。另外，首先确定关键帧块，然后将关键帧块的中间帧选为关键帧，此策略虽比较简单，但可能会造成选择的关键帧在此帧块中并不是最具有代表性的，所以后续考虑更高效的选择策略来进一步提高算法的性能。

参考文献：

[1]　王娟，蒋兴浩，孙锬锋. 视频摘要技术综述 [J]. 中国图象图形学报，2014，19(12)：1685-1695.
WANG Juan，JIANG Xinghao，SUN Tanfeng. Review of video abstraction [J]. Journal of Image and Graphics，2014，19(12)：1685-1695.

[2]　郝雪，彭国华. 基于 SVD 和稀疏子空间聚类的视频摘要 [J]. 计算机辅助设计与图形学学报，2017，29(3)：485-492.
HAO Xue，PENG Guohua. Video summarization

based on SVD and sparse subspace clustering [J]. Journal of Computer-Aided Design and Computer Graphics, 2017, 29(3): 485-492.

[3] MUNDUR P, RAO Y, YESHA Y. Keyframe-based video summarization using Delaunay clustering [J]. International Journal on Digital Libraries, 2006, 6(2): 219-232.

[4] FURINI M, GERACI F, MONTANGERO M, et al. STIMO: STIll and MOving video storyboard for the web scenario [J]. Multimedia Tools and Applications, 2010, 46(1): 47-69.

[5] AVILA S E F D, LOPES A P B, JUZ A D L, et al. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method [J]. Pattern Recognition Letters, 2011, 32(1): 56-68.

[6] 郑慧君, 陈俞强. 基于 MapReduce 的快速视频镜头边界检测算法 [J]. 图学学报, 2017, 38(1): 76-81.
ZHENG Huijun, CHEN Yuqiang. Fast video shot boundary detection algorithm based on MapReduce [J]. Journal of Graphics, 2017, 38(1): 76-81.

[7] 王瑞佳, 牛之贤, 宋春花, 等. 一种改进的基于互信息量的镜头边界检测算法 [J]. 科学技术与工程, 2018, 18(8): 228-236.
WANG Runjia, NIU Zhixian, SONG Chunhua, et al. An improved shot boundary detection algorithm based on mutual information [J]. Science Technology and Engineering, 2018, 18(8): 228-236.

[8] LI J, YAO T, LING Q, et al. Detecting shot boundary with sparse coding for video summarization [J]. Neurocomputing, 2017, 266: 66-78.

[9] MADEMLIS I, TEFAS A, PITAS I. A salient dictionary learning framework for activity video summarization via key-frame extraction [J]. Information Sciences, 2018, 432: 319-331.

[10] KUMAR M, LOUI A C. Key frame extraction from consumer videos using sparse representation [C] // Proceedings of the International Conference on Image Processing. Piscataway, NJ, USA: IEEE, 2011: 2437-2440.

[11] CONG Y, YUAN J, LUO J. Towards scalable summarization of consumer videos via sparse dictionary selection [J]. IEEE Transactions on Multimedia, 2012, 14(1): 66-75.

[12] MEI S, GUAN G, WANG Z, et al. Video summarization via minimum sparse reconstruction [J]. Pattern Recognition, 2015, 48(2): 522-533.

[13] CONG Y, LIU J, SUN G, et al. Adaptive greedy dictionary selection for web media summarization [J]. IEEE Transactions on Image Processing, 2017, 26(1): 185-195.

[14] MA M, MEI S, HOU J, et al. Nonlinear kernel sparse dictionary selection for video summarization [C] // Proceedings of the International Conference on Multimedia and Expo. Piscataway, NJ, USA: IEEE, 2017: 637-642.

[15] GAO S, TSANG I W, CHIA L T. Sparse representation with kernels [J]. IEEE Transactions on Image Processing, 2013, 22(2): 423-434.

[16] WU J, REHG J M. CENTRIST: a visual descriptor for scene categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1489-1501.

[17] VIDAL R. See all by looking at a few: sparse modeling for finding representative objects [C] // Proceedings of the Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2012: 1600-1607.

[18] MEI S, GUAN G, WANG Z, et al. $L_{2,0}$ constrained sparse dictionary selection for video summarization [C] // Proceedings of the International Conference on Multimedia and Expo. Piscataway, NJ, USA: IEEE, 2014: 1-6.

[本刊相关文献链接]

王浩,梁煜,张为. 一种均匀化稀疏表示的图像压缩感知算法. 2019,53(2):136-141. [doi:10.7652/xjtuxb201902018]

吴俊峰,车轩沁. L_1 范数字典约束的感兴趣区域 CT 图像重建算法. 2019,53(2):163-169. [doi:10.7652/xjtuxb201902022]

宋长明,王赟. 融合低秩和稀疏表示的图像超分辨率重建算法. 2018,52(7):18-24. [doi:10.7652/xjtuxb201807003]

刘鑫,张钊强,姚佳文,等. 低秩矩阵和结构化稀疏分解的视频背景差分方法. 2016,50(6):23-29. [doi:10.7652/xjtuxb201606004]

杨艺芳,王宇平. 一种鉴别稀疏局部保持投影的人脸识别算法. 2016,50(6):54-60. [doi:10.7652/xjtuxb201606009]

张鹏,陈湘军,阮雅端,等. 采用稀疏 SIFT 特征的车型识别方法. 2015,49(12):137-143. [doi:10.7652/xjtuxb201512022]

郝雯洁,齐春. 一种鲁棒的稀疏信号重构算法. 2015,49(4):98-103. [doi:10.7652/xjtuxb201504016]

(编辑 陶晴)