基于局部敏感核稀疏表示的视频跟踪

黄宏图*①③ 毕笃彦① 高 山① 查宇飞① 侯志强②
①(空军工程大学航空航天工程学院 西安 710038)
②(空军工程大学信息与导航学院 西安 710077)
③(中国人民解放军 95972 部队 酒泉 735018)

摘 要：为了解决 ℓ_1 范数约束下的稀疏表示判别信息不足的问题，该文提出基于局部敏感核稀疏表示的视频目标跟踪算法。为了提高目标的线性可分性，首先将候选目标的 SIFT 特征通过高斯核函数映射到高维核空间，然后在高维核空间中求解局部敏感约束下的核稀疏表示，将核稀疏表示经过多尺度最大值池化得到候选目标的表示，最后将候选目标的表示代入在线的 SVMs，选择分类器得分最大的候选目标作为目标的跟踪位置。实验结果表明，由于利用了核稀疏表示下数据的局部性信息，使得算法的鲁棒性得到一定程度的提高。

关键词：视频跟踪；核稀疏表示；局部敏感约束；支持向量机
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2016)04-0993-07
DOI: 10.11999/JEIT150785

Visual Tracking via Locality-sensitive Kernel Sparse Representation

HUANG Hongtu①③ BI Duyan① GAO Shan① ZHA Yufei① HOU Zhiqiang②
①(*Aeronautics and Astronautics Engineering College*, *Air Force Engineering University*, *Xi'an* 710038, *China*)
②(*Information and Navigation Institute*, *Air Force Engineering University*, *Xi'an* 710077, *China*)
③(95972 *Troops of PLA*, *Jiuquan* 735018, *China*)

Abstract: In order to solve the problem of lack of discriminability in the ℓ_1-norm constraint sparse representation, visual tracking via locality-sensitive kernel sparse representation is proposed. To improve the linear discriminable power, the candidates' Scale-Invariant Feature Transform (SIFT) is mapped into high dimension kernel space using the Gaussian kernel function. The locality-sensitive kernel sparse representation is acquired in the kernel space. The candidates' representation are obtained after multi-scale maximum pooling. Finally, the candidates' representation is put into the classifier and the candidate with the biggest Support Vector Machines (SVMs) score is recognized as the target. And the experiments demonstrate that the robustness of the proposed algorithm is improved due to the use of the data locality under the kernel sparse representation.

Key words: Visual tracking; Kernel sparse representation; Locality-sensitive constraint; Support Vector Machine (SVM)

1 引言

视频目标跟踪是计算机视觉领域的基础问题之一[1]，广泛应用于视频监控、机器人导航、人机交互和精确制导等领域，是各种后续高级处理，如目标识别、行为分析、视频图像压缩编码和应用理解等高层视频处理和应用的基础。跟踪面临的挑战从内外两个方面来说包括目标内部变化和外界变化，其中目标内部变化包括旋转、尺度变化和形变等，外界变化包括光照变化、遮挡和噪声等。由于目标自身和外界环境变化的复杂性和不可预知性，使得鲁棒实时的视频目标跟踪仍然是亟待解决的问题。

SRC(Sparse Representation-based Classifier)模型已经广泛应用于人脸识别、图像分类、图像去噪、图像分割、超分辨率重建、目标检测和特征提取等计算机视觉领域[2]。得益于 SRC 模型在人脸识别上的成功应用，以及视频本身帧与帧之间存在的冗余性，2009 年 ICCV 上，文献[3]首次将其应用到视频目标跟踪中，后续出现了大量基于稀疏表示的视频目标跟踪算法，并且取得了较好的跟踪性能。文献[4]在稀疏表示模型中引入了微模板系数的 ℓ_2 范数约束，并将加速最近梯度算法引入到模型求解中，提高了算法的鲁棒性和速度，但是由于其模型更新方式导致一旦跟踪失败后续将不可能跟踪上目标。文献[5]将基于稀疏表示的判别式模型和生成式模型结合提出了基于稀疏表示的混合式跟踪算法，在生成式模型中引入了基于重构误差的遮挡检

收稿日期：2015-06-29；改回日期：2015-11-27；网络出版：2016-01-14
*通信作者：黄宏图 huanghongtu@sina.cn
基金项目：国家自然科学基金(61175029, 61379104, 61372167)，国家自然科学基金青年科学基金(61203268, 61202339)
Foundation Items: The National Natural Science Foundation of China (61175029, 61379104, 61372167), The Young Scientists Fund of the National Natural Science Foundation of China (61203268, 61202339)

测。文献[6]将深度学习引入到视频目标跟踪。文献[7]是基于高斯过程回归的迁移学习跟踪算法。其中文献[5]和文献[7]在现有的公开数据库上取得了较好的跟踪效果。目前大多数基于稀疏表示的跟踪算法是基于 ℓ_1 范数约束下的 SRC 模型，然而 SRC 模型存在以下局限性[8]：(1)模型必须是线性的，即各类样本可以用线性子空间建模，同类的样本属于同一子空间；(2)SRC 是通过选择部分训练样本来实现的，需要找到能很好表示各类子空间的字典原子，使得测试样本可用该类的原子有效表示或逼近；(3)约束项中仅含有表示系数的稀疏性先验，没有考虑字典中原子之间的相似性，无法获取数据的局部结构信息，导致获得的稀疏表示判别信息不足；(4)模型算法复杂度高。

文献[9]根据实验结果指出稀疏编码的结果倾向于局部性，即非零系数通常分配给与待编码数据较近的基，稀疏编码是在由待编码数据的最近邻形成的局部坐标系下进行。理论上指出在一些特定的条件下，局部性是比稀疏性更加本质的东西，并且局部性可以通过控制最近邻的数量产生稀疏解，反之稀疏性却不一定能够产生局部性表示。

因此本文针对复杂场景下的视频目标跟踪问题，将 SIFT 特征与核稀疏表示相结合，利用核函数将线性稀疏表示扩展到核空间，在核空间中求解目标基于局部敏感约束的核稀疏表示，使得稀疏表示系数中同时集成了数据的稀疏性和局部性信息，从而增强字典和稀疏表示系数在特征层的类判别能力，实验结果表明提高了基于稀疏表示的判别式跟踪算法的鲁棒性。

2 基于稀疏表示的视觉先验字典学习

大量实验表明，相比直接使用预先指定的字典，使用从训练数据中学习得到的字典将会得到更为紧凑的表示，从而便于后续的压缩编码和分类识别。视觉先验字典学习旨在获取大量同类目标的相似特征信息，所以字典的学习过程需要大量的训练图像。而一般在视频目标跟踪中除了第 1 帧中的目标信息可以利用外，并无其它可利用的有关目标的准确信息。因此如图 1 所示，选用 Caltech101 数据库[10]中的图像进行字典学习。

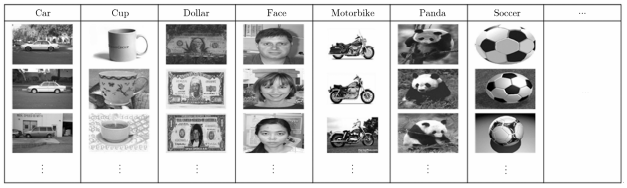


图 1 用于学习视觉先验字典的图像

首先在 101 类目标和 1 类背景的灰度图像上使用固定大小的滑动窗(16×16)，以步长 8 个像素来提取部分重叠图像块的 SIFT 特征 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbb{R}^{m\times n}$，其中 $m=128$ 为 SIFT 特征的维数，$n=50000$ 为提取的 SIFT 特征数量。$\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_k]\in\mathbb{R}^{m\times k}(m\ll k)$ 为要学习得到的字典，字典学习的过程为无监督的离线学习过程，目标函数为[11]

$$[\boldsymbol{D},\boldsymbol{a}_i]=\operatorname{argmin}\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{D}\boldsymbol{a}_i\|_2^2+\beta\sum_{i=1}^{n}\|\boldsymbol{a}_i\|_1,\quad \text{s.t.}\ \ \|\boldsymbol{d}_j\|_2^2\le 1 \tag{1}$$

其中，$j=1,2,\cdots,k$，$\beta=0.15$。当 $\boldsymbol{D}$ 固定时，式(1)为标准的稀疏编码问题，按照文献[12]中的特征-符号(Feature-Sign)搜索算法求解；当 $\boldsymbol{a}_i$ 固定时，式(1)为最小二乘问题，存在解析解，所以可以通过固定一个变量求解另外一个变量的交替优化方法得到字典。字典 $\boldsymbol{D}$ 的初始化通过生成服从高斯分布的随机矩阵完成，迭代次数为 50 次，最终学习得到的字典维数为 $k=1024$。

3 基于局部敏感约束的核稀疏表示

3.1 核稀疏表示

核方法能够捕获非线性特征的相似性，有助于寻找非线性特征的稀疏表示。核函数将样本映射到高维特征空间后可以改变样本的分布，在合适的核函数投影下，数据在高维特征空间将具有更好的线性可分性，样本将可能更准确地由同类的训练样本线性表示，即样本的稀疏表示系数中的非零值将更多地对应于同类训练样本，所以样本的稀疏表示系数中包含更强的判别信息[13]。核稀疏表示本质上是在高维核空间中求解投影特征在投影基下的稀疏表示。给定特征 $\boldsymbol{x}\in\mathbb{R}^m$，$\boldsymbol{x}=\boldsymbol{D}\boldsymbol{v}$。假定由特征投影函数 $\phi:\mathbb{R}^m\to\mathbb{R}^{\mathcal{F}}$ 定义的核 $\kappa(\bullet,\bullet)$，其中 $m\ll\mathcal{F}$。投影函数 $\phi(\bullet)$ 将特征和基投影到高维核空间[13]：

$$\boldsymbol{x}\to\phi(\boldsymbol{x})$$

$$\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_k]\to\boldsymbol{\mathcal{D}}=[\phi(\boldsymbol{d}_1),\phi(\boldsymbol{d}_2),\cdots,\phi(\boldsymbol{d}_k)] \tag{2}$$

然后将投影后的特征和基替换稀疏编码中相应的变量，可以得到核稀疏表示的目标函数：

$$\boldsymbol{v}=\operatorname{argmin}\|\phi(\boldsymbol{x})-\boldsymbol{\mathcal{D}}\boldsymbol{v}\|_2^2+\lambda\|\boldsymbol{v}\|_1,\quad \text{s.t.}\ \ \kappa\left(\boldsymbol{d}_j,\boldsymbol{d}_{j'}\right)\le 1 \tag{3}$$

令 $\boldsymbol{K}_{\boldsymbol{DD}}$ 为 $k\times k$ 的矩阵，其元素为 $\{\boldsymbol{K}_{\boldsymbol{DD}}\}_{jj'}=\kappa(\boldsymbol{d}_j,\boldsymbol{d}_{j'})$，其中 $j=1,2,\cdots,k$，$j'=1,2,\cdots,k$。$\boldsymbol{K}_{\boldsymbol{D}(\boldsymbol{x})}$ 为 k 维向量，$\{\boldsymbol{K}_{\boldsymbol{D}(\boldsymbol{x})}\}_j=\kappa(\boldsymbol{d}_j,\boldsymbol{x})$，核稀疏表示的目标函数式(3)展开为

$$\boldsymbol{v}=\arg\min\kappa(\boldsymbol{x},\boldsymbol{x})+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{K}_{\boldsymbol{DD}}\boldsymbol{v}-2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{K}_{\boldsymbol{D}(\boldsymbol{x})}+\lambda\|\boldsymbol{v}\|_1,\quad \text{s.t.}\ \ \kappa(\boldsymbol{d}_j,\boldsymbol{d}_{j'})\le 1 \tag{4}$$

当选择线性核函数时，$\boldsymbol{K}_{\boldsymbol{DD}}=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}$，$\boldsymbol{K}_{\boldsymbol{D}(\boldsymbol{x})}=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{x}$，此时核稀疏表示退化为标准的稀疏表示，因此在核函数的选择过程中不考虑线性核[13]。常见的非线性核有多项式核、高斯核和直方图交叉核等，这里选择高斯核函数$\kappa(\boldsymbol{d}_j,\boldsymbol{d}_{j'})=\exp(-1/k\,||\,\boldsymbol{d}_j-\boldsymbol{d}_{j'}\,||_2^2)$。

3.2 局部敏感约束的核稀疏表示

给定当前经过核函数映射后的字典$\boldsymbol{\mathcal{D}}=[\phi(\boldsymbol{d}_1),\phi(\boldsymbol{d}_2),\cdots,\phi(\boldsymbol{d}_k)]\in\mathbb{R}^{k\times k}$，在候选目标图像上以相同的方式提取$p$个部分重叠图像块的 SIFT 特征$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_p]\in\mathbb{R}^{m\times p}$，经过核函数投影后的候选目标特征$\boldsymbol{\mathcal{Y}}=[\phi(\boldsymbol{y}_1),\phi(\boldsymbol{y}_2),\cdots,\phi(\boldsymbol{y}_p)]\in\mathbb{R}^{k\times p}$。基于局部敏感约束的核稀疏表示的目标函数为[14]

$$\boldsymbol{C}=\arg\min\sum_{q=1}^{p}\left\|\phi(\boldsymbol{y}_q)-\boldsymbol{\mathcal{D}}\boldsymbol{c}_q\right\|_2^2+\lambda\left\|\boldsymbol{b}_q\odot\boldsymbol{c}_q\right\|_2^2,\quad \text{s.t. }\ \mathbf{1}^{\mathrm{T}}\boldsymbol{c}_q=1 \tag{5}$$

其中$\odot$表示对应元素相乘，$\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_p]$为一个候选目标中$p$个图像块的核稀疏表示，$\lambda=0.03$。根据$\phi(\boldsymbol{y}_q)$和$\boldsymbol{\mathcal{D}}$中基向量之间的相似性，$\boldsymbol{b}_q\in\mathbb{R}^k$赋予基向量不同的自由度：

$$\boldsymbol{b}_q=\exp\left(\frac{\mathrm{dist}(\phi(\boldsymbol{y}_q),\boldsymbol{\mathcal{D}})}{\sigma}\right) \tag{6}$$

其中，$\mathrm{dist}(\phi(\boldsymbol{y}_q),\boldsymbol{\mathcal{D}})$为$\phi(\boldsymbol{y}_q)$和$\boldsymbol{\mathcal{D}}$中原子之间的欧式距离，$\sigma$用来调整局部敏感约束因子权重的衰减速度。

对式(5)中的$\boldsymbol{c}_q$求一阶导数：

$$\begin{aligned}&\frac{d}{dc_q}\left(\left\|\phi(\boldsymbol{y}_q)-\boldsymbol{\mathcal{D}}\boldsymbol{c}_q\right\|_2^2+\lambda\left\|\boldsymbol{b}_q\odot\boldsymbol{c}_q\right\|_2^2\right)\\&=\frac{d}{dc_q}\left(\kappa(\boldsymbol{y}_q,\boldsymbol{y}_q)+\boldsymbol{c}_q^{\mathrm{T}}\boldsymbol{\mathcal{K}}\boldsymbol{c}_q-2\kappa(\bullet,\boldsymbol{y}_q)^{\mathrm{T}}\boldsymbol{c}_q\right.\\&\quad\left.+\lambda\boldsymbol{c}_q^{\mathrm{T}}\mathrm{diag}(\boldsymbol{b}_q)^2\boldsymbol{c}_q\right)\\&=2\left(\boldsymbol{\mathcal{K}}+\lambda\mathrm{diag}(\boldsymbol{b}_q)^2\right)\boldsymbol{c}_q-2\kappa\left(\bullet,\boldsymbol{y}_q\right)^{\mathrm{T}}\end{aligned} \tag{7}$$

其中，$\boldsymbol{\mathcal{K}}=\boldsymbol{\mathcal{D}}^{\mathrm{T}}\boldsymbol{\mathcal{D}}$，$\kappa(\bullet,\boldsymbol{y}_q)=\boldsymbol{\mathcal{D}}^{\mathrm{T}}\phi(\boldsymbol{y}_q)$。令上述一阶导数为0得：$\boldsymbol{c}_q=(\boldsymbol{\mathcal{D}}^{\mathrm{T}}\boldsymbol{\mathcal{D}}+\lambda\mathrm{diag}(\boldsymbol{b}_q)^2)^{-1}\phi(\boldsymbol{y}_q)^{\mathrm{T}}\boldsymbol{\mathcal{D}}$。因此相比$\ell_1$范数约束下的稀疏表示，基于敏感约束下的稀疏表示在获得更加具有判别力表示的同时存在解析解，因此其求解速度比ℓ_1范数约束下的稀疏表示快得多[14]。当字典维数为$\boldsymbol{\mathcal{D}}\in\mathbb{R}^{k\times k}$时，基于$\ell_1$范数约束下的稀疏表示和基于敏感约束下的稀疏表示的算法时间复杂度如表 1 所示，显然基于敏感约束下的稀疏表示的算法复杂度低得多。

表 1 算法时间复杂度比较

	ℓ_1 范数约束	局部敏感约束
时间复杂度	$O(k^{3.5})$	$O(k^2)$

得到目标基于局部敏感约束的核稀疏表示后，沿着图像的不同位置和不同空间尺度对每个单元内的核稀疏表示进行最大值池化，使得池化后的特征对于局部空间转换具有鲁棒性[14]。假定单元区域有ξ个图像块特征，经过最大值池化后，单元区域由k维向量$\boldsymbol{\mu}$表示：

$$\boldsymbol{\mu}=\max\begin{bmatrix}|\boldsymbol{c}_{11}| & |\boldsymbol{c}_{21}| & \cdots & |\boldsymbol{c}_{\xi 1}|\\ |\boldsymbol{c}_{12}| & |\boldsymbol{c}_{22}| & \cdots & |\boldsymbol{c}_{\xi 2}|\\ \vdots & \vdots & \ddots & \vdots\\ |\boldsymbol{c}_{1k}| & |\boldsymbol{c}_{2k}| & \cdots & |\boldsymbol{c}_{\xi k}|\end{bmatrix} \tag{8}$$

其中，$\max(\bullet)$表示依次取矩阵中每行元素的最大值得到k维列向量，$\boldsymbol{c}_\xi$是该单元区域内第ξ个图像块特征的核稀疏表示，$\boldsymbol{c}_{\xi j}$是$\boldsymbol{c}_\xi$的第j个元素。

为了保存空间信息，使用 3 层空间金字塔匹配，将每个候选目标图像分成$1\times1, 2\times2, 4\times4$个子区域，然后对于每个子区域内的核稀疏表示系数使用最大值池化，最后将每层经过最大值池化后的表示系数等权重串联得到目标最终表示。则有候选目标$\boldsymbol{Y}$中所有M个单元区域内的核稀疏表示经过最大值池化后连接起来得到目标的最终表示：

$$\boldsymbol{\rho}=[\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{\mu}_2^{\mathrm{T}},\cdots,\boldsymbol{\mu}_M^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{kM} \tag{9}$$

其中$M=21$。

4 分类器的初始化和更新

4.1 分类器的初始化

在当前帧跟踪位置基础上，在目标周围按照高斯分布提取一定数量的正负样本，其中正样本中心坐标满足$\{\boldsymbol{l}_{\mathrm{pos}}\,|\,\boldsymbol{l}_{\mathrm{pos}}\sim\mathcal{N}(\boldsymbol{l}_t,\delta_1)\}$，负样本中心坐标满足$\{\boldsymbol{l}_{\mathrm{neg}}\,|\,\boldsymbol{l}_{\mathrm{neg}}\sim\mathcal{N}(\boldsymbol{l}_t,\delta_2)\bigcap||\,\boldsymbol{l}_{\mathrm{neg}}-\boldsymbol{l}_t\,||_2^2>\tau\}$，其中$\delta_1$和$\delta_2$为高斯分布的标准差且$\delta_1<\delta_2$，$\boldsymbol{l}_t$为当前帧的跟踪位置。

分别计算正负样本在当前字典下的表示$\boldsymbol{\rho}_r$，将正负样本的表示和其对应的标签$\{\boldsymbol{\rho}_r,l_r=\pm1\}_{r=1}^{R}$代入支持向量机进行训练，基于 SVM 的分类器的目标函数为[15]

$$J(\boldsymbol{\omega})=\arg\min_{\boldsymbol{\omega}}\frac{1}{R}\sum_{r=1}^{R}\ell\left(l_r,\boldsymbol{\omega},\boldsymbol{\rho}_r\right)+\frac{\gamma}{2}\|\boldsymbol{\omega}\|_2^2 \tag{10}$$

其中，$\boldsymbol{\omega}$是分类器参数，$\gamma=1/150$为正则化参数，R为样本个数，其中 50 个正样本，100 个负样本。算法中损失函数$\ell(l_r,\boldsymbol{\omega},\boldsymbol{\rho}_r)$定义为

$$\ell\left(l_r,\boldsymbol{\omega},\boldsymbol{\rho}_r\right)=\left(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\rho}_r^{'}-l_r\right)^2 \tag{11}$$

其中，$\boldsymbol{\rho}_r^{'}=[\boldsymbol{\rho}_r^{\mathrm{T}},1]^{\mathrm{T}}$。

对于每个样本的表示系数$\boldsymbol{\rho}_r$，其分类器响应为

$$h(\boldsymbol{\rho}_r)=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\rho}_r^{'} \tag{12}$$

4.2 基于分类器响应的模型在线更新

考虑到跟踪中目标的变化，模型的在线更新主要是分类器的在线更新，同时为了降低模型更新中由于误差累积导致的漂移，将候选目标在第 1 帧中获得的分类器响应和重新训练得到的分类器响应进行线性加权作为候选目标最终的分类器响应：

$$h(\boldsymbol{\rho}_r)=\eta h_1(\boldsymbol{\rho}_r)+(1-\eta)h_t(\boldsymbol{\rho}_r) \tag{13}$$

其中，$\eta=0.4$，$h_1(\bullet)$ 为第 1 帧中训练得到的初始分类器。由于在高维的特征表示下，SVMs 具有较好的模型泛化能力[16]，因此根据跟踪结果的分类器响应，选择分类器响应大于阈值(0.3)的跟踪结果将其存储下来和初始帧中目标作为正样本，同时将候选目标中分类器响应最小的 5 个存储下来作为负样本，当存储的正样本达到一定数量(10)时，将正负样本带入分类器进行训练得到当前分类器 $h_t(\bullet)$。

5 跟踪算法

算法是在粒子滤波框架[17]下完成。粒子滤波原理实质是用所有已知信息来构造系统状态变量的后验概率密度，即用系统状态转移模型预测状态的先验概率密度，再使用最近的观测值进行修正，得到后验概率密度。这样通过观测数据 $\boldsymbol{I}_{1:t}$ 来递推计算状态 $\boldsymbol{S}_t$ 取不同值时的置信度 $p(\boldsymbol{S}_t\mid\boldsymbol{I}_{1:t})$，由此获得状态的最优估计。给定目标的观察变量集合 $\boldsymbol{I}_{1:t}=\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_t\}$，目标的状态变量 $\boldsymbol{S}_t$ 可以通过最大后验估计得到：

$$\hat{\boldsymbol{S}}_t=\arg\max_{\boldsymbol{S}_t^r} p\left(\boldsymbol{S}_t^r\mid\boldsymbol{I}_{1:t}\right) \tag{14}$$

其中 $\boldsymbol{S}_t^r$ 表示第 t 帧中第 r 个样本的状态。后验概率 $p(\boldsymbol{S}_t\mid\boldsymbol{I}_{1:t})$ 可以由贝叶斯定理递归得到：

$$p\left(\boldsymbol{S}_t\mid\boldsymbol{I}_{1:t}\right)\propto p\left(\boldsymbol{I}_t\mid\boldsymbol{S}_t\right)\int p\left(\boldsymbol{S}_t\mid\boldsymbol{S}_{t-1}\right)\cdot p\left(\boldsymbol{S}_{t-1}\mid\boldsymbol{I}_{1:t-1}\right)\mathrm{d}\boldsymbol{S}_{t-1} \tag{15}$$

其中，$p(\boldsymbol{S}_t\mid\boldsymbol{S}_{t-1})$ 表示动态模型，$p(\boldsymbol{I}_t\mid\boldsymbol{S}_t)$ 表示观察模型。动态模型描述相邻两帧之间目标状态的时间相关性和空间连续性，这里使用 6 参数的仿射变换来模拟相邻两帧之间目标的运动，$\boldsymbol{S}_t=(x_t,y_t,\theta_t,\upsilon_t,\chi_t,\varphi_t)$，分别表示 t 时刻目标 x 方向和 y 方向位移，旋转角度，尺度，宽高比，倾斜角度[17]。状态转换公式为 $p(\boldsymbol{S}_t\mid\boldsymbol{S}_{t-1})=\mathcal{N}(\boldsymbol{S}_t;\boldsymbol{S}_{t-1},\boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}$ 是仿射变换参数的标准差组成的协方差矩阵，这里假设仿射变换参数相互独立且不随时间变化。即以 $\boldsymbol{S}_{t-1}$ 为均值，以 $\boldsymbol{\Sigma}$ 为标准差生成一组满足高斯分布的候选目标。

观察模型 $p(\boldsymbol{I}_t\mid\boldsymbol{S}_t)$ 表示观察变量 $\boldsymbol{I}_t$ 位于状态 $\boldsymbol{S}_t$ 的可能性，算法中构建的观察模型为

$$p(\boldsymbol{I}_t\mid\boldsymbol{S}_t)\propto h(\boldsymbol{\rho}_r) \tag{16}$$

其中，$h(\boldsymbol{\rho}_r)$ 表示候选目标的分类器响应。

因此，本文跟踪算法如表 2 所示。

表 2 基于局部敏感核稀疏表示的视频目标跟踪算法

输入 目标初始位置 $\boldsymbol{S}_1$，仿射变换参数 $\boldsymbol{\Sigma}$。

初始化

(1)从 Caltech101 中提取图像的 SIFT 特征，按照式(1)学习得到字典 $\boldsymbol{D}$，经过高斯核映射后得到核字典 $\boldsymbol{\mathcal{D}}$；

(2)提取正负样本的 SIFT 特征，按照式(5)计算样本的核稀疏表示，经过多尺度最大值池化后得到样本的最终表示 $\boldsymbol{\rho}_r$；

(3)将样本的表示和标签 $\{\boldsymbol{\rho}_r,l_r\}_{r=1}^{R}$ 代入分类器进行训练，得到初始线性分类器 $h_1(\bullet)$；

跟踪

for $t=2$

(1)按照 $p(\boldsymbol{S}_t\mid\boldsymbol{S}_{t-1})=\mathcal{N}(\boldsymbol{S}_t;\boldsymbol{S}_{t-1},\boldsymbol{\Sigma})$ 进行高斯采样，生成候选目标；

(2)提取候选目标的 SIFT 特征，按照式(5)计算候选目标的核稀疏表示，经过多尺度最大值池化后得到候选目标的最终表示；

(3)按照式(13)计算候选目标的分类器响应，选择分类器响应最大的候选目标作为目标的位置 $\boldsymbol{S}_t$。

end

模型更新

if $h(\boldsymbol{\rho}_{r^*})\geq 0.3$

(1)将当前的跟踪结果存储下来和初始帧中的目标作为正样本，同时选择候选目标中分类器响应最小的 5 个作为负样本；

(2)存储的跟踪结果达到 10 个，将正负样本和标签代入分类器进行训练得到新分类器 $h_t(\bullet)$；

end

6 实验结果及分析

6.1 跟踪结果及分析

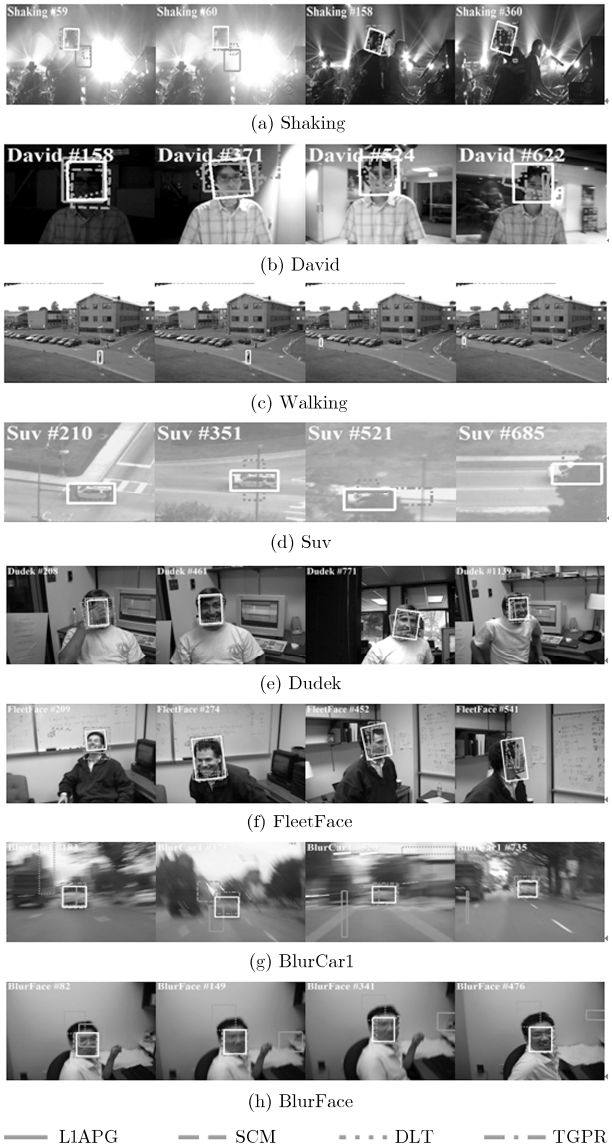
测试视频来自文献[1]，视频数据及目标特征描述如表 3 所示，8 个视频共 5579 帧。实验在 Dual-Core 3.20 GHz，内存 3 GB 的台式计算机上通过 Matlab(R2013a)软件实现。Shaking 仿射变换参数的标准差为[4,4,0.03,0,0,0]，粒子个数为 100 个，David 仿射变换参数的标准差为[5,5,0.01,0.02,0.002,0.001]，粒子个数为 600 个，Walking 仿射变换参数的标准差为[4,5,0.005,0,0,0]，粒子个数为 300 个，Suv 仿射变换参数的标准差为[8,3,0.02,0,0.005,0.001]，粒子个数为 600 个，Dudek 仿射变换参数的标准差为[9, 9, 0.05, 0.05, 0.005, 0.001]，粒子个数为 400 个，FleetFace 仿射变换参数的标准差为[10, 10, 0.06, 0.005, 0.04, 0.001]，粒子个数为 600 个，BlurCar1 仿射变换参数的标准差为[25,15,0.1,0,0.08,0]，粒子个数为 400 个。BlurFace 仿射变换参数的标准差为[30,30,0.01,0,0,0]，粒子个数为 600 个。经过仿射变换后目标区域大小为 32×32。

实验的部分跟踪结果如图 2 所示，其中白色实线为本文算法跟踪结果，其它算法跟踪结果如图例

表 3 视频数据及目标特征描述

视频	描述	分辨率	帧数
Shaking	光照变化，遮挡，旋转	624×352	365
David	光照变化，遮挡，旋转，表情变化，尺度变化	320×240	770
Walking	严重遮挡，尺度变化	768×576	412
Suv	严重遮挡	320×240	945
Dudek	旋转，遮挡，表情变化，尺度变化	720×480	1145
FleetFace	旋转，表情变化，尺度变化	720×480	707
BlurCar1	运动模糊	640×480	742
BlurFace	运动模糊，表情变化	640×480	493

所示。比较算法分别为：基于加速最近梯度的快速 ℓ_1 跟踪算法(L1 tracker using Accelerated Proximal Gradient approach, L1APG)[4]，基于稀疏性的混合式跟踪算法(Sparse Collaborative Model, SCM)[5]，



(a) Shaking　(b) David　(c) Walking　(d) Suv　(e) Dudek　(f) FleetFace　(g) BlurCar1　(h) BlurFace

图 2 部分实验跟踪结果

基于深度学习的跟踪算法(Deep Learning Tracking, DLT)[6]，基于高斯过程回归的迁移学习跟踪算法(Transfer learning with Gaussian Process Regression, TGPR)[7]。由于上述算法都是在粒子滤波框架下利用仿射变换模型完成，因此所有算法均采用相同的初始位置、相同的粒子个数和相同的仿射变换参数标准差，其余参数采用代码中的默认参数。

光照变化 (Shaking#59, #60; David#158, #371)：由于 SIFT 特征对梯度幅值直方图进行了归一化因而能够对光照变化具有一定的不变性，加之目标的表示中同时集成了数据的局部性和稀疏性信息，使得在高维的特征表示下目标和背景更加线性可分。

遮挡(Shaking#360; Walking#87; Suv; Dudek #208)：由于算法提取的局部图像块的 SIFT 特征对于部分遮挡具有一定的鲁棒性，并且在分类器的在线更新中通过设定响应阈值避免了将遮挡物信息引入到模型更新中，所以能够较好地处理跟踪中的遮挡问题。

共面旋转(Dudek#771)：算法每次生成不同旋转角度的候选框，并且 SIFT 特征本身对于共面旋转具有不变性，因此能够解决跟踪中的共面旋转问题。

异面旋转 (Shaking#158; Dudek#1139; FleetFace#274, #452, #541)：对于异面旋转由于目标的视觉特征发生改变，因此主要是通过分类器的在线更新对目标的变化作出自适应响应。

尺度变化(David, Walking, Dudek, FleetFace, BlurCar1)：在粒子滤波框架下按照设定的仿射变换的标准差每次生成不同尺度的候选框，所以能够较好地处理跟踪中目标的尺度变化。

运动模糊(BlurCar1; BlurFace)：图像模糊等效于模糊核与清晰图像的卷积，显然模糊前后目标的 SIFT 特征是不同的，经过高斯核映射后提高了目标的线性可分性，并且目标的稀疏表示中集成了数据的局部性信息，因而能够将模糊后的目标图像与背景分开。

表情变化(David#158; Dudek#461; BlurFace 209; BlurFace#476)：由于人脸的表情变化导致人脸面部的非线性运动，导致目标的 SIFT 特征发生改变。在高维的特征表示下 SVMs 具有较好的泛化能力，所以能够将表情变化后的人脸与背景分开。

6.2 跟踪精度

6.2.1 重叠率　$\mathrm{R_T}$ 是算法标定的跟踪区域，$\mathrm{R_G}$ 是人工标定的真实目标区域，$\mathrm{area}(\mathrm{R_T} \cap \mathrm{R_G})$ 是二者重

叠面积，$\text{area}(\text{R}_\text{T} \cup \text{R}_\text{G})$是二者面积之和。定义重叠率[1]为$o = \text{area}(\text{R}_\text{T} \cap \text{R}_\text{G}) / \text{area}(\text{R}_\text{T} \cup \text{R}_\text{G})$。不同算法下 8 个视频重叠率随帧数变化曲线如图 3 所示。

6.2.2 中心误差 中心误差定义为算法跟踪框的中心与人工标定的真实的中心之间的欧氏距离(像素)[1]，中心误差的统计特征如表 4 所示，其中每个视频对应的第 1 行为中心误差的均值，第 2 行为中心误差的标准差，中心误差的均值表示算法的平均性能，中心误差的标准差表示算法的稳定性，在均值相同的情况下，标准差越小表示算法的稳定性越好。从表 4 中可以看出本文算法整体上优于其它 4 种算法。

6.3 跟踪鲁棒性

如果在一帧中重叠率$o \geq 0.5$，认为跟踪成功[1]。跟踪成功率定义为跟踪成功的帧数与视频总帧数的比值。不同算法下视频的跟踪成功率如表 5 所示，其中最后一行为算法在 8 个视频上的跟踪成功率的平均值。从表 5 中可以看出本文算法的跟踪成功率整体上高于其它 4 种算法。

6.4 算法处理速度比较

算法处理速度比较如表 6 所示。可以看出 Matlab 环境下基于稀疏表示的跟踪算法(L1APG, SCM，本文算法)目前还很难达到实时处理(0.04 s/帧)的要求。由于本文算法需要对每个候选目标提取 9 个局部图像块的 SIFT 特征，所以特征提取过程是影响算法速度的主要因素

表 4 不同算法下各视频中心误差的均值和标准差(像素)

	L1APG	SCM	DLT	TGPR	本文
Shaking	75.1	13.7	37.0	16.8	11.3
	27.5	12.7	23.0	5.4	9.0
David	14.7	24.3	10.8	19.1	12.2
	10.7	15.2	6.5	18.0	7.9
Walking	2.2	2.4	15.1	5.8	4.5
	1.1	1.6	22.9	1.7	2.5
Suv	88.4	80.3	19.9	85.0	28.9
	105.3	101.1	38.5	94.8	60.5
Dudek	17.4	14.5	91.4	13.1	13.3
	11.6	11.9	157.1	10.6	10.1
FleetFace	43.3	19.9	39.7	26.4	19.4
	86.4	15.2	47.2	18.2	17.0
BlurCar1	139.4	16.6	280.4	9.6	6.6
	103.4	35.8	75.9	7.4	9.1
BlurFace	208.0	5.8	129.6	6.0	6.7
	103.6	2.4	8.9	2.8	3.9
平均值	68.0	26.2	80.9	26.4	14.3
	95.3	51.3	118.1	48.8	27.7

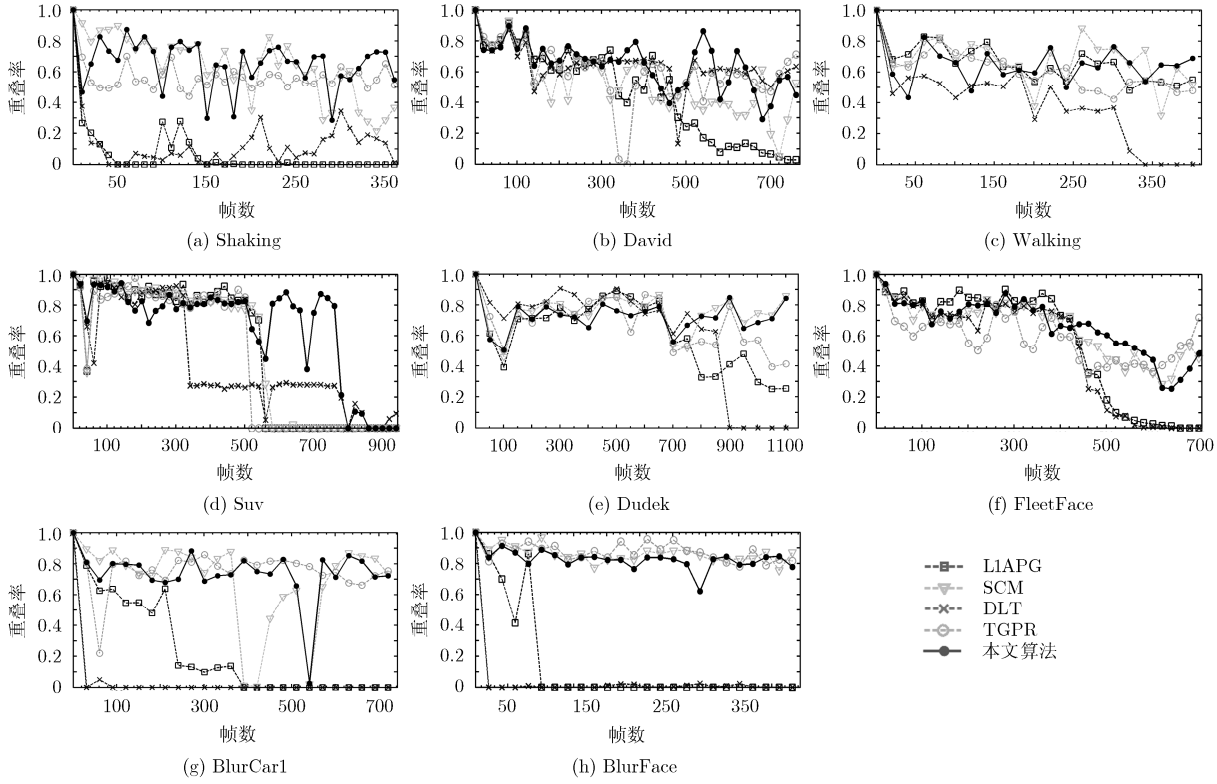


图 3 各视频重叠率随帧数的变化曲线

表 5 不同算法下各视频跟踪成功率(%)

	L1APG	SCM	DLT	TGPR	本文算法
Shaking	1.37	76.99	1.10	78.90	89.04
David	53.12	44.94	92.60	78.57	84.29
Walking	90.29	96.12	34.95	80.10	90.53
Suv	56.51	57.35	35.24	53.33	80.21
Dudek	67.60	94.24	76.33	83.49	96.33
FleetFace	63.65	72.56	62.66	66.76	82.32
BlurCar1	21.56	88.14	2.83	98.11	98.25
BlurFace	15.62	100	0.20	100	99.59
平均值	49.85	77.15	45.40	78.44	89.80

表 6　算法平均处理速度比较(平均处理速度)

算法	L1APG	SCM	DLT	TGPR	本文算法
平均处理速度	0.0835	5.0856	0.3976	3.1638	1.9515

7　结束语

为了克服 ℓ_1 范数约束下的稀疏表示无法获取数据的局部信息，从而导致稀疏表示系数判别力受限的不足，本文提出局部敏感约束的核稀疏表示视频目标跟踪算法。通过在核空间中同时集成数据的稀疏性和局部性信息，从而使获得的稀疏表示系数具有良好判别力，最终使得跟踪中目标和背景更加线性可分。实验结果表明算法的鲁棒性得到了一定程度的提高。

参 考 文 献

[1] WU Yi, LIM J, and YANG Minghsuan. Object tracking Benchmark[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(6): 1442–1456.

[2] WRIGHT J, MA Yi, MAIRAL J, *et al*. Sparse representation for computer vision and pattern recognition[J]. *Proceedings of the IEEE*, 2010, 98(6): 1031–1044.

[3] MEI X and LING H. Robust visual tracking using L_1 minimization[C]. 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2009: 1436–1443.

[4] BAO Chenglong, WU Yi, LING Haibin, *et al*. Real time robust L_1 tracker using accelerated proximal gradient approach[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012: 1830–1837.

[5] ZHONG Wei, LU Huchuan, and YANG Minghsuan. Robust object tracking via sparse collaborative appearance model[J]. *IEEE Transactions on Image Processing*, 2014, 23(5): 2356–2368.

[6] WANG N and YEUNG D Y. Learning a deep compact image representation for visual tracking[C]. Advances in Neural Information Processing Systems, Nevada, 2013: 809–817.

[7] GAO Jin, LING Haibin, HU Weiming, *et al*. Transfer Learning Based Visual Tracking with Gaussian Processes Regression[M]. Computer Vision-ECCV 2014, Zurich: Springer International Publishing, 2014: 188–203.

[8] 王瑞, 杜林峰, 孙督, 等. 复杂场景下结合 SIFT 与核稀疏表示的交通目标分类识别[J]. 电子学报, 2014, 42(11): 2129–2134.
WANG Rui, DU Linfeng, SUN Du, *et al*. Traffic object recognition in complex scenes based on SIFT and kernel sparse representation[J]. *Acta Electronica Sinica*, 2014, 42(11): 2129–2134.

[9] YU K, ZHANG T, and GONG Y. Nonlinear learning using local coordinate coding[C]. Advances in Neural Information Processing Systems. Vancouver, 2009: 2223–2231.

[10] LI Feifei, FERGUS R, and PERONA P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories[J]. *Computer Vision and Image Understanding*, 2007, 106(1): 59–70.

[11] WANG Qing, FENG Chen, YANG Jimei, *et al*. Transferring visual prior for online object tracking[J]. *IEEE Transactions on Image Processing*, 2012, 21(7): 3296–3305.

[12] LEE H, BATTLE A, RAINA R, *et al*. Efficient sparse coding algorithms[C]. Advances in Neural Information Processing Systems, Vancouver, 2006: 801–808.

[13] GAO Shenghua, TSANG I W, and CHIA Liangtien. Sparse representation with kernels[J]. *IEEE Transactions on Image Processing*, 2013, 22(2): 423–434.

[14] WANG Jinjun, YANG Jianchao, YU Kai, *et al*. Locality-constrained linear coding for image classification[C]. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA. 2010: 3360–3367.

[15] CHANG Chihchung and LIN Chihjen. LIBSVM: A library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): 1–27.

[16] SMOLA A J and SCHOLKOPF B. A tutorial on support vector regression[J]. *Statistics and Computing*, 2004, 14(3): 199-222.

[17] ROSS M A, LIM Jongwoo, LIN Ruei-Sung, *et al*. Incremental learning for robust visual tracking[J]. *International Journal of Computer Vision*, 2008, 77(1/3): 125–141.

黄宏图：　男，1986 年生，博士生，研究方向为视频目标跟踪.

毕笃彦：　男，1962 年生，博士，教授，研究方向为图像处理和模式识别.

高　山：　女，1983 年生，博士，讲师，研究方向为图像处理.

查宇飞：　男，1979 年生，博士，副教授，研究方向为计算机视觉和机器学习.