E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中图法分类号：TP391　　**文献标识码**：A　　**文章编号**：1006-8961(2016)12-1622-09

论文引用格式：Dai B, Hou Z Q, Yu W S, Hu D, Fan S Y. Robust visual tracking via fast deep learning[J]. Journal of Image and Graphics, 2016, 21(12):1662-1670.[戴铂, 侯志强, 余旺盛, 胡丹, 范舜奕. 快速深度学习的鲁棒视觉跟踪[J]. 中国图象图形学报, 2016, 21(12):1662-1670.]

[DOI:10.11834/jig.20161211]

快速深度学习的鲁棒视觉跟踪

戴铂, 侯志强, 余旺盛, 胡丹, 范舜奕

空军工程大学信息与导航学院, 西安　710077

摘　要：**目的**　基于深度学习的视觉跟踪算法具有跟踪精度高、适应性强的特点,但是,由于其模型参数多、调参复杂,使得算法的时间复杂度过高。为了提升算法的效率,通过构建新的网络结构、降低模型冗余,提出一种快速深度学习的算法。**方法**　鲁棒特征的提取是视觉跟踪成功的关键。基于深度学习理论,利用海量数据离线训练深度神经网络,分层提取描述图像的特征;针对网络训练时间复杂度高的问题,通过缩小网络规模得以大幅缓解,实现了在GPU驱动下的快速深度学习;在粒子滤波框架下,结合基于支持向量机的打分器的设计,完成对目标的在线跟踪。**结果**　该方法精简了特征提取网络的结构,降低了模型复杂度,与其他基于深度学习的算法相比,具有较高的时效性。系统的跟踪帧率总体保持在22帧/s左右。**结论**　实验结果表明,在目标发生平移、旋转和尺度变化,或存在光照、遮挡和复杂背景干扰时,本文算法能够实现比较稳定和相对快速的目标跟踪。但是,对目标的快速移动和运动模糊的鲁棒性不够高,容易受到相似物体的干扰。

关键词:视觉跟踪;深度学习;支持向量机;粒子滤波;自编码

Robust visual tracking via fast deep learning

Dai Bo, Hou Zhiqiang, Yu Wangsheng, Hu Dan, Fan Shunyi

Institute of Information and Navigation, Air Force Engineering University, Xi'an　710077, China

Abstract: **Objective** Deep learning-based trackers can always achieve high tracking precision and strong adaptability in different scenarios. However, because the number of the parameter is large and finetuning is challenging, the time complexity is high. To improve efficiency, we proposed a tracker based on fast deep learning through construction of a new network with less redundancy. **Method** The feature extractor plays the most important role in a visual tracking system. Based on the theory of deep learning, we proposed a deep neural network to describe essential features of images. Fast deep learning can be achieved by restricting the network size. With the help of GPU (graphics processing unit), the time complexity of the network training is released to a large extent. Under the framework of particle filter, the proposed method combined the deep learning extractor with a support vector machine scoring professor to distinguish the target from the background. **Result** The condensed network structure reduced the complexity of the model. Compared with other deep learning-based trackers, the proposed method can achieve higher efficiency. The frame rate is kept at 22 frames per second on average. **Conclusion** Experiments on an open tracking benchmark demonstrate that both the robustness and timeliness of the proposed tracker is

收稿日期:2016-05-16;**修回日期**:2016-07-21

基金项目:国家自然科学基金项目(61473309);陕西省自然科学基金项目(2015JM6269,2016JM6050)

第一作者简介:戴铂(1992—　),男,中国人民解放军空军工程大学信息与通信工程专业硕士研究生,主要研究方向为计算机视觉与机器学习。E-mail: daybright.daibo@gmail.com

Supported by: National Natural Science Foundation of China(61473309); Natural Science Foundation of Shaanxi Province, China(2015JM6269, 2016JM6050)

promising when the appearance of the target changes contains translation, rotation, and scale, or the interference contains illumination, occlusion, and cluttered background. Unfortunately, the tracker is not robust enough when the target moves fast or the motion blur and some similar objects exist.

Key words: visual tracking; deep learning; support vector machine (SVM); particle filter; autoencoder

0 引　言

基于视觉的目标跟踪是计算机视觉领域中的重要分支，也是物体识别、人机交互、视频监控等应用的关键技术。一个完整的视觉跟踪系统需要选择合适的搜索方法并建立观测模型。

常用的搜索方法有粒子滤波[1]、Mean Shift[2]等。粒子滤波能有效的处理目标跟踪中普遍存在的非线性、非高斯性的问题，它不仅适用于静态视觉平台，而且适用于移动视觉平台，广泛应用于算法研究与工程实践。

可靠的观测模型是提升跟踪鲁棒性的关键，常用的有颜色直方图[3]、纹理[4]、轮廓[5]等，以及后期出现的对这些特征的融合[6-7]。其中，基于高层视觉线索的目标表示模型在处理目标的旋转、尺度变化和背景干扰时效果较好，但当目标发生遮挡时容易产生漂移；基于低层视觉线索的目标表示模型具有较强的抗遮挡能力，但很容易受复杂背景干扰而丢失目标。深度学习[8]能够有效地将高层视觉线索和低层视觉线索相结合，更好地表示数据的特征，同时由于模型的层次、参数足够，对于数字图像处理[9-10]、语音识别[11-12]、目标识别[13]等问题，模型能够在大规模训练数据上取得较好的效果。

近年来，深度学习框架已成功用于处理视觉跟踪问题。Wang[14]应用经过预训练的深度神经网络(DNN)提取目标的深度特征，实现对目标的跟踪；Li[15]应用卷积神经网络(CNN)实现了对目标的在线跟踪。然而，由于深度神经网络的构建和更新过程计算量大、时间复杂度高，跟踪算法的时效性需要加强。为此，在粒子滤波的框架下，通过构建新的网络结构，结合支持向量机(SVM)[16]，提出基于快速深度学习的目标跟踪算法(FDST)。

1 基于深度学习的特征提取

从模式识别的角度理解，目标跟踪是将跟踪目标从背景中分离的模式分类问题，而特征的选择和提取是决定分类效果的关键。本文应用深度学习方法作为特征提取机，通过多层非线性变换描述特征。

1.1 自动编码器

自动编码器(AE)[17]是实现深度学习的重要理论基础，由输入层、隐藏层和输出层构成。设输入数据的向量表示为 $\boldsymbol{x}$，通过以下方式进行数据的编码和译码，即

$$\boldsymbol{h} = sigm(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

$$\hat{\boldsymbol{x}} = sigm(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{h} + \boldsymbol{c}) \tag{2}$$

式中，$sigm(z) = \dfrac{1}{1+\exp(-z)}$ 为 *sigmoid* 函数；与传统的多层神经网络不同，自动编码器使用的权重矩阵 $\boldsymbol{W}$、$\boldsymbol{W}^{\mathrm{T}}$互为转置；$\boldsymbol{b}$ 和 $\boldsymbol{c}$ 分别为编码函数和译码函数的偏置向量。

自动编码器的学习目标是使输出层最大程度的还原输入层的状态，目标函数为

$$\min \sum_{i=1}^{N} \|\boldsymbol{x}^i - \hat{\boldsymbol{x}}^i\|_2^2 + \eta(\|\boldsymbol{W}\|_{\mathrm{F}}^2 + \|\boldsymbol{W}^{\mathrm{T}}\|_{\mathrm{F}}^2) \tag{3}$$

式中，η 用于平衡重构损失部分和权重惩罚部分，$\|\cdot\|_{\mathrm{F}}$ 为弗罗贝尼乌斯范数(Frobenius norm)[18]，定义为

$$\|\boldsymbol{W}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|\boldsymbol{w}_{ij}|^2} \tag{4}$$

1.2 堆栈式消噪自动编码器

为了提升自动编码器表达特征的鲁棒性，消噪自动编码器(DAE)[19]在输入数据中加入噪声，即使用 $\tilde{\boldsymbol{x}}$ 代替 $\boldsymbol{x}$ 作为输入，$\tilde{\boldsymbol{x}}$ 定义为

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} \cdot \boldsymbol{B} \tag{5}$$

式中，矩阵 $\boldsymbol{B}$ 的元素独立采样于一个伯努利分布。训练的过程与自动编码器相同，采用梯度下降算法对加噪前数据进行还原，流程如图 1 所示。

通过逐层贪心算法的预训练，多个 DAE 相堆叠形成一个深度的模型，构成堆栈式消噪自动编码器(SDAE)。

1.3 网络的改进

DLT 算法应用深度神经网络实现了较好的视觉跟踪效果，但是其 15 帧/s 的跟踪速度依然很难满

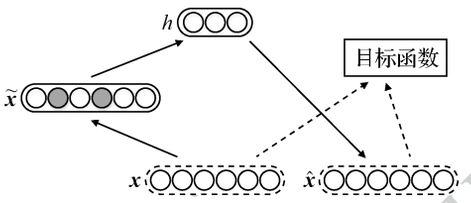


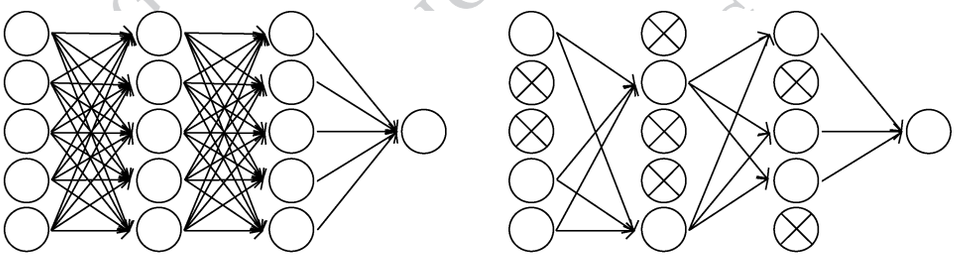
图 1　DAE 工作原理

Fig. 1　Working principle of DAE

足实际需求。结合近年有关网络优化的新思想，对堆栈式消噪自编码器跟踪网络做了部分调整和改进，以更好地适应视觉跟踪场景。

1）在训练一个大型的深度网络时，训练数据的不足会使模型在测试集上的精度降低，出现过拟合。采用比 DLT 跟踪网络更窄的网络架构，通过减少每一层神经元的个数，不仅有效地缓解了过拟合，而且大幅度地提升了网络更新的速度。

2）深度学习理论的重要开拓者 Hinton 提出 dropout 思想[20]，即在每次训练的过程中，使 50% 的特征检测器闲置，通过阻止部分特征的协同，提升网络的泛化能力。图 2 描述了引入 dropout 前后的编码网络架构。



(a) 原始编码网络　　(b) 引入dropout的编码网络

图 2　dropout 思想

Fig. 2　The principle of dropout((a) the original decoding network; (b) the network with dropout)

2　基于 SVM 的打分器设计

本文算法的搜索策略是通过布朗运动理论在空域状态建立粒子滤波。跟踪目标的位置坐标由一个 6 维的仿射变换[21]构成，即

$$\boldsymbol{x}_t = (x_t, y_t, s_t, \theta_t, \alpha_t, \phi_t) \tag{6}$$

式中，x_t 和 y_t 为横、纵坐标的当前偏移量，s_t、θ_t 描述尺度变换，α_t、ϕ_t 分别描述旋转变换和错切变换。跟踪目标的候选区在当前帧的状态参数是服从以上一帧状态为中心的高斯分布，且各参数相互独立。

为了对粒子滤波采样的候选区进行评价，本文设计基于 SVM 的打分器，分数取值 0－1，取分数最高粒子为跟踪结果。打分器通过计算判别函数工作，SVM 的判决函数为

$$f(\boldsymbol{x}) = \mathrm{sgn}\{\boldsymbol{w}^* \boldsymbol{x} + \boldsymbol{b}^*\}$$

式中，$\boldsymbol{x}$ 代表当前输入的待处理特征向量，$\boldsymbol{w}^*$ 和 $\boldsymbol{b}^*$ 为 SVM 的权值向量和偏置向量。通过 Lagrange 变换，得到其对偶表达式

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{l} \boldsymbol{a}_i^* y_i (\boldsymbol{x}_i \boldsymbol{x}) + \boldsymbol{b}^*\}$$

式中，$(\boldsymbol{x}_i, y_i)$ 代表第 i 组训练样本，$\boldsymbol{x}_i$ 为描述样本的特征向量；y_i 为样本对应的标签，在本文的二分类问题中，$y_i \in \{+1, -1\}$，即正样本取 1，负样本取 －1。$\boldsymbol{a}_i^*$ 是在进行 Lagrange 变换时引入的参数，可通过 $\boldsymbol{w}^* = \sum_{i=1}^{l} \boldsymbol{a}_i^* y_i \boldsymbol{x}_i$ 求解。为了实现 SVM 打分功能，将其输出调整为

$$P(C_{+1} \mid \boldsymbol{x}) = \frac{1}{1 + \mathrm{e}^{-g(\boldsymbol{x})}}$$
$$P(C_{-1} \mid \boldsymbol{x}) = \frac{1}{1 + \mathrm{e}^{g(\boldsymbol{x})}} \tag{7}$$

式中，$P(C_{+1} \mid \boldsymbol{x})$ 表示判定为目标的概率，$P(C_{-1} \mid \boldsymbol{x})$ 表示判定为背景的概率，$g(x) = \sum_{i=1}^{l} \boldsymbol{a}_i^* y_i (\boldsymbol{x}_i \boldsymbol{x}) + \boldsymbol{b}^*$，将结果归一化，生成得分

$$S(\boldsymbol{x}) = \frac{P(C_{+1} \mid \boldsymbol{x})}{P(C_{+1} \mid \boldsymbol{x}) + P(C_{-1} \mid \boldsymbol{x})} \tag{8}$$

得分越高表示粒子作为目标的可信度越高。

3　快速深度学习的鲁棒视觉跟踪

结合基于深度学习的特征提取机和基于 SVM 的打分器，提出一种基于快速深度学习的鲁棒视觉

跟踪算法。

算法的观测模型由离线训练和在线跟踪两个阶段组成。其中，离线训练阶段通过训练 cifar-10 数据集[22]建立 SDAE 网络，构建生成式模型，使网络具备描述图像特征的能力；在线跟踪阶段利用生成式模型获得的参数作为搜索起点，通过有效的正、负样本采样调整网络权值，结合 SVM 分类器构建判别式模型，进行特定场景的目标跟踪。跟踪流程如图 3 所示。

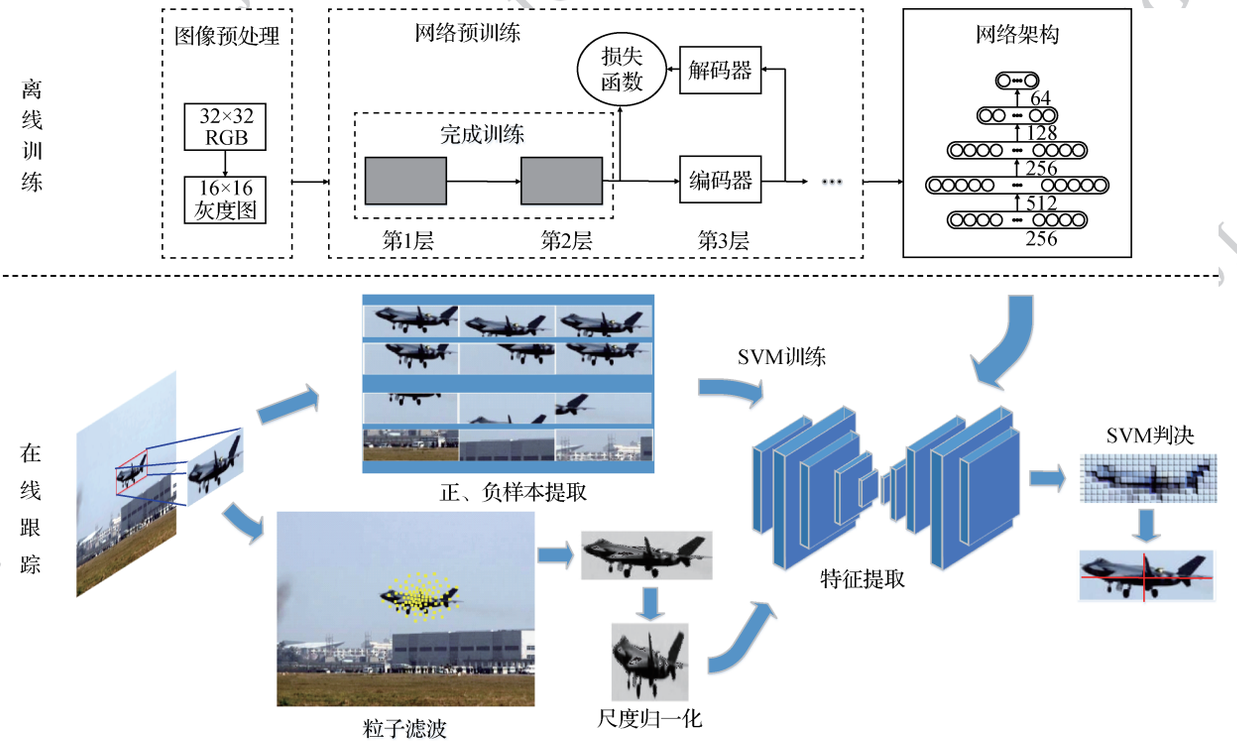


图 3　算法流程

Fig. 3　Pipeline of the algorithm

3.1　离线训练

由于 SDAE 网络层级深、层与层之间的参数多，因此，通过使用海量数据对网络进行预训练，使其对图像特征具备一定的表达力，有利于构建更加鲁棒的视觉跟踪系统。本文采用 cifar-10 数据集进行网络的预训练，该数据集由 60 000 幅32 × 32 像素的 RGB 彩色图像构成，包含 50 000 幅训练图像和 10 000幅测试图像。本文预训练应用给定的训练图像作为无标签数据，通过非监督方式学习特征。

应用图像处理技术对 32 × 32 像素的原始图像进行随机的剪切，处理后获得 16 × 16 像素的灰度图像集，将每幅图像按列展开作为输入，进行网络的预训练。由于多层整体调参会扰乱已完成训练的层级，因此，深度学习的预训练采用自底上升、逐层固定的方式。应用引入 dropout 的网络增加了模型稀疏性的约束，使得学习到的特征更加鲁棒。

3.2　在线跟踪

跟踪过程在粒子滤波[23]的框架下进行，采样粒子对应目标候选区，应用 SVM 为候选区打分，通过有监督学习的模式适时的更新网络，使 SDAE 网络和 SVM 打分器与当前跟踪任务相适应。离线部分应用大规模数据集完成了对 SDAE 编码器的训练；由于打分器与编码器连接的参数相对少，因此，该层的训练对样本数量的要求低，适于在跟踪过程中进行。在线跟踪的具体步骤如下：

1）将输入图像尺寸归一化为 16 × 16 像素，使跟踪系统对尺度变化鲁棒，同时与网络的架构相匹配。

2）第 1 帧时，以跟踪目标的标定位置为中心，提取 10 个正样本、50 个负样本，设置学习速率步长为 0.01，利用前向神经网络的输出残差结合 BP 算法进行批梯度下降计算，设置 10 个样本为一批次，迭代次数为 10。

3）进行跟踪时，应用传统的粒子滤波算法在当前帧选取多个候选区，通过 SDAE 编码器网络对其进行多层非线性变换。

4）结合式（7）（8），将结果输入 SVM 打分器进

行打分，记 S_{max} 为最高得分，认为其对应的跟踪框为跟踪结果。

5）设置阈值 $T=0.9$ 用于评估网络的可信度。当 $S_{max} \geqslant T$ 时，更新正样本，继续下一帧的跟踪；当 $S_{max} < T$ 时，在更新正样本的同时，重新采集负样本，设置迭代次数为5，进行网络的调整，保证目标观测和目标模型之间的相似度，在新的目标模型上进行后续帧的目标跟踪。

4 实验结果与分析

实验在处理器为 Intel Xeon 2.4 GHz 的计算机上进行，采用 MATLAB2014a 编写程序对 FDST 进行测试。实验包括 10 个具有挑战性的图像序列[24]，其内容涵盖了目标的尺度、旋转、形变、遮挡以及光照变化等情况。对比算法为基于学习机制的经典算法 MIL（multiple instances learning）[25]、TLD（tracking-learning-detection）[26]、DLT（deep learning tracker）以及目前适用性较强的 CT（compressive tracking）[27]和 DFT（distribution fields for tracking）[28]。

4.1 定性对比

图4给出部分跟踪结果。

1）平移、旋转和尺度变化。图像序列“dudek”和“david2”测试了算法对目标平移、旋转的适用性，FDST 优于其他对比算法，虽然 MIL 和 TLD 也能够完成跟踪，但其结果的中心位置存在一定程度的偏移；图像序列“singer1”的目标存在较大尺度的变化，由于利用了多尺度图像特征，FDST 和 DLT 都能较好的适应。

2）光照、遮挡和复杂背景干扰。图像序列“cardark”和“singer1”的目标存在较大的光照变化，“faceocc1”和“faceocc2”存在严重的遮挡，“car4”和“cardark”的背景比较复杂。在上述外在干扰情况下进行测试，FDST 能够提供更为稳定和相对精确的结果。

3）目标快速移动和运动模糊。图像序列“mountainbike”、“basketball”和“boy”的目标的快速移动贯穿整个测试，而且，“basketball”存在与目标相似物体的干扰，“boy”存在运动模糊。FDST 能够完成上述序列的跟踪，但是，“basketball”的结果会出现轻微的漂移，“boy”的跟踪没有 DLT 稳定。

图4 跟踪算法性能的定性比较

Fig. 4 Comparison of 6 trackers on 10 video sequences

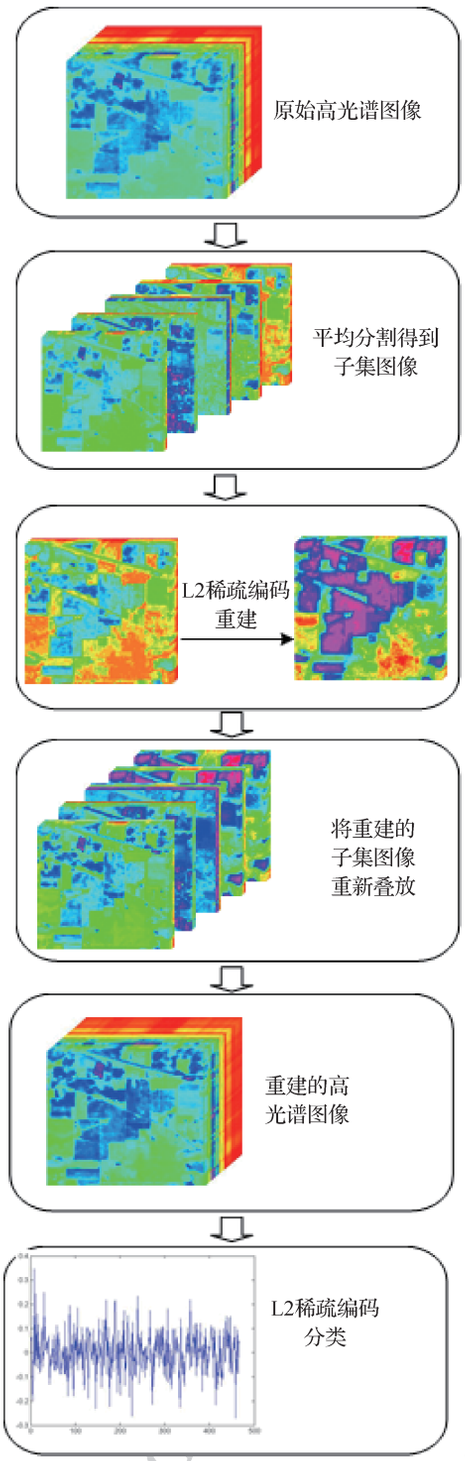


图 5　双重 L2 稀疏编码方法流程图

Fig. 5　The flow chart of double L2 sparse coding

建滑窗内中心点像元，将重建好的 5 个子集图像按原来的波段顺序连接形成新的高光谱数据。在分类阶段，随机选择每类样本的 5%、10%、15%、25%、35%、45%、50% 作为训练样本构造字典 $\boldsymbol{D}$，剩余样本作为测试样本，用 L2 稀疏编码作为分类器，根据重建误差实现高光谱图像分类。为了确定分类器中 λ_{fl} 参数的设置，随机选择每类样本的 15% 作为训练样本，λ_{fl} 分别取 0.01、0.1、1、10、100 进行实验，分类结果如表 2 所示，根据实验结果，本文所有实验中稀疏编码分类器的正则化系数 $\lambda_{fl}=1$。每次分类实验将用不同的训练与测试样本循环执行 20 次，取平均值作为最后总的分类精度。

表 2　不同 λ_{fl} 值的分类精度（15% 的训练样本情况下）

Table 2　The classification accuracy of methods using different λ_{fl} for classification (15% training data)

	λ_{fl} 取值				
	0.01	0.1	1	10	100
分类精度/%	98.90	98.93	99.10	98.88	98.64

3.1　高光谱图像的重建阶段

实验中所用的滑动窗口的尺寸大小不同，重建像元的速度有很大的差异，同时给分类的精度也带来了影响。用不同尺寸滑窗采样对高光谱图像进行重建所用的时间如表 3。

表 3　重建图像所需的时间（不同尺寸滑窗）

Table 3　The time of construction (using sliding windows of different sizes)

/s

	滑窗尺寸 S					
	3	5	7	9	11	13
重建时间	11.26	19.76	36.92	66.44	159.20	295.94

由表 3 可见，当滑窗尺寸每增加一次，重建所需的时间近似增加一倍。针对不同尺寸滑窗重建的图像数据，随机选取每类样本的 15% 作为训练样本时，分别用 L2 稀疏编码，L1 稀疏编码，KNN 和 SVM 算法进行分类，分类精度如图 6 所示。

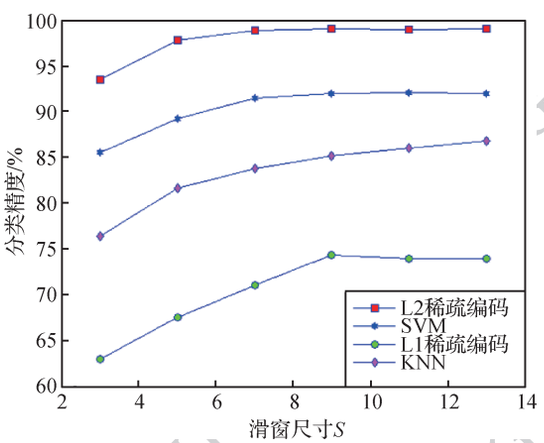


图 6　高光谱图像分类结果（不同尺寸滑窗）

Fig. 6　The classification results of hyperspectral image (using sliding windows of different sizes)

4.2　定量分析

采用中心位置误差、平均覆盖率和成功率作为统计量，对跟踪结果进行定量的分析。中心位置误差是图像中跟踪结果的中心位置与实际的中心位置之间的欧氏距离，结果如图 5 所示；覆盖率是跟踪结果和目标真实值的区域之间重叠部分所占的比率，结果如表 1 所示；成功率指的是在完整的跟踪过程中覆盖率大于阈值 T 的帧数占总帧数的比率，结果如表 2 所示。

分析实验结果，FDST 在上述实验的大部分图像

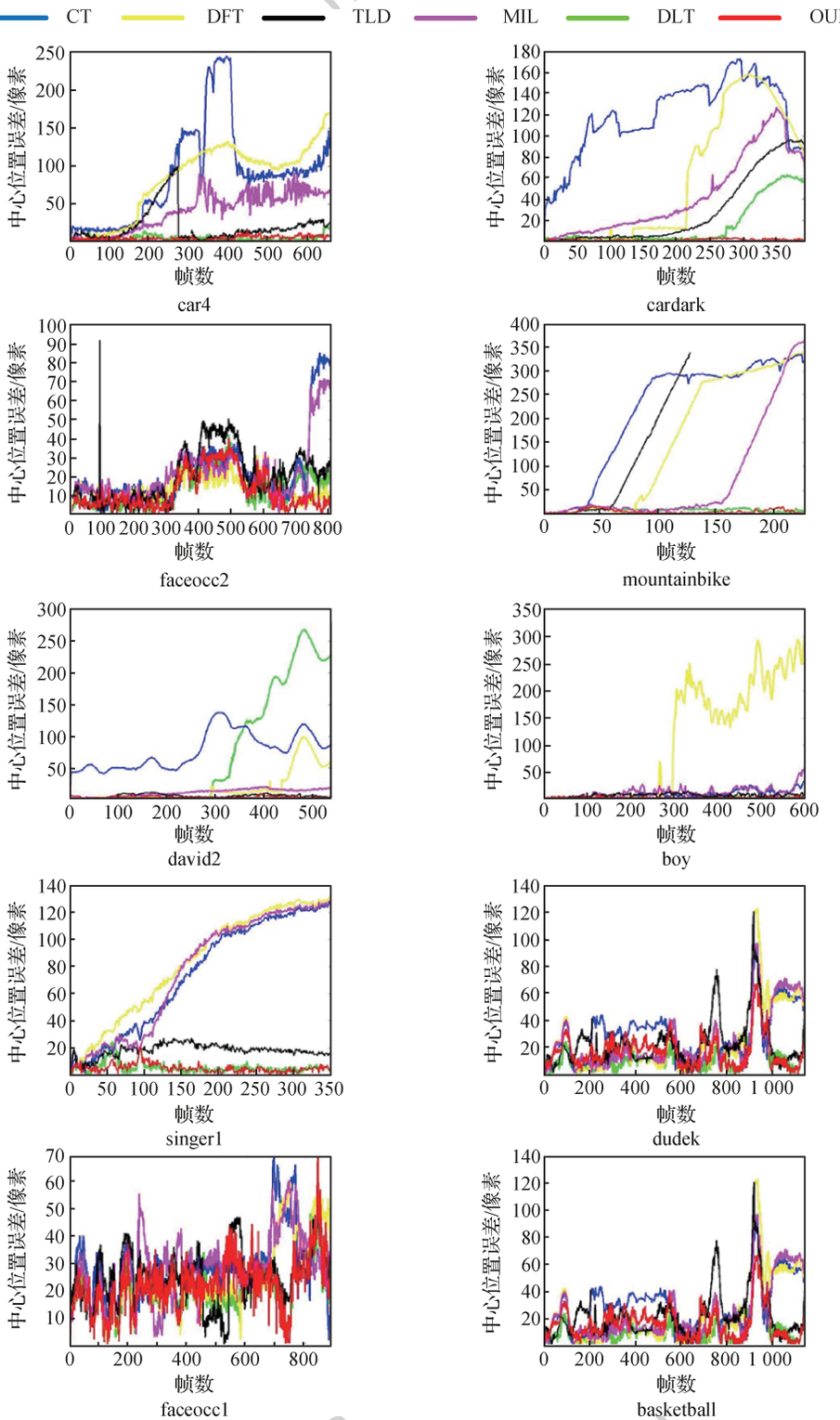


图 5　中心位置误差比较

Fig. 5　Frame-by-frame comparison of 6 trackers on 10 video sequences

序列中保持了较低跟踪误差,表现出优于其他对比算法的跟踪精度,具有较好的跟踪鲁棒性。

4.3 算法实时性分析

在视觉跟踪测试中,帧率是衡量算法运行速度的指标。在型号为 TITANX 的 GPU 驱动下,对比同样基于深度学习的两种算法 FDST 和 DLT 在 10 组实验中的运行速度,结果如表 3 所示。

由于 FDST 提取特征的网络各层维度均减小为 DLT 的 1/4,算法的计算复杂度大幅降低。跟踪成功的平均帧率为 22 帧/s,基本满足实时要求。

表 1 平均覆盖率

Table 1 Overlap percentage

/%

图像序列	CT	DFT	TLD	MIL	DLT	本文算法
car4	20.3	20.2	—	34.6	*86.1*	**84.8**
cardark	1.3	39.1	43.2	20.4	*57.1*	**87.9**
faceocc2	57.1	66.7	57.9	59.6	*67.7*	**71.8**
mountainbike	13.8	30.6	5.3	45.7	*74.4*	**84.2**
david2	1.5	54.4	*66.3*	45.5	45.2	**78.0**
boy	63.0	40.8	64.4	50.6	**84.3**	*82.7*
singer1	31.3	30.5	72.6	35.1	*83.7*	**83.8**
dudek	74.0	81.2	—	80.5	**88.9**	*80.8*
faceocc1	68.7	**78.7**	68.7	72.6	76.2	*77.2*
basketball	25.6	**60.7**	—	21.9	48.8	*51.7*

注:黑体表示每个视频的最优结果,斜体为次优结果,—表示算法在视频中失效。

表 2 成功率与平均中心位置误差值

Table 2 Success rate, while the number in parentheses denotes the central-pixel error

图像序列	成功率/%(平均中心位置误差值/像素)					
	CT	DFT	TLD	MIL	DLT	本文算法
car4	24.6(87.1)	24.9(82.5)	—	29.6(40.2)	*100*(*6.0*)	**100**(**4.6**)
cardark	1.0(119.1)	33.6(58.7)	52.7(28.3)	17.6(42.9)	*69.7*(*15.3*)	**100**(**1.3**)
faceocc2	62.4(22.6)	83.6(14.2)	59.8(20.2)	72.6(21.3)	*75.9*(*13.9*)	**85.1**(**12.4**)
mountainbike	16.6(216.6)	35.1(156.2)	26.3(316.4)	59.7(74.4)	**100**(*8.0*)	**100**(**5.0**)
david2	1.2(76.8)	54.1(17.0)	*87.3*(*5.9*)	32.7(11.0)	54.0(73.1)	**99.4**(**3.7**)
boy	71.5(8.7)	48.3(106.7)	89.0(8.0)	46.6(13.1)	**100**(**2.7**)	**100**(*3.0*)
singer1	23.2(74.4)	29.3(83.4)	99.7(17.9)	35.9(76.5)	**100**(**5.2**)	**100**(*8.0*)
dudek	83.7(29.2)	77.8(25.0)	—	83.2(24.7)	*96.0*(12.0)	**97.4**(*16.4*)
faceocc1	85.9(28.4)	85.8(24.0)	90.5(24.6)	81.7(28.6)	**99.7**(**20.4**)	*98.3*(*21.6*)
basketball	25.9(89.1)	**71.5**(*18.1*)	—	27.5(91.9)	51.5(23.4)	*56.0*(**17.8**)

注:黑体表示每个视频的最优结果,斜体为次优结果,—表示算法在视频中失效。

表 3 DLT 算法与本文算法的效率对比

Table 3 Comparison of efficiency with DLT method

/(帧/s)

算法	car4	cardark	david2	faceocc1	faceocc2	boy	singer1	basket	Mtbike	dudek	average
DLT	14.7	14.5	14.8	13.9	13.8	12.3	13.3	10.1	11.4	14.1	13.3
本文算法	24.8	24.6	23.8	23.5	23.8	19.1	19.7	18.0	17.7	23.2	21.8

5 结 论

本文提出了一种新的单目标视觉跟踪算法(FDST)。该算法基于深度学习理论,应用海量数据离线训练 SDAE 网络,使其具备描述图像特征的能力;利用改进的 SVM 作为打分器,对粒子滤波框架下的目标候选区进行打分,生成跟踪结果;结合阈值判定的模板更新策略,完成对目标的在线跟踪。实验表明,在目标发生平移、旋转和尺度变化,或存在光照、遮挡和复杂背景干扰时,FDST 能够实现比较稳定和相对快速的目标跟踪。

实验结果表明了本文算法的有效性,但算法依然存在许多问题。例如,算法对目标的快速移动和运动模糊的鲁棒性不够高,容易受到相似物体的干扰。此外,FDST 是在 GPU 驱动下运行的,现阶段的工程适用性受到一定的限制。后续研究的重点将从这些问题出发,探索更加鲁棒的视觉跟踪系统。

参考文献(References)

[1] Arulampalam M S, Maskell S, Gordon N, et al. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking[J]. IEEE Transactions on Signal Processing, 2002, 50(2): 174-188. [DOI: 10.1109/78.978374]

[2] Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619. [DOI: 10.1109/34.1000236]

[3] Adam A, Rivlin E, Shimshoni I. Robust fragments-based tracking using the integral histogram[C]//Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006: 798-805. [DOI: 10.1109/CVPR.2006.256]

[4] Mairal J, Bach F, Ponce J, et al. Discriminative learned dictionaries for local image analysis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE, 2008: 1-8. [DOI: 10.1109/CVPR.2008.4587652]

[5] Fujimura K, Nanda H. Visual tracking using depth data: US, 7372977[P]. 2008-05-13.

[6] Xin J, Liu X D, Ran B J, et al. Adaptive multiple cues integration for robust outdoor vehicle visual tracking[C]//Proceedings of the 34th Chinese Control Conference. Hangzhou, China: IEEE, 2015: 4913-4918. [DOI: 10.1109/ChiCC.2015.7260402]

[7] Ma L, Lu J W, Feng J J, et al. Multiple feature fusion via weighted entropy for visual tracking[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 3128-3136. [DOI: 10.1109/ICCV.2015.358]

[8] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends® in Machine Learning, 2009, 2(1): 1-127. [DOI: 10.1561/2200000006]

[9] Xie S N, Yang T B, Wang X Y, et al. Hyper-class augmented and regularized deep learning for fine-grained image classification[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 2645-2654. [DOI: 10.1109/CVPR.2015.7298880]

[10] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 1-9. [DOI: 10.1109/CVPR.2015.7298594]

[11] Vesely K, Ghoshal A, Burget L, et al. Sequence-discriminative training of deep neural networks[C]//Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013: 2345-2349.

[12] Yu D, Hinton G, Morgan N, et al. Introduction to the special section on deep learning for speech and language processing[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 4-6. [DOI: 10.1109/TASL.2011.2173371]

[13] Zheng Y, Chen Q Q, Zhang Y J. Deep learning and its new progress in object and behavior recognition[J]. Journal of Image and Graphics, 2014, 19(2): 175-184. [郑胤, 陈权崎, 章毓晋. 深度学习及其在目标和行为识别中的新进展[J]. 中国图象图形学报, 2014, 19(2): 175-184.] [DOI: 10.11834/jig.20140202]

[14] Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking[C]//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: MIT Press, 2013: 809-817.

[15] Li H X, Li Y, Porikli F. DeepTrack: learning discriminative feature representations by convolutional neural networks for visual tracking[C]//Proceedings of the British Machine Vision Conference. Nottingham, UK: BMVA Press, 2014: 110-119. [DOI: 10.5244/C.28.56]

[16] Yu H, Kim J, Kim Y, et al. An efficient method for learning nonlinear ranking SVM function[J]. Information Sciences, 2012, 209: 37-48. [DOI: 10.1016/J.INS.2012.03.022]

[17] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507. [DOI: 10.1126/science.1127647]

[18] Custódio A L, Rocha H, Vicente L N. Incorporating minimum Frobenius norm models in direct search[J]. Computational Opti-

mization and Applications, 2010, 46(2): 265-278. [DOI: 10.1007/s10589-009-9283-0]

[19] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.

[20] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems 25. Harrahs and Harveys, Lake Tahoe: MIT Press, 2012: 1097-1105.

[21] Ross D A, Lim J, Lin R S, et al. Incremental learning for robust visual tracking[J]. International Journal of Computer Vision, 2008, 77(1-3): 125-141. [DOI: 10.1007/s11263-007-0075-7]

[22] Krizhevsky A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009.

[23] Julier S J, Uhlmann J K. Unscented filtering and nonlinear estimation[J]. Proceedings of the IEEE, 2004, 92(3): 401-422. [DOI: 10.1109/JPROC.2003.823141]

[24] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013: 2411-2418. [DOI: 10.1109/CVPR.2013.312]

[25] Babenko B, Yang M H, Belongie S. Visual Tracking with Online Multiple Instance Learning[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 983-990. [DOI: 10.1109/CVPR.2009.5206737]

[26] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409-1422. [DOI: 10.1109/TPAMI.2011.239]

[27] Zhang K H, Zhang L, Yang M H. Real-time compressive tracking[C]//Proceedings of the 12th European Conference on Computer Vision. Berlin Heidelberg: Springer, 2012: 864-877. [DOI: 10.1007/978-3-642-33712-3_62]

[28] Sevilla-Lara L, Learned-Miller E. Distribution Fields for Tracking[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012: 1910-1917. [DOI: 10.1109/CVPR.2012.6247891]