基于卷积神经网络与一致性预测器的稳健视觉跟踪

高 琳[1]，王俊峰[2]，范 勇[1]，陈念年[1]

[1]西南科技大学计算机科学与技术学院，四川 绵阳 621010；

[2]四川大学计算机学院，四川 成都 610065

摘要 针对视频序列的稳健性目标跟踪问题，提出一种基于卷积神经网络(CNN)与一致性预测器(CP)的视觉跟踪算法。该算法通过构建一个双路输入 CNN 模型，同步提取帧采样区域和目标模板的高层特征，利用逻辑回归方法区分目标与背景区域；将 CNN 嵌入至 CP 框架，利用算法随机性检验评估分类结果的可靠性，在指定风险水平下，以域的形式输出分类结果；选择高可信度区域作为候选目标区域，优化时空域全局能量函数获得目标轨迹。实验结果表明，该算法能够适应目标遮挡、外观变化以及背景干扰等复杂情况，与当前多种跟踪算法相比具有更强的稳健性和准确性。

关键词 机器视觉；目标跟踪；卷积神经网络；一致性预测器；时空域能量函数

中图分类号 TP391.4 **文献标识码** A

doi：10.3788/AOS201737.0815003

Robust Visual Tracking Based on Convolutional Neural Networks and Conformal Predictor

Gao Lin[1], Wang Junfeng[2], Fan Yong[1], Chen Niannian[1]

[1] *School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan* 621010, *China*;

[2] *College of Computer Science, Sichuan University, Chengdu, Sichuan* 610065, *China*

Abstract On the issues about the robustness in visual object tracking, a novel visual tracking algorithm based on convolutional neural network (CNN) and conformal predictor (CP) is proposed. A two-input CNN model is constructed to extract the high level features from the sampled image patches and target template simultaneously, and the logistic regression is used to separate the object from the background. The CNN classifier is embedded into the CP framework, and the reliability of classification is evaluated via algorithms randomness testing. The classification result with credibility is obtained by region prediction at a specified significance level. The image patches with high credibility are selected as candidate objects, thus, the target trajectory is obtained through spacetime optimization. Experimental results show that the proposed algorithm can adapt to the occlusion, target appearance changes and complex background, and it has a better robustness and higher precision than the current algorithms.

Key words machine vision; target tracking; convolutional neural networks; conformal predictor; spacetime optimization

OCIS codes 150.0155; 150.1135; 100.4999

1 引 言

视觉目标跟踪是计算机视觉领域中的一个基本问题，其任务是确定目标在视频中的运动状态，包括位置、速度以及运动轨迹等。尽管近年来视觉跟踪技术取得了较大进展，但是在目标遮挡、姿态变化、混杂背景等复杂情况下，要实现稳健性较强的跟踪仍然面临巨大挑战。

视觉跟踪问题中的目标特征表达是影响跟踪性能的重要因素之一。用来表达目标的特征应具有适应目

收稿日期：2017-02-24；**收到修改稿日期**：2017-04-23

基金项目：国家自然科学基金(91338107，91438119，91438120)、教育部博士点基金(20130181110095)

作者简介：高 琳(1976—)，男，博士，讲师，主要从事计算机视觉、模式识别方面的研究。E-mail：gaolinscu@163.com

标外观变化以及对背景具有较好的区分性的特点。大量的特征提取方法被应用于视觉跟踪，如 Harr[1]、HOG[2]等，这些通过手工设计的底层特征，具有较强的针对性，但是对目标变化不具备稳健性。

近年来，深度学习技术中的卷积神经网络(CNN)广泛应用于目标检测、图像分类、语义分割等领域[3-6]。相比于传统的手工特征，基于 CNN 的自动学习特征能够捕捉目标的高层次语义信息，对目标外观变化具有较强的稳健性，因此逐渐被引入到目标跟踪问题的求解[7-9]。应用深度特征进行跟踪时，需要大量样本进行训练并更新 CNN 参数，而对于视觉跟踪任务，通常难以预先获得大量跟踪目标的训练样本，因此，CNN 参数的有效训练与更新是其应用于跟踪所面对的主要问题。

另一方面，利用 CNN 提取目标特征后，通常以判别式方法实现跟踪[7-8]，其基本思想是将目标跟踪视为图像区域的二值分类问题，通过分类器将图像区域分为目标和背景区域，根据每帧的分类结果获得最终轨迹。然而，分类结果的可靠性是决定跟踪成败的关键，目前的分类算法大都缺少对输出结果的可靠性分析，即通过量化的可信度评价结果在多大程度上是正确的这一过程。如果能够有效评估每个时刻的分类结果，并为目标状态估计以及特征模型参数更新提供可靠的信息依据，将会大幅提高跟踪的准确性和稳健性。

本文在 CNN 特征提取与分类研究的基础上，引入一致性预测器(CP)对分类结果进行可靠性分析，并提出一种基于分类可信度的视觉跟踪算法。首先，利用 CNN 提取图像中采样样本区域的高层特性，通过逻辑回归判别目标和背景区域；然后，采用 CP 评估分类结果的可信度，基于可信度选择每帧中的候选目标区域；最后，通过时空域全局能量优化实现目标跟踪。

2　算法概述

本文算法流程如图 1 所示，主要分为两个阶段：1)初始化阶段——构建一个双路输入的 CNN，其中的卷积层参数利用现有图像数据集预先训练得到，其他层则利用首帧中手工采样的样本进行训练并得到模型的初始参数；2)跟踪阶段——对序列图像逐帧进行区域采样，利用 CNN 提取样本的高层特征，通过逻辑回归计算样本属于目标或背景的回归值，进而采用 CP 获得指定风险水平下样本的类别，然后选择可信度高的目标样本建立候选目标集，优化定义在候选目标集上的时空能量函数获得最终的目标轨迹。采用一种半离线方式处理长序列的目标跟踪，将整个视频序列分段，依次处理每个序列段的跟踪并连接分段轨迹，同时在跟踪过程中逐段对 CNN 的模型参数进行在线更新。

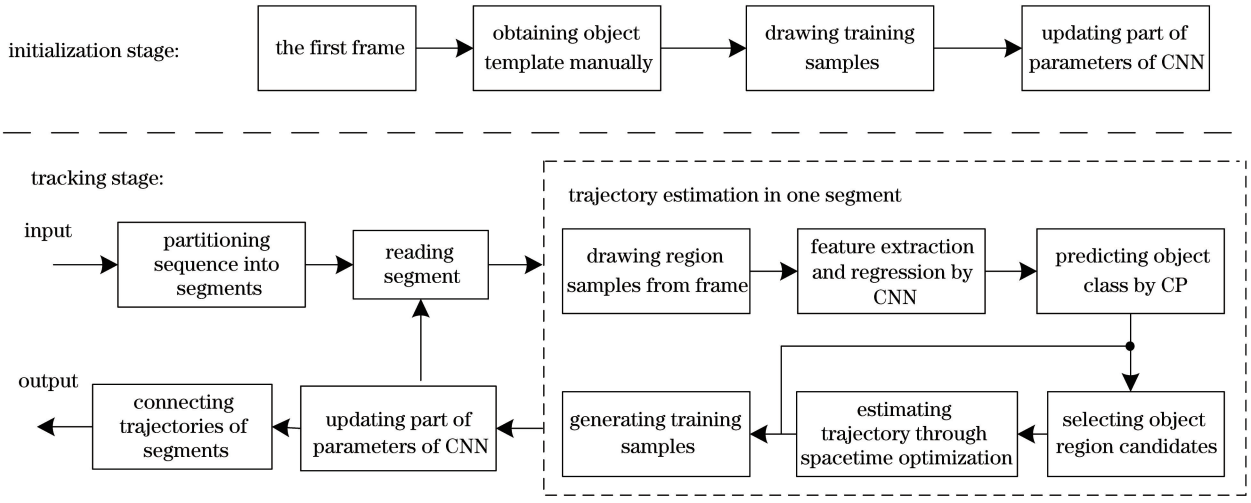

图 1　跟踪算法框图

Fig. 1　Flow chart of tracking algorithm

3　CNN 目标特征提取与分类

CNN 是一种专门处理栅格结构数据的多层神经网络，其通过卷积核隐式地提取图像局部特征，并具有良好的位移、缩放以及其他类型形变的不变性。而对于跟踪问题，网络结构和参数训练方式是影响 CNN 性

能的关键因素，因此，必须对两者进行充分设计。

3.1　CNN 网络结构

在目标识别应用中，CNN 通常需要经过大量数据进行训练后才能准确地表达目标特征，而对于一个特定的跟踪任务，往往难以预先获得充分的训练数据，因此，应用于目标识别的 CNN 难以直接应用于目标跟踪，还需进行调整与改进。

与目标识别不同，目标跟踪中不必关注目标的具体种类，只要能与背景区分即可，为此，采用一种双路输入的 CNN 网络结构，如图 2 所示。将目标模板与待识别图像两路信息同时输入网络，卷积层提取特征后，在全连接(FC)层融合形成判别特征，最终在输出层进行逻辑回归实现分类。其中，目标模板可通过手工在序列图像首帧中获得，而待识别图像则是在序列图像中采样的局部区域；网络中包含了两套独立的卷积层，为简化模型，两套卷积层共享同样的结构和参数；两路输入在经过卷积层后被映射为高层特征，然后在 FC 层中进行融合，进一步映射为对目标与背景具有区分性的特征；输出层为 Logistic 回归分类器，通过逻辑回归预测输入样本的类别，即目标或是背景。

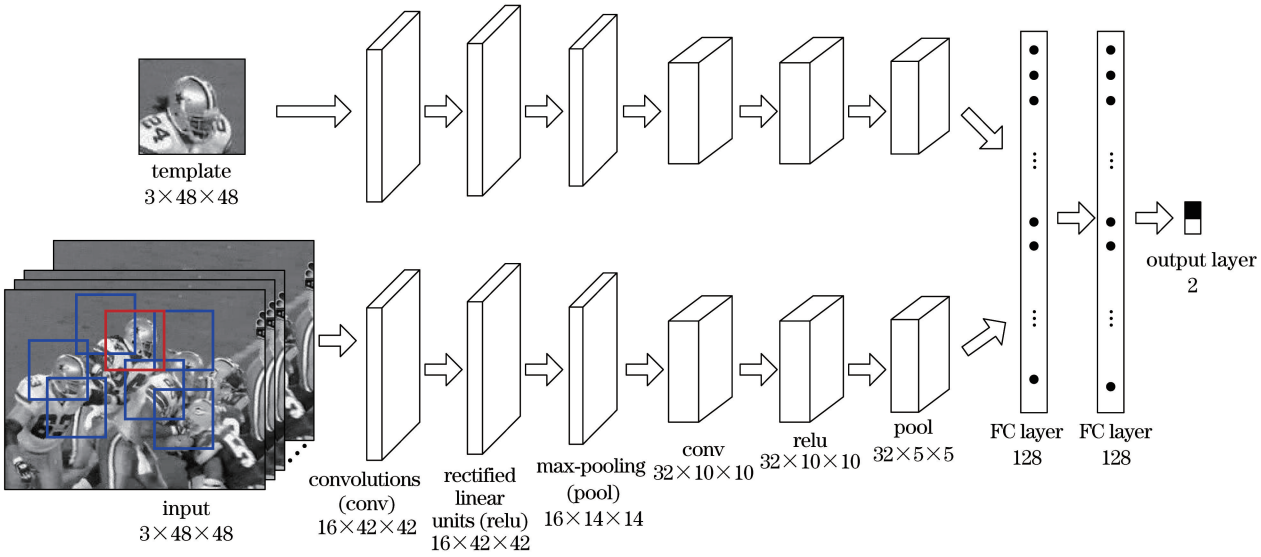

图 2　CNN 结构

Fig. 2　Structure of CNN

3.2　网络参数训练

文中 CNN 卷积层预先在 CIFAR-10 数据集[11]上进行离线训练，使之能够提取通用目标特征。在预训练时，CNN 网络结构简化为单输入结构，训练后的参数被两套卷积层共享。此外，针对 CIFAR-10 数据的 10 分类问题，CNN 的输出层设为 10 个单元，当预训练结束后，再将输出层替换为 1 个单元，以对应跟踪任务的二分类问题。预训练后的 CNN 将根据实际跟踪任务进行参数微调。在跟踪过程中，为了提高参数调整效率，将预训练后的卷积层参数固定，仅对 FC 层和输出层参数进行在线更新，以适应目标和背景的变化。

对于训练集的建立，在跟踪初始化阶段，手工选取首帧中的目标区域，根据目标区域采样正负训练样本，以样本与目标区域的覆盖率(设定阈值为 0.5)来判断其正负属性。为了提高训练样本数量，对样本进行随机的尺度和旋转变换以实现数据增强。在后续跟踪中，通过分类结果的风险评估，选取满足可信度条件的跟踪结果(选取方法见 5.2 节)作为中心进行训练样本采样。

训练集为 $T=\{[x^{(1)},y^{(1)}],\cdots,[x^{(n)},y^{(n)}]\}$，其中 $y^{(i)}\in\{C^-=0,C^+=1\}$，类标签 C^- 为背景、C^+ 为目标；$x^{(i)}\in Z^d$ 为目标状态向量，包括位置和尺度。在输出层利用 Logistic 回归计算样本属于目标或背景的概率，即

$$R(y\,|\,x;\theta)=h_\theta(x)^y\cdot[1-h_\theta(x)]^{1-y},\tag{1}$$

式中 $h_\theta(x)=\dfrac{\exp(\theta^{\mathrm{T}}x)}{1+\exp(\theta^{\mathrm{T}}x)}$，$\theta$ 为网络模型参数。利用训练集 T 训练模型，使得对数似然损失函数 $L(\theta)$ 达

到最小，

$$L(\theta)=-\frac{1}{m}\sum_{i=1}^{m}\{y_i\ln[h_\theta(x_i)]+(1-y_i)\ln[1-h_\theta(x_i)]\}。\tag{2}$$

采用随机梯度下降法沿着 $L(\theta)$ 的负梯度方向调整网络权值和偏置值，通过反向传播方法对卷积层以上的各层参数同时迭代更新。

4　基于 CP 的候选目标选择

Logistic 回归值为样本类别预测提供了依据，但 Logistic 回归值本身无法对预测错误的风险进行理论评估。为了实现预测结果的可靠性分析，将 CNN 模型嵌入至 CP 框架中，根据算法随机性水平计算样本类别的可信度，进而选择候选目标。

4.1　CP

通过量化的可信度可以评价预测结果在多大程度上是正确的，而目前的机器学习算法对于预测结果大都缺乏该过程，有效可信度的衡量标准是可校准性[10]。CP 是一种能够有效输出可信度的机器学习范式，其利用假设检验方法进行预测，并对预测结果提供可靠性评估。

传统的 CP 算法计算量很大，为提高运算效率，采用 CP 的改进算法预测样本类别，即归纳一致性预测器(ICP)[11]。在 ICP 算法中，首先假定训练集中的样本服从独立同分布，将训练集 $T=\{[x^{(1)},y^{(1)}],\cdots,[x^{(n)},y^{(n)}]\}$ 划分为两个部分：前 m 个样本组成正常训练集 $T_a=\{[x^{(1)},y^{(1)}],\cdots,[x^{(m)},y^{(m)}]\}$；后面 q 个样本组成校准集 $T_b=\{[x^{(m+1)},y^{(m+1)}],\cdots,[x^{(m+q)},y^{(m+q)}]\}$，$n=m+q$。$T_a$ 用于更新 CNN 参数，T_b 与待识别样本一起构成检验序列，利用算法随机性检验确定样本类别。

算法随机性检验方法为：首先定义映射函数 $A:Z^{(q-1)}\times Z\to R$，将 T_b 中的每个样本一一映射至奇异值空间，得到奇异值序列 $\alpha_{m+1},\cdots,\alpha_{m+q}$。奇异值反映了该样本与样本整体分布的不一致性。令待识别样本的目标状态为 x^s，分别赋予 x^s 类别标签 C^- 和 C^+，从而构成两个检验样本 (x^s,y_i)，$i=0,1$。计算检验样本的奇异值 $\alpha_s^{y_i}$ 后，与 T_b 对应的奇异值一起构成两个检验序列 $\alpha_{m+1},\cdots,\alpha_{m+q},\alpha_s^{y_i}$，$i=0,1$。为了获得序列的算法随机性水平，计算检验统计量

$$p_s(y_i)=\frac{|\{j=m+1,\cdots,m+q,s:\alpha_j\geqslant\alpha_s^{y_i}\}|}{q+1},i=0,1,\tag{3}$$

式中 $p_s(y_i)$ 为目标状态 x^s 被标记为 y_i 时的 p 值，将其作为 x^s 属于类别 y_i 的可信度。指定算法风险水平阈值 ε，将 p 值大于 ε 的假设作为 ICP 的输出，即

$$\Gamma_s^\varepsilon=\{y_i:p_s(y_i)>\varepsilon,\ i=0,1\}。\tag{4}$$

当 x^s 的真实类别 y^s 不在 Γ_s^ε 中时，可认为出现了预测错误，根据 CP 有效性定理[10]，其错误率不会大于算法风险水平 ε，即

$$P[p_s(y^s)\leqslant\varepsilon]\leqslant\varepsilon,\tag{5}$$

因此，ICP 的预测域具有可校准性。

4.2　样本奇异函数

序列的算法随机性检验需要先定义奇异值映射函数，用来度量待检验样本隶属于整体样本分布的一致性程度。根据 CNN 输出的回归值分析一致性，样本特征对应于真实类别的回归值越大，认为该样本与校准集序列的一致性越强，奇异值函数定义为

$$a_i=\frac{1-R^y[x^{(i)}]}{R^y[x^{(i)}]+\gamma},\tag{6}$$

式中 $R^y[x^{(i)}]$ 为由(1)式得到的 $x^{(i)}$ 对应于类别 y 的回归值；参数 γ 用于调节奇异值 a_i 对回归值变化的敏感度，γ 越小，a_i 对 $R^y[x^{(i)}]$ 的变化越敏感。

4.3　候选目标选择

ICP 输出的结果为一个集合，其中可能包含多个类别。对于待识别样本的二分类问题，ICP 输出的结果

有 4 种可能性，即 ϕ、$\{C^+\}$、$\{C^-\}$、$\{C^+,C^-\}$。每个输出结果中，除了类别信息，还附带有可信度 p 值。从所有样本的域预测结果中选择可信度高的样本作为每帧的候选目标。即对于 t 时刻的图像帧，将该帧中输出为$\{C^+\}$或$\{C^+,C^-\}$的样本按照可信度 $p(C^+)$值进行排序，选取最大的 N_c 个样本建立候选目标集 O_t，可知$|O_t|\leqslant N_c$。

5　目标跟踪算法

5.1　时空域能量函数

候选目标集 O_t 包含了 t 时刻目标的若干个可能状态，目标将从 O_t 中的某个状态转换到下一时刻候选目标集 O_{t+1} 中的某个状态，因此，可以将目标跟踪视为寻找最优路径问题。为了获得最优路径，定义时空域能量函数 E_{Track} 描述目标轨迹，通过优化能量函数即可得到目标轨迹

$$\mathcal{T}_{1:N}=\arg\min_{x_{1:N}}E_{\text{Track}}=\arg\min_{x_{1:N}}(E_{\text{Local}}+E_{\text{Pairwise}}),\tag{7}$$

式中 $\mathcal{T}_{1:N}=\{x_1^*,\cdots,x_N^*\}$，$E_{\text{Track}}$ 包含局部代价项 E_{Local} 和逐对代价项 E_{Pairwise} 两部分。E_{Local} 定义为每个时刻的目标状态 x_t 对应于背景的 CNN 输出值之和。由于目标部分遮挡情况会降低局部代价项的可靠性，为此，引入稳健估计算子来降低出格点数据对函数优化的影响，表示为

$$E_{\text{Local}}=\sum_{t=2}^{N}\rho[R_t^-(x)],\tag{8}$$

式中 $R_t^-(x)$为目标状态 x 对应于背景时的回归值；$\rho(\cdot)$为 Huber 算子，用于增强局部代价项的可靠性，其定义为

$$\rho(a)=L\delta(a)=\begin{cases}a^2/2, & |a|\leqslant\delta\\ \delta\cdot(|a|-\delta/2), & \text{otherwise}\end{cases}。\tag{9}$$

式中 E_{Pairwise} 描述目标状态的变化程度。当序列中出现目标遮挡、杂乱背景或是目标姿态变化时，目标状态会由于估计误差较大而出现跳跃式变化。假定目标的运动是连贯的，E_{Pairwise} 的作用是在能量函数优化时对轨迹中的突变点进行惩罚，使得轨迹具有一定的平滑性。其定义为

$$E_{\text{Pairwise}}=\sum_{t=2}^{n}\|x_t-x_{t-1}\|^2。\tag{10}$$

采用动态规划方法对(7)式中的能量函数进行优化[19]，即可得到最优的运动轨迹。

5.2　训练样本更新

跟踪过程中，利用上一个序列段的跟踪结果更新 CNN 模型参数，然后处理下一个序列段。为避免出现模型漂移，仅在可靠性高的跟踪结果上采集训练样本。对于时刻 t 的跟踪结果 x_t^*，根据其可信度 p 值进行选择，若 p 大于设定的阈值 α，则基于 x_t^* 采样正负训练样本，否则进入下一时刻进行判断选择。

训练集中的负样本普遍存在冗余现象，冗余的负样本对模型训练贡献很小，浪费计算资源。为此，通过挖掘难负样本[13]优化训练集，提高训练效率。实验中发现，域预测结果为$\{C^+,C^-\}$的样本(记为 $x_t^{\pm}$)通常会出现在背景物与目标易混淆的情况下，因此，可以从这类样本中选择难负样本。可以采用一种简单的选择方式，即判断 $x_t^{\pm}$ 与当前跟踪结果 x_t^* 之间是否存在区域交叠，若没有交叠，则将 $x_t^{\pm}$ 作为负样本添加至训练集中。

5.3　跟踪算法步骤

所提出的 CNN 与 CP 的稳健视觉跟踪算法具体步骤如下：

输入：目标初始状态 x_0，预训练的 CNN，长度为 N 的序列图像

输出：目标运动轨迹 $\mathcal{T}_{1:N}=\{x_1^*,\cdots,x_N^*\}$

初始化阶段：

1) 将 x_0 对应的图像区域作为 CNN 的输入模板；

2) 在 x_0 处采集正负样本，建立训练集 T，并将其划分为正常训练集 T_a 和校准集 T_b；

3) 利用 T_a 对 CNN 中的 FC 层和输出层进行训练调整。

跟踪阶段：

4) 将图像序列划分为 $K=\lceil N/n_l \rceil$ 个片段，依次对第 $k=1,\cdots,K$ 片段进行处理；

5) 估计第 k 个片段的目标轨迹，其处理过程分为以下两步：

第一步，建立所有帧的候选目标集合 $O_{1:n_l}=\{O_1,\cdots,O_{n_l}\}$

①令当前时刻为 t，$O_t=\phi$，以时刻 t-1 图像中 p 值最高的目标状态 $\hat{x}_t$ 为中心，在位置和尺度上进行高斯分布随机采样，获得 M 个样本 $x_t^{(j)}$，$j=1,\cdots,M$，高斯分布的协方差为对角阵 Diag($0.1r^2$，$0.1r^2$，0.2)，r 为 $\hat{x}_t$ 的长和宽的平均值；

②利用 CNN 计算样本的回归值 $R^y[x_t^{(j)}]$；

③根据 T_b，利用(3)式计算 $x_t^{(j)}$ 的可信度 $p[x_t^{(j)}]$；

④根据风险阈值 ε，利用(4)式获得 $x_t^{(j)}$ 的域预测结果 $\Gamma^{\varepsilon}[x_t^{(j)}]$；选取输出结果为 $\{C^+\}$ 或 $\{C^+,C^-\}$，且可信度 $p(C^+)$ 值排在前 N_c 个的样本 $x_t^{(j)}$，加入至候选目标集 O_t，$O_t \leftarrow O_t \cup \{x_t^{(j)}\}$；

⑤令 $t=t+1$，若 $t>n_l$，则第一步处理结束，否则转至步骤①；

第二步，通过优化能量函数 E_{Track} 获得第 k 个片段的目标轨迹 $\mathcal{T}_k=\underset{x\in O_{1:t}}{\arg\min} E_{\text{Track}}$；

6) 更新训练集 T：根据 p 值选择可信度高的跟踪结果更新训练集，并挖掘难负样本加入至 T 中；

7) 连接目标轨迹 $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{T}_k$，若已处理完最后一个片段，输出最终的轨迹 $\mathcal{T}$；否则，令 $k=k+1$，转入步骤 5)。

6　实验结果与分析

为验证算法的有效性，利用 Matlab 软件进行仿真实验，硬件平台的 CPU 为 3.4 GHz intel-i7-6700，内存为 8 GB。算法的各项参数设置为：正常训练集 T_a 规模 $m=300$，校准集 T_b 规模 $q=30$，算法风险水平 $\varepsilon=0.4$，样本奇异值参数 $\gamma=0.5$，候选目标集规模上限 $N_c=20$，稳健函数参数 $\delta=0.4$，训练样本更新参数 $\alpha=0.6$。整个实验中算法参数保持不变，算法的平均处理速度约为 8 frame/s。

选用公开数据集 TOP100[14] 中的视频序列作为实验对象，并对当前多种主流跟踪算法实验效果进行对比，这些算法包括 VTS[15]、LOT[16]、STRUCK[1]、MIL[17] 和 KCF[2]。为了验证 CP 的有效性，在实验中测试了本文算法的一个简化版本，该版本中未引入 CP，而是直接根据 CNN 输出的回归值，选择最大的 N_c 个样本作为候选区域。实验中采用覆盖率和中心点位置误差两个标准[18] 比较各算法的性能。覆盖率定义为 $C_r=(R_s \cap R_t)/(R_s \cup R_t)$，其中 R_s、R_t 分别为跟踪结果区域和真实目标区域；中心点位置误差是指跟踪结果中心点与真值中心点之间的欧氏距离。图 3 给出了部分实验结果，选取的视频序列中包含了目标遮挡、外观变换、光照变化和复杂背景等典型的复杂情况。

图 3(a)给出了 FaceOcc1 视频序列的部分跟踪结果，跟踪目标为一名女子的脸部。该序列图像中人脸多次被书本遮挡，并且遮挡的部位和程度不同。从图中可以看出，在有遮挡的情况下，LOT 算法的跟踪结果仅局限于未被遮挡部分，尺度误差较大；而 MIL 算法在目标被严重遮挡时(第 834 帧中)出现了较大飘移。图 4(a)中的中心点位置误差和图 5(a)中的覆盖率数据均表明，KCF 算法以及本文算法始终能够准确定位目标，对遮挡具有较强的稳健性。

视频序列 Bolt 是短跑比赛场景，跟踪目标为其中一名运动员。该序列的挑战在于目标的姿态不断变化，同时随着镜头的转动图像中运动员从正面逐渐转向背面，因此目标外观变化很大。图 3(b)中，VTS、STRUCK、MIL 算法从序列开始不久就发生偏移，在第 48 帧中均脱离目标；KCF、LOT 和本文算法能够保持跟上目标，但 LOT 算法和未引入 CP 的算法在目标发生形变时(第 222 帧)出现了较大的尺度误差。本文算法利用高层特征，受目标外观变化的影响不大，并通过可靠性分析来更新模型，避免了漂移的出现，从图 4(b)和图 5(b)中的误差分析可以看出，本文算法的跟踪结果误差最小。

Football 视频序列是美式足球比赛场景，跟踪目标为一名球员的头部。该序列的难点是背景中有许多外观十分相似的球员，他们之间频繁交互运动，对目标跟踪造成了干扰。图 3(c)中，VTS、MIL、KCF、STRUCK 和未引入 CP 的算法多次出现较大漂移；在第 360 帧时，VTS、MIL 和 KCF 算法则完全跟踪到其

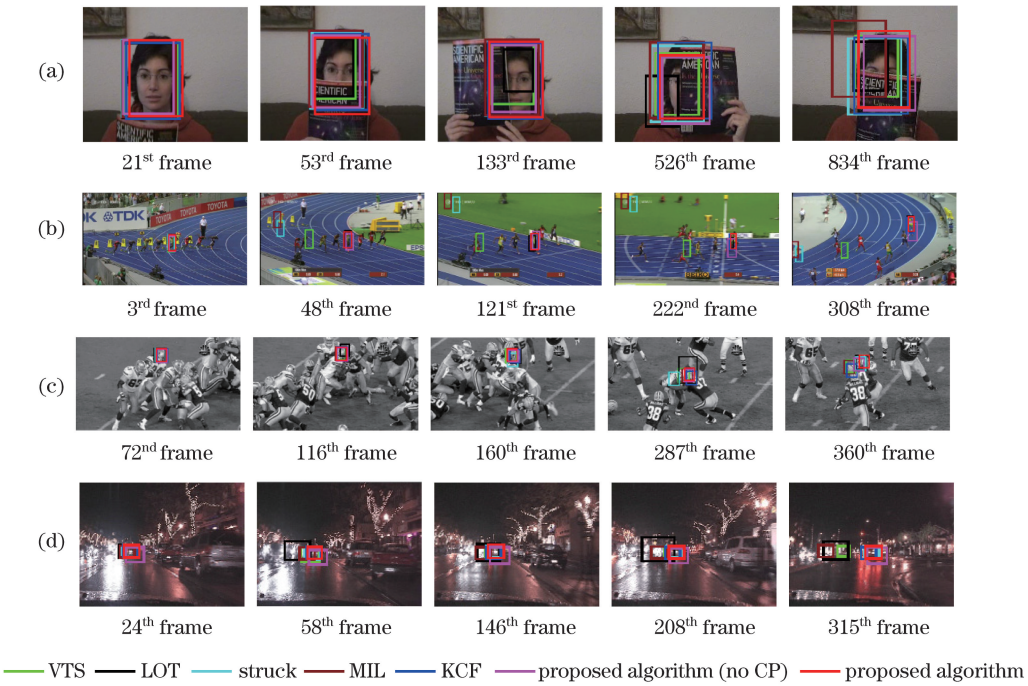

图 3　视频序列跟踪结果。(a) FaceOcc1；(b) Bolt；(c) Football；(d) CarDark

Fig. 3　Tracking results of video sequences. (a) FaceOcc1; (b) Bolt; (c) Football; (d) CarDark

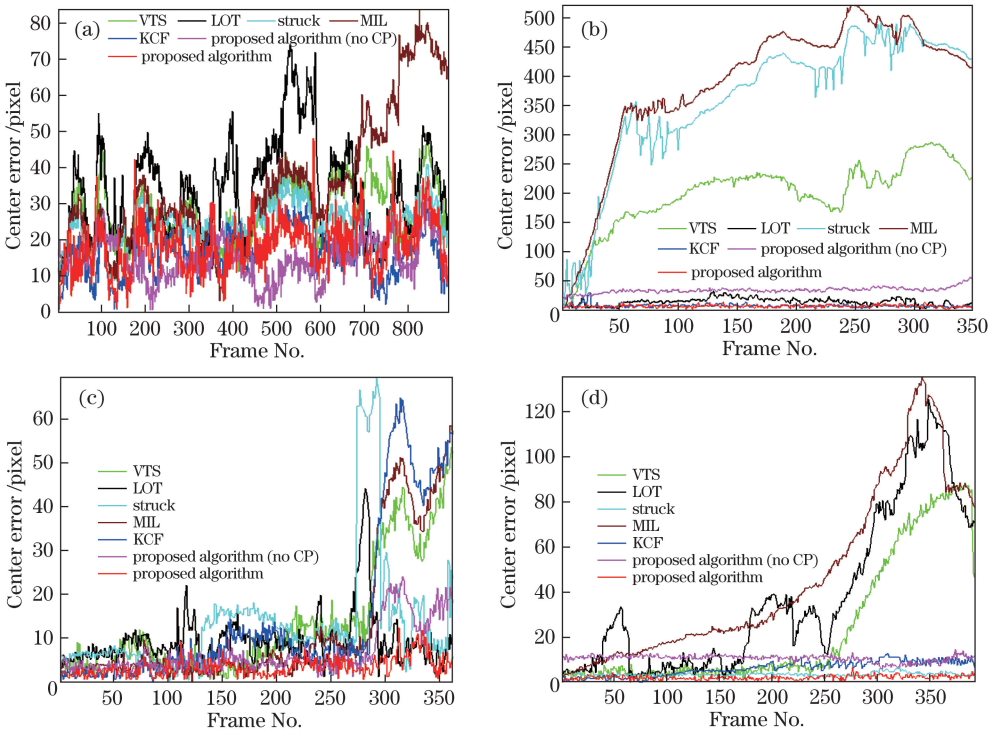

图 4　跟踪结果的中心点位置误差。(a) FaceOcc1；(b) Bolt；(c) Football；(d) CarDark

Fig. 4　Center position error of tracking results. (a) FaceOcc1; (b) Bolt; (c) Football; (d) CarDark

他球员；本文算法通过时空轨迹优化确保轨迹平滑性，降低了相似目标干扰的影响。图 4(c)和图 5(c)中的结果表明，本文算法在该序列的跟踪中保持了最低的跟踪误差。

CarDark 视频序列是夜景下汽车行驶场景，跟踪目标是一辆汽车的尾部，该序列的特点是光照剧烈变化，背景混杂且图像分辨率低。图 3(d)中，在第 58 帧，LOT 算法受到目标左侧出现的亮光干扰，严重偏离目标，同时尺度误差较大；MIL 和 VTS 算法也出现了一定程度的漂移，随着左侧亮光的不断干扰，MIL 和

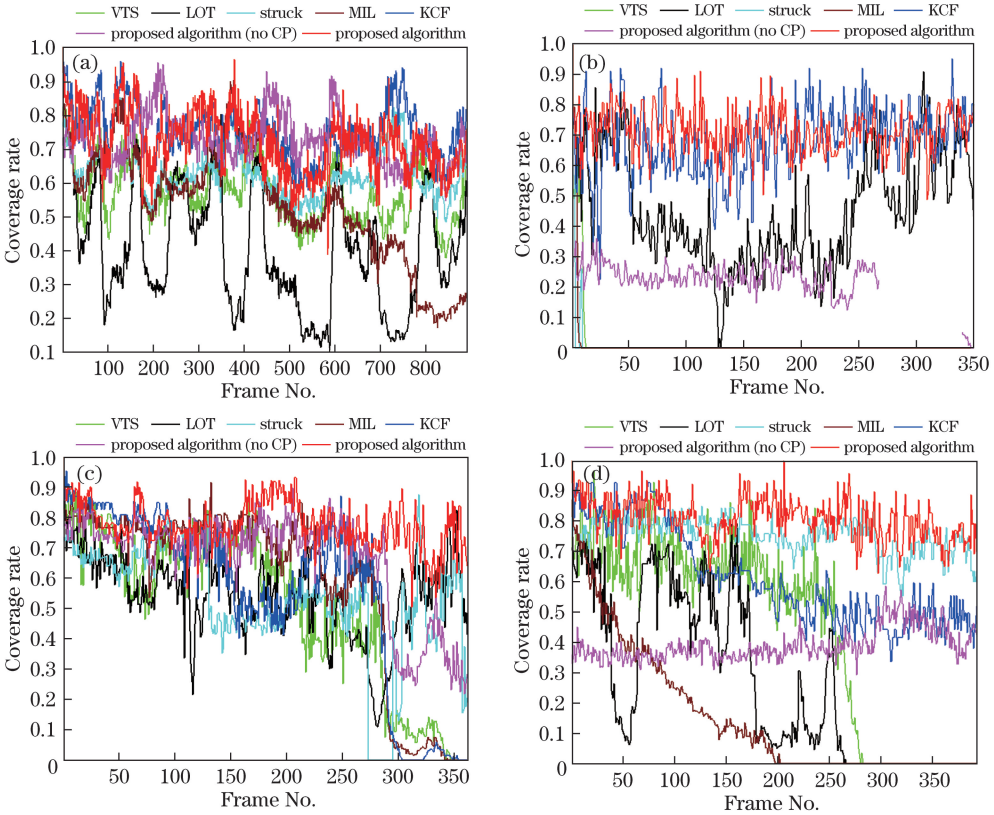

图 5　跟踪结果的覆盖率。(a) FaceOcc1；(b) Bolt；(c) Football；(d) CarDark

Fig. 5　Coverage rate of tracking results. (a) FaceOcc1; (b) Bolt; (c) Football; (d) CarDark

LOT 算法在第 208 帧时已丢失目标；而在第 315 帧时，路面上的倒影亮斑也造成了 VTS 算法丢失目标；STRUCK、KCF 和本文算法在跟踪车尾过程中保持稳定，但未引入 CP 的算法普遍存在较大的尺度误差。该序列的跟踪误差分析如图 4(d)和图 5(d)所示，其中 KCF 和 STRUCK 算法的跟踪精度略低于本文算法。

为了比较 7 种算法的整体性能，图 6 中给出了这些算法在所有测试序列上的一次通过评价结果[14]，包括位置精度图和覆盖成功率图。以曲线下面积的大小对算法的性能进行排序，可以看出，本文算法在位置精度和覆盖成功率上都高于其他算法，其中 KCF 的性能与本文算法最接近，而在未引入 CP 的条件下算法性能出现了下滑，尤其是在覆盖成功率上较为明显，如图 6(b)所示。

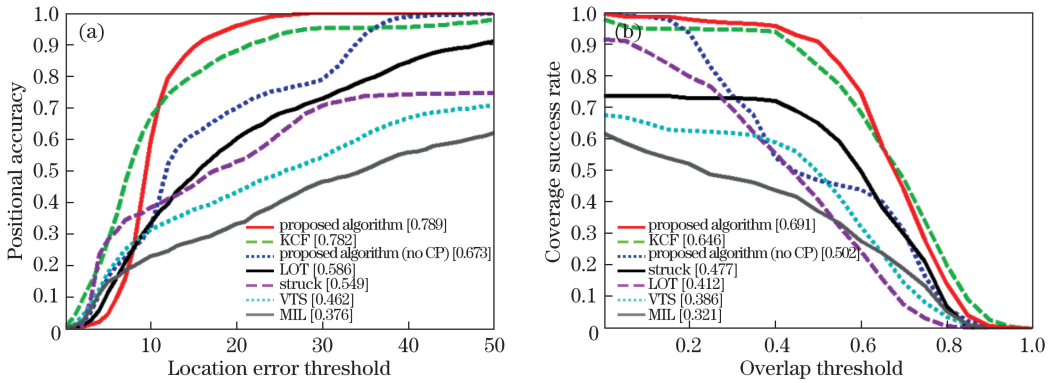

图 6　一次通过评价结果。(a)位置精度图；(b)覆盖成功率图

Fig. 6　One-pass evaluation. (a) Positional accuracy diagram; (b) coverage success rate diagram

表 1 给出了 7 种算法的平均中心点位置误差和平均覆盖率，可见本文算法的性能指标最优，表明本文算法中的 CNN 网络提取的深度特征能够很好地区分目标和背景，通过结合 ICP 对分类结果进行可信度评价，可以有效地保证跟踪结果的可靠性，并在多种典型复杂情况的视频序列上表现出良好的性能。

表 1　平均中心点位置误差和覆盖率

Table 1　Average center position error and coverage rate

Sequence	VTS	LOT	Struck	MIL	KCF	Proposed algorithm (no CP)	Proposed algorithm
Football	13.27(0.51)	9.23(0.54)	13.33(0.53)	12.55(0.58)	14.60(0.55)	6.73(0.66)	5.36(0.68)
FaceOcc1	27.42(0.57)	34.22(0.40)	24.50(0.63)	34.86(0.54)	15.98(0.75)	14.29(0.73)	20.43(0.70)
CarDark	23.45(0.45)	37.43(0.26)	3.42(0.75)	45.69(0.15)	6.05(0.61)	10.54(0.39)	3.04(0.74)
Bolt	197.53(0.02)	13.18(0.45)	360.19(0.01)	387.04(0.01)	6.37(0.68)	33.25(0.22)	8.65(0.65)
Total average	65.41(0.39)	23.52(0.41)	100.36(0.48)	120.03(0.32)	10.75(0.65)	16.20(0.50)	9.37(0.69)

7　结　论

提出了一种基于 CNN 与 CP 的目标跟踪算法。该算法采用 CNN 提取图像高层特征，克服了底层特征对目标外观变换敏感的缺点。为了提高跟踪稳健性，引入 CP 对分类结果进行可靠性分析，选择满足可信度条件的分类结果作为候选目标区域，最后通过时空域全局能量函数优化获得最终的目标轨迹。

在公开数据集上进行实验，并与多种目前流行的跟踪算法进行对比，结果表明，本文算法具有更优的跟踪稳健性和准确性。

参　考　文　献

[1] Hare S, Golodetz S, Saffari A, *et al*. Struck: structured output tracking with kernels[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2096-2109.

[2] Henriques J F, Caseiro R, Martins P, *et al*. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[3] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems, 2012: 1097-1105.

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations (ICLR), 2015.

[6] Xu Lu, Zhao Haitao, Sun Shaoyuan. Monocular infrared image depth estimation based on deep convolutional neural networks[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
许　路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报, 2016, 36(7): 0715002.

[7] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 4293-4302.

[8] Ma C, Huang J B, Yang X, *et al*. Hierarchical convolutional features for visual tracking[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015: 3074-3082.

[9] Wang L, Ouyang W, Wang X, *et al*. Visual tracking with fully convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015: 3119-3127.

[10] Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world[M]. New York: Springer, 2005.

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

[12] Papadopoulos H. Inductive conformal prediction: theory and application to neural networks[M]. Rijeka: Tools in Artificial Intelligence. InTech, 2008: 330-332.

[13] Sung K K, Poggio T. Example-based learning for view-based human face detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(1): 39-51.

[14] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013: 2411-2418.

[15] Kwon J, Lee K M. Tracking by sampling trackers[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011: 1195-1202.

[16] Oron S, Bar-Hillel A, Levi D, *et al*. Locally orderless tracking[J]. International Journal of Computer Vision, 2015,

111(2): 213-228.

[17] Babenko B, Yang M H, Belongie S. Visual tracking with online multiple instance learning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009: 983-990.

[18] Everingham M, Gool L V, Williams C K I, *et al*. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[19] Buchanan A, Fitzgibbon A. Interactive feature tracking using K-D trees and dynamic programming[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, 1: 626-633.