# 基于定位-分类-匹配模型的目标跟踪方法

刘大千[1*]，刘万军[2]，费博雯[3]

[1]辽宁工程技术大学电子与信息工程学院，辽宁 葫芦岛 125105；

[2]辽宁工程技术大学软件学院，辽宁 葫芦岛 125105；

[3]辽宁工程技术大学工商管理学院，辽宁 葫芦岛 125105

**摘要** 近年来卷积神经网络框架被成功地应用到目标跟踪领域，并取得了较为稳健的跟踪结果。基于此思想，提出一种基于定位-分类-匹配模型的目标跟踪方法。首先，在定位模型中，利用前一帧的位置信息预测当前帧中的候选目标区域。然后，采用已训练的深度特征对候选区域进行类间筛选，选出 $N$ 个次优目标区域。最后，利用常规颜色特征对次优目标区域进行类内寻优匹配，从而确定最终的跟踪目标。与此同时，分别对定位、分类中的网络进行更新，并对建立的匹配模型进行在线实时更新，使得其对目标的描述更加准确。在 OTB50 和 OTB100 标准数据库上进行实验测试，结果表明，提出的跟踪方法在快速运动、相似物体干扰、复杂背景等条件下具有较好的跟踪稳健性。

**关键词** 机器视觉；卷积神经网络；定位模型；类间筛选；寻优匹配；目标跟踪

**中图分类号** TP391.41 **文献标识码** A **doi**：10.3788/AOS201838.1115003

# Target Tracking Method Based on Location-Classification-Matching Model

Liu Daqian[1*], Liu Wanjun[2], Fei Bowen[3]

[1] *School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning* 125105, *China*;

[2] *School of Software, Liaoning Technical University, Huludao, Liaoning* 125105, *China*;

[3] *School of Business and Management, Liaoning Technical University, Huludao, Liaoning* 125105, *China*

**Abstract** Recently, the framework of convolution neural network has been successfully applied to the target tracking, and has achieved robust tracking results. On the basis of this conception, a target tracking method based on location-classification-matching model is proposed. First of all, in the location model, the candidate target region of the current frame is predicted by using location information of previous frame. Secondly, the trained deepth features are used to inter-class screen the candidate regions, and $N$ sub-optimal target regions are selected. Finally, we use conventional color features to perform intra-class optimization matching for sub-optimal target regions, so as to determine the final tracking target. Meanwhile, the network in the location and the classification is updated separately, and the established target model is updated online and real-time to ensure that the model describes the target accurately. Experimental tests are performed on OTB50 and OTB100 standard databases, the experimental results show that the proposed tracking method has better tracking robustness under the conditions of fast motion, similar object interference, and complex background.

**Key words** machine vision; convolution neural network; location model; inter-class screen; optimization matching; target tracking

**OCIS codes** 150.1135; 100.4996; 100.4999

## 1 引 言

目标跟踪是计算机视觉领域的重要组成部分[1-5]，被广泛应用于智能交通、机器视觉、运动捕捉等方面。现实场景中多种因素(光照变化、快速运动及复杂背景等)的影响很容易造成跟踪准确性的降低，因此如何有效地克服这些问题成为目标跟踪这一研究领域的关键。

**收稿日期**：2018-04-17；**修回日期**：2018-06-10；**录用日期**：2018-06-19

**基金项目**：国家自然科学基金(61172144)

* **E-mail**：liudaqianlntu@163.com

近年来，基于深度学习(DL)的方法被应用于跟踪领域。2013 年，Wang 等[6]提出深度学习跟踪(DLT)方法，将深度学习应用于目标跟踪领域。该算法首先使用栈式降噪自编码器(SDAE)在大规模自然图像数据集上进行无监督的离线预训练，获得通用的物体表征能力；然后取离线 SDAE 的编码部分叠加到 sigmoid 分类层组成分类网络，采用粒子滤波的方式估计当前帧的候选目标区域；最后采用限定阈值的方式进行目标模型更新。该方法在 OTB50[7]视频库中取得了非常稳健的跟踪效果。此后，一些流行的深度学习框架被成功应用到目标跟踪中，卷积神经网络 (CNN)是其中一种比较成熟的框架。Wang 等[8]利用非跟踪数据预训练与在线微调策略相结合的方法解决跟踪过程中训练数据不足的问题，使用 CNN 作为获取特征和分类的网络模型。该算法作为 CNN 在目标跟踪领域的一次成功应用，取得了非常优异的成绩。Ma 等[9]提出分层卷积特征的概念，将卷积神经网络与相关滤波器结合，在首帧中利用 Conv3_4、Conv4_4、Conv5_4 特征的插值分别训练，得到 3 个相关滤波器，然后利用上一帧目标位置获取的三个卷积层的特征做插值，并通过每层的相关滤波器预测区域置信度，最后从 Conv5_4 开始逐层预测，把最低层的预测结果作为跟踪的目标。Hong 等[10]将 CNN 与支持向量机(SVM)结合，利用 SVM 的分类属性对 CNN 进行训练，实验结果证明了该训练方式的有效性。Wang 等[11]直接使用在大规模分类数据库 ImageNet 上训练出的 CNN 获得目标的特征表示，然后利用观测模型进行分类以获得最终的跟踪结果。该方法不仅避免了跟踪时直接训练 CNN 样本不足的困境，而且还充分利用了深度特征强大的表征能力。Nam 等[12]提出了一种基于学习的多域网络(MDNet)，该网络分为共享层和特定域层两部分，将每个训练序列当成一个单独的特定域，每个域都有一个针对它的二分类层用于区分当前序列的前景和背景，而网络之前的所有层都是共享的，该共享层达到了学习视频序列中目标特征表达的目的，同时，特定域层又解决了不同训练序列分类目标不一致的问题。

尽管基于 CNN 的方法在目标跟踪方面取得了理想的效果，但也存在以下不足之处：1)基于传统 CNN 的跟踪方法仅利用前一帧的目标位置做定位，当目标快速运动或变形时，这类方法易发生跟踪漂移；2)在跟踪目标的过程中，仅利用分类网络确定候选目标区域，当背景中存在相似物体干扰时，跟踪方法易发生错误匹配从而导致丢失目标；3)为了提高跟踪的准确率，一些跟踪方法的网络结构比较复杂，时间复杂度较高。

本文提出一种新的基于 CNN 的目标跟踪方法，该方法包括定位、分类、匹配 3 个模型。在定位模型中，利用前一帧目标的位置信息估计生成若干个当前帧的候选目标区域。在分类模型中，采用已训练的 CNN 特征对候选区域进行类间筛选，排除无关类别物体的干扰。在匹配模型中，利用双向相似匹配(BSM)方法计算候选目标区域与建立的目标模型之间的相似性，从而完成类内寻优过程。与此同时，本文还分别对 3 个模型进行在线更新，保证跟踪的准确性。在 OTB50 和 OTB100[13]标准视频库中开展定性、定量分析，实验结果验证了本文方法的有效性。

## 2　总体框架结构

基于定位-分类-匹配(LCM)模型的目标跟踪方法的整体框架如图 1 所示。在定位模型中，首先利用前一帧目标区域确定当前帧的搜索范围，然后引入 CNN，利用前一帧目标特征信息估计当前帧的目标位置，最后利用估计的位置信息进行随机采样，获得若干个候选目标区域。在分类模型中，采用已训练的 CNN 特征对获得的候选区域进行类间筛选，计算每个区域的分类得分，并选择前 $N$ 个得分高的候选区域作为次优分类结果。在匹配模型中，为了避免相似物体干扰、复杂背景等影响，采用基于颜色特征和距离约束的双向最优相似匹配，确定匹配相似度最高的区域为最优目标区域，从而完成寻优匹配过程。在更新模型中，对 CNN 网络实行长期/短期更新策略，并提出一种置信决策方法对匹配中的目标模型进行在线更新。

## 3　基于 LCM 模型的目标跟踪方法

### 3.1　定位模型

传统的定位网络模型令 $X_t=\{x,y,\delta\}_{i=1}^{M}$ 表示第 $t$ 帧的候选区域，其中，$(x,y)$为坐标值，$\delta$ 表示目标的尺度，$M$ 表示随机采样数。$S(X_t;\tilde{x}_{t-1})$为转换模型，用于当前帧的位置估计，$\tilde{x}_{t-1}$为前一帧的目标位置，$S(\cdot)$的输出结果即为 $M$ 个候选目标区域 $X_t$。根据视频图像序列中目标通常保持平滑移动的性质，可令目标转换模型 $S(\cdot)$服从标准正态分布，即 $S_{\mathrm{norm}}(X_t;\tilde{x}_{t-1})=\mathbf{N}(X_t;\tilde{x}_{t-1},\boldsymbol{\Omega})$，其中，$\boldsymbol{\Omega}$ 为对角协方差矩阵，表示样本位置的方差。

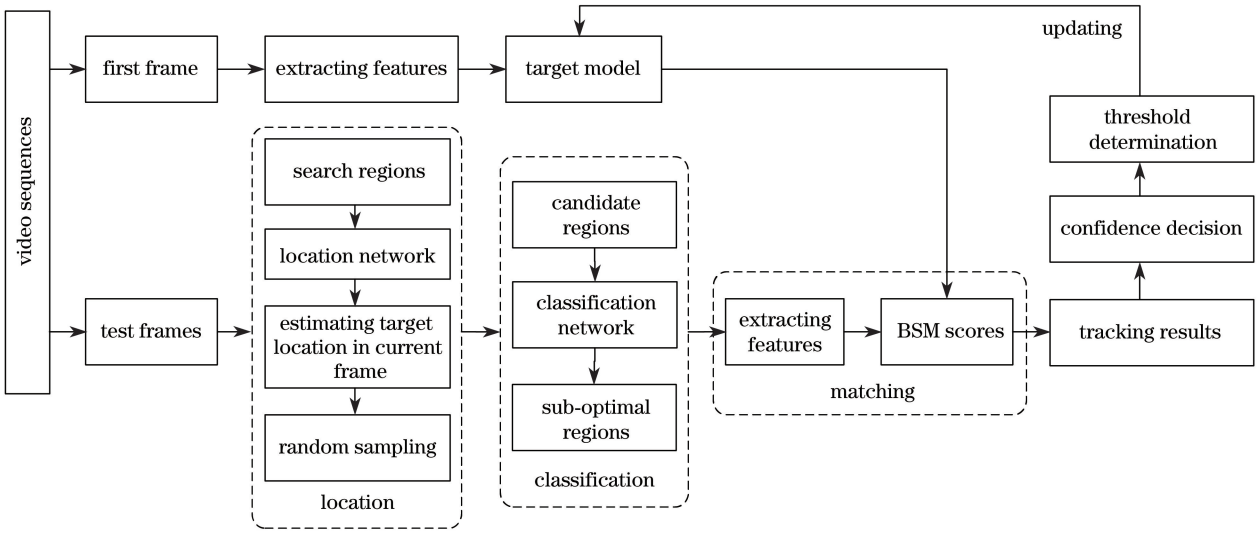


图 1　LCM 模型的整体框架图

Fig. 1　Overall framework of LCM model

然而，当目标快速运动时，基于上述模型的跟踪方法易发生跟踪漂移。为解决此问题，本研究引入一个简单的 2 层 CNN 结构：首先，根据前一帧的目标区域 $\bar{x}_{t-1}$ 确定当前帧的搜索区域，并利用预训练的 CNN 特征表示该区域；然后，利用双卷积层在滑动窗口中计算特征局部区域概率，得到搜索区域的得分；最后，采用得分最大的局部区域作为当前帧的目标位置。文中设置 $\bar{x}_t$ 的矩形框大小与 $\bar{x}_{t-1}$ 相同。与此同时，利用当前帧的估计位置求解中心随机采样 $M$ 个候选目标区域 $X_t$，即 $S_{\text{norm}}(X_t;\bar{x}_t)$。本研究选用 VGG-M[11,14] 特征作为预训练特征，定位模型包括 2 个 1×1 卷积层，输出搜索区域的特征得分，定位过程如图 2 所示。

### 3.2　分类模型

根据定位模型输出的候选区域特征得分，一种典型的分类方法是使用传统的决策模型，如 SVM 或多层感知器（MLP）。这种方法首先将特征图矢量化，然后将其输入一个或多个全连接层，最后再输入到 SVM 或 Softmax 回归层。这种方法往往需要大量的训练参数，且在跟踪的过程中全连接层易发生过拟合。为了解决这一问题，引入 Yang 等[15] 设计的 3 层 CNN 网络对获得的候选区域进行类间筛选，计算每个区域的分类得分，并选择前 $N$ 个得分高的候选区域作为次优分类结果。

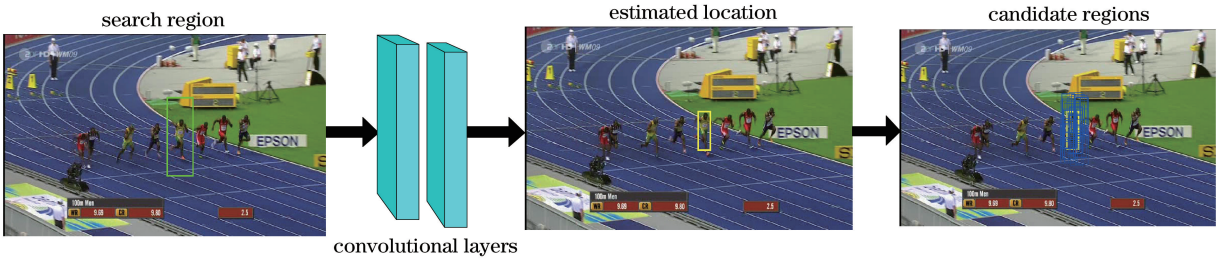


图 2　定位过程示意图

Fig. 2　Diagram of positioning process

为了捕获更多的类别信息，选用 VGG-M 模型的 $\text{ReLU}_3$ 层进行特征提取，输出每个区域的特征映射图，然后送入分类网络。分类网络分为 3 个 1×1 卷积层，输出每个区域的得分，得分越高表示该区域越接近目标区域。虽然分类模型可以较准确地跟踪目标，但当出现相似物体干扰时，这类跟踪方法易发生错误匹配而导致目标丢失。为了弥补分类网络的不足，对生成的各个区域得分采用降序排序，并选择前 $N$ 个得分高的区域进行再匹配。

### 3.3　匹配模型

在匹配模型中，受 Dekel 等[16] 的最优相似匹配对思想启发，改进并提出了一种 BSM 方法，计算前 $N$ 个次优候选区域与目标模型之间的相似性。需要指出的是，目标模型通过提取首帧人工选定的目标区域内的 RGB 颜色特征而建立，并随着跟踪的深入而不断进行在线实时更新。

设目标模型的像素集为 $P=\{p_i\}$，$i\in 1,\cdots,n$，其中，$p_i$ 为集合内的像素点。同理，设候选目标区域集合为 $Q=[Q^1,Q^2,\cdots,Q^N]$，其中，$Q^s=\{q_j^s\}$，$j\in 1,\cdots,m$，表示任意一个候选区域的像素集合，$s\in 1,\cdots,N$。下面以任意一个候选区域 $Q^s$ 为例，说明与目标模型之间匹配相似性的计算过程，即计算 $N_{\mathrm{BSM}}(P,Q^s)$。

$N_{\mathrm{BSM}}(P,Q^s)$ 的核心是计算像素集合中最优匹配对(BMP)个数，最优匹配对的计算方式为

$$N_{\mathrm{BMP}}(p_i,q_j^s,P,Q^s)=\begin{cases}1, & NN(p_i,Q^s)=q_j^s \cap NN(q_j^s,P)=p_i\\ 0, & NN(p_i,Q^s)\neq q_j^s \cup NN(q_j^s,P)\neq p_i\end{cases},\tag{1}$$

式中：$N_{\mathrm{BMP}}(p_i,q_j^s,P,Q^s)$ 为一个二进制函数，若 $p_i$ 与 $q_j^s$ 为最近邻匹配对，则 $N_{\mathrm{BMP}}(\cdot)$ 值为 1，反之，$N_{\mathrm{BMP}}(\cdot)$ 为 0；$\cap$ 操作符表明采用的是双向匹配过程，不仅要求 $p_i$ 与集合 $Q^s$ 之间的最优匹配对为 $q_j^s$，同时要求 $q_j^s$ 与集合 $P$ 之间的最优匹配对为 $p_i$；$NN(p_i,Q^s)$ 表示最近邻匹配距离，具体计算公式为

$$NN(p_i,Q^s)=\arg\min_{q^s\in Q^s} d(p_i,q^s),\tag{2}$$

式中，$d(p_i,q^s)$ 为像素点 $p_i$ 与 $q_j^s$ 之间的匹配距离，表达式为

$$d(p_i,q^s)=\left\|p_i^{(R)}-q^{s(R)}\right\|_2^2+\left\|p_i^{(L)}-q^{s(L)}\right\|_2^2,\tag{3}$$

式中 $R$ 表示 RGB 颜色特征，$L$ 为像素之间的欧氏距离。由此可以计算出集合 $P$ 与 $Q^s$ 之间的 $N_{\mathrm{BSM}}(\cdot)$ 为

$$N_{\mathrm{BSM}}(P,Q^s)=\frac{1}{\min\{n,m\}}\sum_{i=1}^{n}\sum_{j=1}^{m}N_{\mathrm{BMP}}(p_i,q_j^s,P,Q^s)。\tag{4}$$

重复计算 $N$ 次 $N_{\mathrm{BSM}}(P,Q^s)$，可以得到最优匹配区域 $\tilde{Q}(\tilde{Q}\in Q)$，计算表达式为

$$N_{\mathrm{BSM}}(P,\tilde{Q})=\max_{Q^s}[N_{\mathrm{BSM}}(P,Q^s)]。\tag{5}$$

本节以 Bolt 视频序列为例，说明相似度匹配的稳健性，匹配过程如图 3 所示。Bolt 视频序列涵盖目标形变、快速运动及复杂背景等影响因素。从图 3可以看出，BSM 采用双向校验的方式进行匹配，不受目标先验形状、模型结构的约束，只利用双向匹配对个数进行统计计算，避免了目标形变、快速运动等因素的影响，具有较高的匹配准确率。

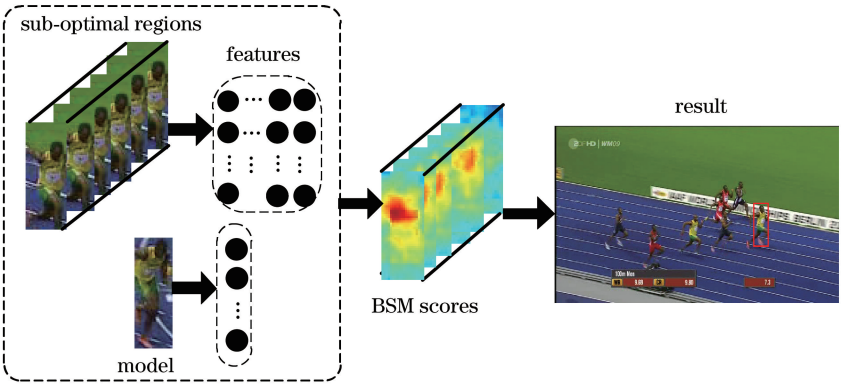


图 3　相似度匹配过程

Fig. 3　Similarity matching process

### 3.4　模型的更新

在更新过程中，采用不同的策略更新网络模型和匹配模型。

更新网络模型：与其他 CNN 跟踪器一样，利用长期/短期($T_{\mathrm{long}}/T_{\mathrm{short}}$)策略进行网络更新[12,15]。类似地，网络中的消极样本始终在 $T_{\mathrm{short}}$ 周期内进行采集，积极样本则各自从 $T_{\mathrm{long}}$、$T_{\mathrm{short}}$ 周期采集，其中，$T_{\mathrm{long}}$ 为常数(在实验中取 20)，$T_{\mathrm{short}}$ 则由 $N_{\mathrm{BSM}}(\cdot)$ 的值决定，即设置一个阈值来决定是否执行短期 $T_{\mathrm{short}}$ 更新。若 $N_{\mathrm{BSM}}(\cdot)$ 小于 $\theta$，则执行短期更新(在实验中设 $\theta=0.6$)。

更新匹配模型：提出一种置信决策方法，类似于网络的短期更新策略，利用 $N_{\mathrm{BSM}}(\cdot)$ 的值决定是否更新目标模型。具体过程为：若 $N_{\mathrm{BSM}}(\cdot)$ 大于等于 0.6，则认为当前帧的匹配结果较为准确，目标模型无须更新；若 $N_{\mathrm{BSM}}(\cdot)$ 取值范围为 0.3～0.6，则对目标模型进行更新；若 $N_{\mathrm{BSM}}(\cdot)$ 小于等于 0.3，则认为当前帧发生了误匹配，在实验中发现，此时更新匹配模型会使后续帧中发生较大位置的跟踪偏移，因此在此情况下不更新目标模型。更新模型的公式为

$$P=\lambda P+(1-\lambda)\tilde{Q},\tag{6}$$

式中 $\tilde{Q}$ 为当前帧中匹配到的目标区域特征分布，$\lambda$ 为更新权重(在实验中取 0.7)。

## 4　实验结果及分析

实验均在 CPU 为 Intel Core i7-6700，3.4 GHz

主频和 16 GB 内存的台式机上进行的，测试开发平台为 MATLAB R2016a、Visual Studio 2015。本研究选取的标准测试集为 OTB50 与 OTB100，主要考虑其客观性和权威性，便于与优秀的算法进行对比分析。首先，详细说明实验的参数设置，以进一步了解算法跟踪的实现过程。然后为了体现本研究方法的整体稳健性，将各个模型进行拆分重组，并在 OTB50 上进行验证说明。最后选用近几年较为流行的目标跟踪方法在 OTB50 和 OTB100 上进行测试分析，说明本研究方法的有效性。

### 4.1　参数设置

在定位模型中，设定当前帧中的搜索区域是前一帧目标区域的 3 倍(区域面积)。为了得到覆盖更为广泛、全面的候选区域，根据当前帧估计位置，随机采样 256 个大小相同的区域作为候选目标区域，即 $M=256$，在分类模型中，设 $N=10$，即选择前 10 个候选区域进行目标模型匹配。对于正态分布 $N(\cdot)$的对角协方差阵，设 $\boldsymbol{\Omega}=(0.09r^2 \quad 0.09r^2 \quad 0.25)$，其中，$r$ 为跟踪得到目标区域大小的平均值。定位网络：第一层包括 96 个卷积核，第二层包括 256 个卷积核，即输出 256 个搜索区域的特征得分。分类网络：第一层包括 512 个滤波器，第二层包括 128 个滤波器，最后一层输出各个搜索区域的类间得分。首帧设置的网络训练最大迭代次数为 10，在线更新迭代次数为 5。

### 4.2　LCM 跟踪性能分析

为了说明本研究方法的有效性，本节将关键的三个模型(定位模型、分类模型及匹配模型)进行拆分重组，分别设计三种跟踪方法：无定位模型(Alg1)，无分类模型(Alg2)，无匹配模型(Alg3)。然后，将这三种跟踪方法与本研究方法在 OTB50 标准视频库上进行验证，证明本研究方法的稳健性。选择的对比属性为光照变化(IV)、尺度变化(SV)、遮挡(OCC)、形变(DEF)、运动模糊(MB)、快速运动(FM)、背景杂波(BC)及低分辨率(LR)。评价指标为距离精确度(DP，阈值设定为 20 pixel)、跟踪重叠率(OR，重叠阈值为 0.5)及运行速度(RS)。

#### 4.2.1　准确性分析

本研究方法与 Alg1、Alg2 及 Alg3 方法在不同属性下的 DP 与 OR 如表 1 所示。同时，选取 5 组具有代表性的图像序列进行跟踪结果展示，其跟踪效果如图 4 所示。

表 1　4 种跟踪方法在不同属性下的 DP 与 OR

Table 1　DP and OR of 4 tracking methods under different attributes　%

| Attribute | Alg1 | | Alg2 | | Alg3 | | LCM | |
|---|---|---|---|---|---|---|---|---|
| | DP | OR | DP | OR | DP | OR | DP | OR |
| IV | 86.2 | 81.4 | 73.8 | 68.5 | 88.7 | 83.2 | 90.5 | 84.6 |
| SV | 82.4 | 73.8 | 70.2 | 62.1 | 88.6 | 79.4 | 88.2 | 78.7 |
| OCC | 81.6 | 76.5 | 67.6 | 60.4 | 84.2 | 78.6 | 91.4 | 81.5 |
| DEF | 86.1 | 77.9 | 68.4 | 60.8 | 88.3 | 79.6 | 89.7 | 82.1 |
| MB | 84.3 | 82.6 | 71.4 | 69.5 | 88.9 | 85.2 | 89.6 | 86.3 |
| FM | 83.8 | 78.6 | 67.5 | 60.8 | 88.7 | 81.4 | 89.8 | 82.7 |
| BC | 87.8 | 82.6 | 75.7 | 68.5 | 89.9 | 83.0 | 91.3 | 84.4 |
| LR | 77.6 | 67.8 | 57.8 | 50.9 | 89.6 | 78.6 | 94.1 | 80.5 |
| Average | 83.7 | 77.6 | 69.0 | 62.6 | 88.3 | 81.1 | 90.5 | 82.6 |

从表 1 可以看出，本研究提出的 LCM 模型在不同属性下的 DP 与 OR 均优于其他跟踪器，按照优劣排列顺序为 LCM、Alg3、Alg1、Alg2。从 Alg1 与 LCM 的对比可以看出，加入定位模型的有效性，在 8 个属性下，LCM 模型的 DP 提高 7%左右，OR 提高 5%左右。同样地，LCM 模型与 Alg2 相比可以验证加入分类模型的重要性，在 8 个属性下，LCM 模型的 DP 提高 21%左右，OR 提高 20%左右。与 Alg3 相比，LCM 模型的 DP 提高 2%左右，OR 提高 1.5%左右。由于 Alg3 本身的平均 DP 和 OR 已经较高，因此加入匹配模型并未体现得十分明显，但也足以说明其重要性。

LCM 模型在 8 种属性下的平均 DP 和 OR 分别为 90.5%，82.6%。从单个属性下的跟踪结果能够分析出本研究方法的稳健性。例如，在遮挡的条件下，LCM 模型的 DP、OR 分别为 91.4%，81.5%，由于使用 BSM 进行相似匹配，并对结果进行双向校验，因此 LCM 方法可以解决该条件下的误匹配问题，从而获得更精确的跟踪结果。在目标形变的情况下，LCM 模型的 DP、OR 分别为 89.7%，

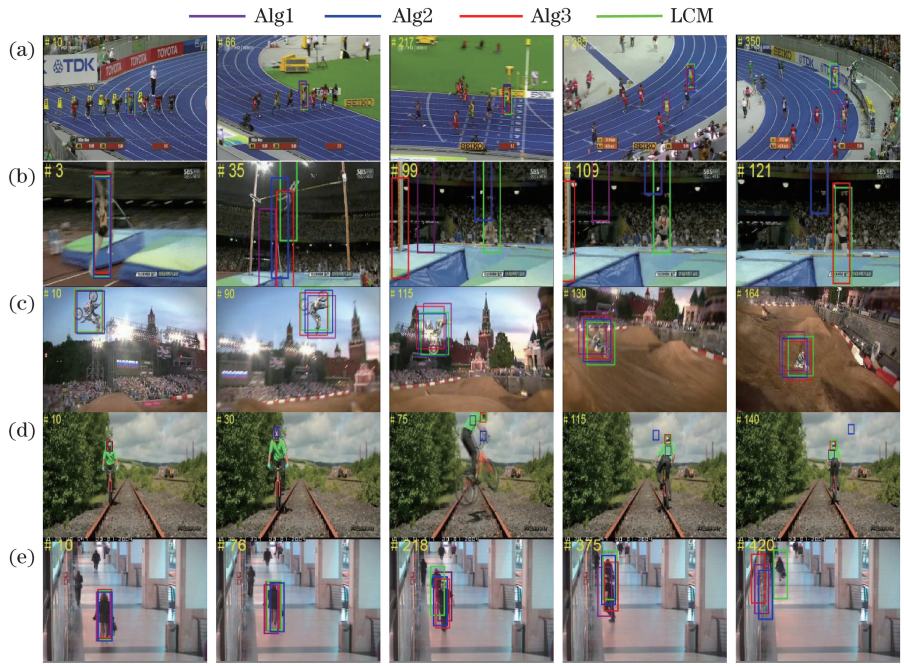


图 4　4 种跟踪方法在 5 组序列中的跟踪结果。(a) Bolt 序列;(b) jump 序列;(c) motorrolling 序列;(d) biker 序列;(e) walking2 序列

Fig. 4　Tracking results of the 4 tracking methods in the 5 sequences. (a) Bolt sequences; (b) jump sequences; (c) motor rolling sequences; (d) biker sequences; (e) walking2 sequences

82.1%。由于研究采用已训练的 CNN 特征进行定位、分类,从 Alg3 可以看出此类方法自身的稳健性,而本研究还加入匹配模型使得 LCM 的跟踪更为可靠。

4.2.2　时效性分析

4 种方法在不同属性下的运行速度如表 2 所示。从表 2 可以看出,本研究提出的 LCM 方法的运行速度维持在 2.6 frame/s 左右,具体的排列顺序为 LCM<Alg1<Alg3<Alg2。LCM 与 Alg1 两种方法的跟踪速度相当,这说明定位模型的运行效率较高,并未因为模型的加入而增加跟踪的时间复杂度。与 Alg3 相比,匹配模型的加入使得 LCM 的整体跟踪运行速度略有降低。与 Alg2 相比,虽然在运行速度上分类模型的时间复杂度较高,但从表 1 中的两个指标可以看出该模型的重要性。综合表 1 和表 2,虽然 LCM 的跟踪准确性较高,但算法的时效性略低,因此在今后的工作中,将重点对分类模型的结构、模型的更新策略的选取及模型之间的协同关系作进一步的研究,提供跟踪方法的整体稳健性。

表 2　4 种跟踪方法在不同属性下的 RS

Table 2　RS of 4 tracking methods in different attributes　frame/s

| Attribute | Alg1 | Alg2 | Alg3 | LCM |
|---|---|---|---|---|
| IV | 2.74 | 64.55 | 3.48 | 2.68 |
| SV | 2.78 | 59.32 | 3.27 | 2.43 |
| OCC | 2.63 | 62.24 | 3.56 | 2.58 |
| DEF | 2.59 | 64.58 | 3.44 | 2.47 |
| MB | 3.22 | 69.44 | 4.06 | 2.81 |
| FM | 2.84 | 67.43 | 3.69 | 2.76 |
| BC | 2.61 | 68.72 | 3.14 | 2.39 |
| LR | 3.13 | 69.46 | 3.57 | 2.66 |
| Average | 2.82 | 65.72 | 3.53 | 2.60 |

### 4.3　对比实验分析

选择被广泛使用的曲线评价方法 OPE(one-pass evaluation) 来客观分析 LCM 的准确性和稳健性。借鉴 Wu 等[7,13]在 OTB 标准库中提供的两大客观性指标:一是精确度图,表示运行跟踪方法得到的目标中心位置与标准库中的实际位置之间的平均欧式距离,用于评价跟踪方法的准确性,精确度图能够清晰地显示出在给定的阈值范围内的准确跟踪帧数占视频总帧数的百分比,在实验中使用的阈值为 20 pixel;二是成功率图,即位置矩形框的覆盖重叠率,设跟踪的矩形框为 $c_t$,实际的矩形框为 $c_a$,重叠率表示为 $S=|c_t \cap c_a|/|c_t \cup c_a|$,为评估跟踪方法的性能,计算重叠率 $S$ 大于指定阈值的帧数,成功率图则显示出阈值在[0,1]之间变化的跟踪成功帧

的百分比。在评估过程中，利用曲线的面积（AUC）对跟踪方法做综合排名。本小节选取近年来表现优秀的跟踪方法与本研究方法进行对比分析，具体方法有 MDNet[12]、MUSTer（multi-store tracker）[17]、CNN-SVM[10]、MEEM（multiple experts using entropy minimization）[18]、TGPR（tracking with Gaussian processes regression）[19]、DSST（discriminative scale space tracker）[20]、KCF（kernelized correlation filters）[21]，下面分别从 OTB50 和 OTB100 两个标准库中的实验结果进行说明分析。

4.3.1　OTB50

OTB50 包括 50 个人工标定的视频序列，涵盖了目标形变、遮挡及背景杂波等挑战性因素，该基准库受到国内外专家学者的一致认可，并得到了广泛的应用。本研究基于该标准库进行测试，所提出的 LCM 模型与其他 7 种优秀的跟踪方法在精确度和成功率上的稳健性评估结果如图 5 所示。从图 5 可以看出，本研究方法在精确度和成功率上与 MDNet 方法相当，精确度仅下降 3%左右，成功率也有 1%左右的降低，但优于其他 6 种跟踪方法。例如与 CNN-SVM 方法相比，LCM 模型的精确度和成功率分别有 4%和 5%的提高，这说明本研究引入的基于 CNN 网络结构的定位、分类模型的有效性，以及设计匹配模型的必要性。综上所述，本研究提出的 LCM 方法在跟踪成功率和精确性方面显示出良好的性能。

4.3.2　OTB100

OTB100 是在 OTB50 的基础上得到的，标准视频库拓展到 100 组，并且涵盖更为复杂的挑战性因素。本节基于此标准库进行测试，所提出的 LCM 与其他 7 种优秀的跟踪方法在精确度和成功率上的稳健性评估结果如图 6 所示。与 OTB50 的标准库验证相似，本研究方法在精确度和成功率上与 MDNet 方法相当，精确度仅降低 1%左右，成功率也有 1%左右的降低。由于面对更加全面、覆盖面更广的 OTB100，MDNet 方法精度下降了 4%，成功率下降了 3%，而本研究提出的 LCM 的精度和成功率都较为稳定，均下降 1%左右，这说明 LCM 的运行稳定性较好。与其他 6 种跟踪方法相比，LCM 在精确度和成功率上均有一定优势，与 MUSTer、CNN-SVM 相比，精确度分别提高了 11%和 8%，成功率分别提高了 9%和 11%。在此标准库中的实验

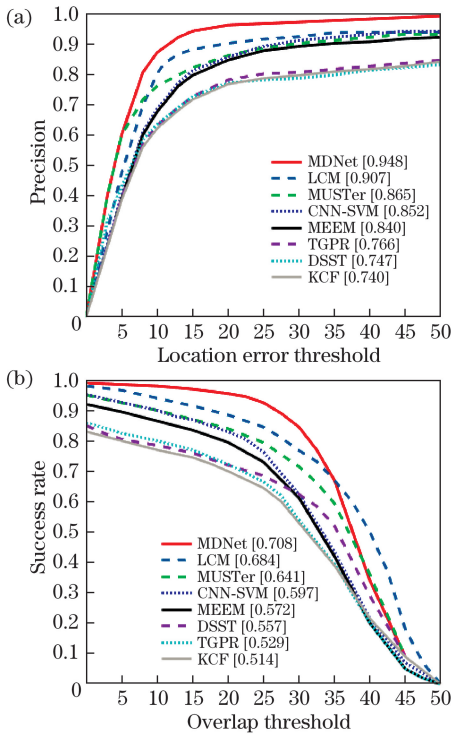


图 5　8 种跟踪方法在 OTB50 上的精确度和成功率。（a）OPE 的精确度图；（b）OPE 的成功率图

Fig. 5　Precision and success plots of 8 tracking methods on OTB50. (a) Precision plots of OPE; (b) success plots of OPE

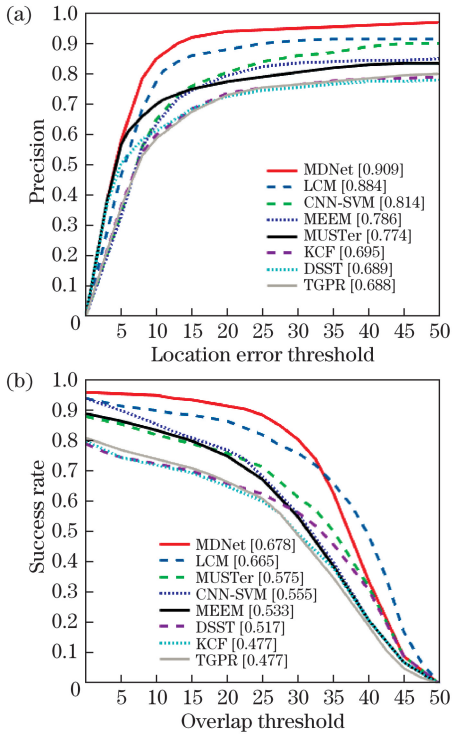


图 6　8 种跟踪方法在 OTB100 上的精确度和成功率。（a）OPE 的精确度图；（b）OPE 的成功率图

Fig. 6　Precision and success plot of 8 trackers on OTB100. (a) Precision plots of OPE; (b) success plots of OPE

结果进一步说明 LCM 方法在跟踪成功率和精确性方面显示出良好的性能且整体稳定性较高。

本研究方法与其他跟踪方法不同之处在于：1)在定位模型中，利用目标在相邻帧的位置相关性估计采样当前帧的目标位置信息，保证候选样本的全面性；2)深度特征与手动提取特征结合，利用双向相似匹配方法对类内目标进行相似匹配，跟踪结果更稳健；3)提出一种在线匹配模型更新算法，引入置信决策方法，保证模型对目标的描述更充分。

本研究将深度特征与人工手动提取特征相融合并应用到目标跟踪中，上述实验证明了其有效性。然而，与 MDNet 方法相比，本研究提出的 LCM 方法仍有不足之处，特别是在匹配的过程中，出现相似颜色特征物体干扰时 LCM 方法的跟踪不理想。因此在今后的研究工作中，本课题组将重点研究将其他手动提取特征[如纹理特征、梯度方向直方图(HOG)特征以及 Haar 等]与深度特征相结合以实现更稳健的跟踪。

## 5 结 论

本文提出了一种基于卷积神经网络的目标跟踪方法，由定位、分类、匹配三个模块组成。在定位模型中，采用前一帧的跟踪结果估计预测当前帧的目标位置，并根据目标位置进行采样，获得若干目标候选区域；在分类模型中，采用已训练的深度特征对获得的候选区域进行类间筛选，计算每个区域的分类分数，并选择前 $N$ 个作为次优候选区域；在匹配过程中，利用 BSM 计算次优区域和目标模型之间的相似性，从而得出当前帧中最优目标区域。在 OTB50 和 OTB100 视频库中与近几年的跟踪方法相比，所提出的跟踪方法显示出良好的跟踪性能。

## 参 考 文 献


[1] Vatavu A, Danescu R, Nedevschi S. Stereovision-based multiple object tracking in traffic scenarios using free-form obstacle delimiters and particle filters[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 498-511.

[2] Zhao G P, Shen Y P, Wang J Y. Adaptive feature fusion object tracking based on circulant structure with kernel[J]. Acta Optica Sinica, 2017, 37(8): 0815001.
赵高鹏, 沈玉鹏, 王建宇. 基于核循环结构的自适应特征融合目标跟踪[J]. 光学学报, 2017, 37(8): 0815001.

[3] Cai Y Z, Yang D D, Mao N, *et al*. Visual tracking algorithm based on adaptive convolutional features[J]. Acta Optica Sinica, 2017, 37(3): 0315002.
蔡玉柱, 杨德东, 毛宁, 等. 基于自适应卷积特征的目标跟踪算法[J]. 光学学报, 2017, 37(3): 0315002.

[4] Zhang B, Long H. Visual target tracking algorithm based on image signature algorithm[J]. Laser & Optoelectronics Progress, 2017, 54(9): 091504.
张博, 龙慧. 基于图像签名算法的视觉目标跟踪算法[J]. 激光与光电子学进展, 2017, 54(9): 091504.

[5] Lin S Z, Zheng Y, Lu X F, *et al*. Adaptive tracking algorithm for aerial small targets based on multi-domain convolutional neural networks and autoregression model[J]. Acta Optica Sinica, 2017, 37(12): 1215006.
蔺素珍, 郑瑶, 禄晓飞, 等. 基于多域卷积神经网络与自回归模型的空中小目标自适应跟踪方法[J]. 光学学报, 2017, 37(12): 1215006.

[6] Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking[C]// Proceedings of the 26th International Conference and Workshop on Neural Information Processing Systems, December 5-10, 2013, Lake Tahoe, Nevada. New York: ACM, 2013, 1: 809-817.

[7] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. 2013: 2411-2418.

[8] Wang N Y, Li S Y, Gupta A, *et al*. Transferring rich feature hierarchies for robust visual tracking[J]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[9] Ma C, Huang J B, Yang X K, *et al*. Hierarchical convolutional features for visual tracking[C]// IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3074-3082.

[10] Hong S, You T, Kwak S, *et al*. Online tracking by learning discriminative saliency map with convolutional neural network[C]// Proceedings of the 32nd International Conference on Machine Learning, July 6-11, 2015, Lille, France. JMLR. org, 2015: 597-606.

[11] Wang L J, Ouyang W L, Wang X G, *et al*. Visual tracking with fully convolutional networks[C]// 2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3119-3127.

[12] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]// 2016 IEEE Conference on Computer Vision and Pattern

Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4293-4302.

[13] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[14] Chatfield K, Simonyan K, Vedaldi A, *et al*. Return of the devil in the details: delving deep into convolutional nets[C]// British Machine Vision Conference, 2014.

[15] Yang L X, Liu R S, Zhang D, *et al*. Deep location-specific tracking[C]// Proceedings of the 2017 ACM on Multimedia Conference, October 23-27, 2017, Mountain View, California, USA. New York: ACM, 2017: 1309-1317.

[16] Dekel T, Oron S, Rubinstein M, *et al*. Best buddies similarity for robust template matching[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 2021-2029.

[17] Hong Z, Chen Z, Wang C, *et al*. Multi-store tracker (MUSTer): a cognitive psychology inspired approach to object tracking[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 749-758.

[18] Zhang J M, Ma S G, Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization[C]// European Conference on Computer Vision 2014, September 6-12, 2014, Zurich, Switzerland. Cham: Springer, 2014, 8694: 188-203.

[19] Gao J, Ling H B, Hu W M, *et al*. Transfer learning based visual tracking with Gaussian processes regression[C]// European Conference on Computer Vision 2014, September 6-12, 2014, Zurich, Switzerland. Cham: Springer, 2014, 8691: 188-203.

[20] Danelljan M, Häger G, Shahbaz K F, *et al*. Accurate scale estimation for robust visual tracking[C]// Proceedings of the British Machine Vision Conference, 2014. BMVA Press, 2014.

[21] 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.