# 结合连续卷积算子的自适应加权目标跟踪算法

罗会兰　石武

江西理工大学信息工程学院，赣州， 341000

**摘 要：目的** 在视觉跟踪领域中，特征的高效表达是鲁棒跟踪的关键，观察到在相关滤波跟踪中，不同卷积层表达了目标的不同方面特征，提出了一种结合连续卷积算子的自适应加权目标跟踪算法。**方法** 针对目标定位不准确的问题，提出连续卷积算子方法，将离散的位置估计转换成连续位置估计，使得位置定位更加准确；利用不同卷积层的特征表达，提高跟踪效果。首先利用深度卷积网络结构提取多层卷积特征，通过计算相关卷积响应大小，决定在下一帧特征融合时各层特征所占的权重，凸显优势特征，然后使用从不同层训练得到的相关滤波器与提取得到的特征进行相关运算，得到最终的响应图，响应图中最大值所在的位置便是目标所在的位置和尺度。**结果** 与目前较流行的 3 种目标跟踪算法在目标跟踪基准数据库（OTB-2013）中的 50 组视频序列进行测试，本文提出的算法平均跟踪成功率达到 85.4%。**结论** 本文算法在光照变化，尺度变化，背景杂波，目标旋转、遮挡和复杂环境下的跟踪具有较高的鲁棒性。

**关键词**：目标跟踪；相关滤波跟踪；连续卷积算子；自适应加权；卷积特征；响应图

# An adaptive weighted object tracking algorithm with continuous convolution operator

Luo Huilan, Shi Wu

*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China*

**Abstract**: **Objective** In the field of visual tracking, efficient representation of features is the key to robust tracking. It is observed that different convolution layers represent different aspects of the target in correlation filter tracking. An adaptive weighted object tracking algorithm with continuous convolution operator is proposed. **Method** Aiming at the problem of inaccurate target location, a continuous convolution operator method is proposed to convert discrete position estimates into continuous position estimates, which makes position location more accurate. The feature representations of different convolution layers are leveraged to improve tracking effect. Different convolutional layer features in deep convolutional neural networks have different expression ability, that is, shallow features have more positional information, while deep features have more semantic features. Therefore, if feature expression and tracking can be carried out by combining them, better tracking effect can be obtained than using only deep or shallow features. Firstly, the multi-layer convolution features are extracted by using the deep convolution network structure, and the weight of each layer features in the fusion features in the next frame is determined by calculating the correlation convolution response, so as to highlight the dominant features and make the target more distinguishable from the background or distractor. Then, the correlation filter trained from different layers is used to perform correlation operation with the extracted features to obtain the final response map. The position of the maximum value in the response map is used to calculate the position and scale of the target. The weights of different convolutional feature layers are adaptively updated through the correlation filtering tracking effect of different convolutional layers, the feature expression ability of different convolutional layers in the convolutional neural network is fully exerted, and the expression scheme is adaptively adjusted according to the different environmental conditions of each frame to improve the tracking performance. **Result** Compared with three state-of-the-art tracking algorithms in 50 video sequences of object tracking benchmark (OTB-2013) dataset, the average success rate of the proposed algorithm is 85.4%. **Conclusion** Experimental results show that the proposed tracking algorithm has good performance

**基金项目**：国家自然科学基金（61462035, 61862031），江西省青年科学家培养项目（20153BCB23010），江西省自然科学资金项目（20171BAB202014）

**Supportred by:** National Natural Science Foundation of China(61462035, 61862031),The Young Scientist Training Project of Jiangxi Province(20153BCB23010),The Natural Science Foundation of Jiangxi Province of China(20171BAB202014)

and can track successfully and efficiently for many complicated situations, such as illumination variation, scale variation, background clutters, object rotation and occlusion.

**Key words:** object tracking; correlation filter tracking; continuous convolution operator; adaptive weighted; convolution features; response map

# 0 引 言

目标跟踪一直都是计算机视觉领域研究的基本问题之一，已经广泛应用于智能控制（如无人机、机器人等），人机交互，自动驾驶[1]等领域。目标跟踪是给定目标第一帧的初始状态（通常是位置和尺度大小）的情况下，并在后续视频序列中估计出目标状态的过程。随着深度学习的出现，视觉目标跟踪技术取得了重大进步和突破性进展。但由于受到目标的快速运动、旋转、外观变化、光照变化、尺度变化、相似背景干扰以及遮挡等一种或者多种因素的影响，高效准确的跟踪仍然极具挑战性。

视觉目标跟踪方法大致可以分为两类，一类是产生式方法，另一类是判别式方法。产生式方法运用学习得到的目标模型描述目标的外观特征，然后在候选目标中寻找与模型最相似的区域作为目标，比较有代表性的算法有基于稀疏表示的目标跟踪算法[2-4]和基于线性子空间的目标跟踪算法[5]。产生式方法突出目标本身的信息却忽略了背景信息，导致在目标自身发生变化或者被遮挡时容易产生漂移[6]。

判别式方法则是通过训练数据学习到一个分类器来区分目标和背景。以目标区域为正样本，背景区域为负样本，进行模型的训练，最高分类器分数所在的候选位置就是目标的位置，这种方法也被称为检测跟踪方法。其中有代表性的算法有基于分类跟踪的深度学习方法[7, 8]和基于支持向量机的跟踪算法[9, 10]等。

在信号处理领域中，可以用相关性来表示两个信号的相似程度。通常情况下，相关性计算使用卷积来实现。在 2010 年，Bolme 等人[11, 12]首次将相关滤波应用于目标跟踪领域，利用卷积定理和快速傅立叶变换的性质，通过在频域中最小化期望响应和滤波器与目标区域的循环相关之间的均方误差之和，得到误差最小平方和滤波器（MOSSE）。由于 MOSSE 跟踪器很慢并且不能准确地估计目标的尺度，Danelljan 等人[12]提出一种快速准确的自适应尺度相关滤波跟踪器，该算法用梯度直方图特征代替灰度特征，利用多尺度搜索的方法估计目标的尺度，提高了跟踪性能。文献[13-16]利用已经预训练好的深度卷积网络模型提取特征，结合了高效鲁棒的深度特征和相关滤波算法，取得了很好的跟踪效果。Ma 等人[13]提出了融合多层卷积特征的相关滤波(HCFT)跟踪算法，融合多层卷积特征，提升了算法的跟踪性能。文献[17]与文献[13]类似，采用多层卷积特征和相关滤波的方法，由原来的三层卷积特征变成六层卷积特征，将搜索区域的六层卷积特征输入到对应的相关滤波器，得到六个响应图，每个响应图有一个最大点位置，每个最大点位置乘以相对应的自适应权重得到目标新位置。Danelljan 等人[15]提出一种连续域卷积相关滤波（CCOT）跟踪器，将时域离散的位置估计转换到连续域上，使位置估计更准确，并且能解决融合不同分辨率特征的问题，实现了传统特征与深度特征的融合。He 等人[16]在文献[15]的基础上，分析了第一和五卷积层的特性，分配第一和五层特征固定权重,两层卷积响应相加以产生最终响度图，跟踪性能得到进一步的提升。

近年来，许多学者提出了各种具有特定结构的神经网络用于目标跟踪。在相关滤波过程中，既要保存滤波器信息，又要提取特征。孪生网络（Siamese Network）的一条网络支路保存滤波器信息，另一条网络支路提取特征，然后把滤波器与特征进行相关操作，得到响应图，根据响应图中最大值位置判断目标状态。Tao 等人[18]应用 Siamese 学习特征进行目标跟踪，利用大量的视频帧学习一个匹配函数，通过后续的视频帧和第一帧匹配，达到跟踪的目的。然而文献[18]中的方法需要候选评估，这个过程很耗时，因此，Held 等人[19]在此基础上提出了一个卷积神经网络模型，直接学习预测目标相对于参考目标的相对位置，避免了候选评估和特征匹配阶段。另一种深度网络结构是递归神经网络（RNN），Cui 等人[20]提出一种循环目标强化跟踪算法，利用 RNN 获取响应图，响应图在目标部分区域具有较高的值，将其作为相关滤波器的系数，增强相关滤波器在跟踪过程的抗干扰能力。Fan 等人[21]提出一种用于目标跟踪的结构感知网络，网络利用 CNN 学习分辨目标物体与背景，利用 RNN 学习分辨目标物体与相似物体，使用跳跃式链接策略获取多层 CNN 特征和 RNN 特征并进行融合，以此提高跟踪器的判别能力。

受到文献[15]和文献[17]的启发，本文在相关滤波跟踪算法的基础上，结合文献[15]和文献[17]的思想，提出了一种结合连续卷积算子的自适应加权的目标跟踪方法，基于不同卷积层特征分别训练滤波器，通过计算特征响应值的大小自适应地分配下一帧各自特征的权重，凸显优势特征，使得目标与背景或干扰物更具有区分度。与文献[17]不同的是，本文利用了连续卷积算子将离散的位置估计转换成连续位置估计，使得位置估计更加准确。深度卷积神经网络中不同卷积层特征具有不同的表达特点，即浅层特征具有更多的位置信息，而深层特征具有更多的语义特征，所以如果能结合它们进行特征表达和跟踪，会得到相较于只利用深层或浅层特征更好的跟踪效果。与文献[15]利用 VGG-M[22]提取的多层卷积特征进行线性均值融合不同的是，本文通过不同卷积层相关滤波跟踪效果自适应更新不同卷积特征层的权重，充分发挥卷积神经网络中不同卷积层特征表达能力，根据每帧的不同环境情况，自适应调整表达方案，提高跟踪性能。本文的第二小节介绍了相关工作，第三小节详细论述了本文提出的连续卷积算子和自适应融合多层卷积特征方法，第四小节对本文提出的方法进行了实验分析比较，最后是结论。

# 1 相关工作

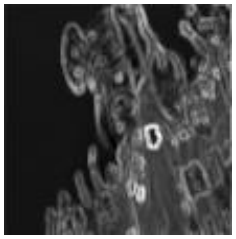
## 1.1 分层卷积特征

卷积神经网络（CNNs）[23, 24]是近年来一种非常典型的深度学习架构，能够学习到平移 、旋转和形变等不变性特征。许多卷积神经网络模型已经成功应用到图像分类和目标检测等中，如 AlexNet[25]、VGG-Net[22]和 ResNet[26]。VGG-Net[22]在 ILSVRC-2014[27]中获得定位任务第一名和分类任务第二名，其突出贡献在于证明使用很小的卷积（3*3），和增加网络深度可以有效提升模型的识别效果，而且 VGG-Net[22]对其他数据集具有很好的泛化能力。由于用于目标跟踪的评估基准数据集和实际应用中的跟踪视频分辨率都较低，适合采用层数较少的小型卷积神经网络[28]，既可以减少图片信息损失，也可以提高计算速度。目前，很多深度目标跟踪算法[8, 15, 16, 29, 30]都采用 VGG-M[22]网络提取特征。VGG-M[22]网络是一种小型神经网络，由 5 个卷积层和 3 个全连接层组成。

图 1 为 VGG-M 模型[22]的不同卷积层特征的可视化表示，从图 1 可以看出，层次越深，卷积特征包

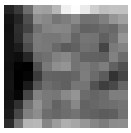
含目标的语义信息越多，这有利于区分目标跟踪过程中不同的对象，但是关于目标的位置空间信息更少。层次越浅，卷积特征保留的空间信息更多，比如目标的位置和尺寸信息，这对目标跟踪的准确定位非常重要。因此，为了能更好运用卷积神经网络的卷积层特征，目前已经有很多学者将多层卷积层特征的融合应用到视觉目标跟踪领域，并且取得了很好跟踪结果。如文献[13]利用卷积神经网络的分层卷积特征，提出了一种由粗到细（Coarse-to-fine）的跟踪框架,融合不同层的特征，提高算法的跟踪精度。文献[14]从大规模分类任务上学习到 CNN 不同层具有不同的特性，通过融合两个不同卷积层，很好缓和了漂移问题。

(a) 输入图像，称为 layer0

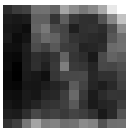


(b)layer1

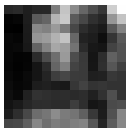
(c) layer2



(d) layer3



(e) layer4



(f) layer

图 1　VGG-M 模型的不同卷积层特征的可视化：

(a)输入图像，称为 layer0；(b) layer1；(c) layer2;(d) layer3；(e) layer4；(f) layer5

Fig.1　Visualization of different convolution features of VGG-M model

(a)input image，called layer0；(b) layer1；(c) layer2;(d) layer3；(e) layer4；(f) layer5

### 1.2 相关滤波跟踪

相关滤波跟踪使用目标位置图像块训练得到滤波器，然后对图像进行滤波处理，在响应图中最大值所在的位置即是目标所在的位置。故可以把相关滤波目标跟踪的过程近似地等效看成是对搜索区域图像块进行相关滤波的过程，寻找目标所在的位置即是寻找滤波器响应图的最大值位置。相关滤波跟踪[13-17, 31]过程概括如下：

1） 将第一帧上给定的目标位置的图像块，作为训练样本，通过最小化损失函数训练得到相关滤波器。

2） 在后续的每一帧，根据前一帧的预测目标位置提取新的图像块作为候选图像块。用预训练好的卷积神经网络从当前帧候选图像块中提取特征，并用余弦窗弱化图像边界对跟踪结果的影响。

3） 对用余弦窗处理过的特征与学习到的相关滤波器进行相关滤波操作。

4） 寻找相关滤波操作响应图的最大值点，响应图最大值所在的位置即是目标的位置。

5） 然后提取预测位置的特征，通过最小化损失函数更新相关滤波器，完成一次跟踪。

## 2. 结合连续卷积算子的自适应加权目标跟踪算法

本文提出的结合连续卷积算子的自适应加权目标跟踪算法的结构示意图如图 2 所示。利用视频第一帧，通过深度卷积网络提取多层卷积特征，有监督训练对应各层的相关滤波器。在目标跟踪阶段，将多层特征输入到对应的相关滤波器，得到多个响应图，根据响应图输出，自适应地决定在下一帧特征融合时各个响应图所占的权重。多个响应图通过自适应得到的权重进行加权求和，得到最终的响应图，通过搜索响应图最大值位置即可以确定跟踪目标位置和尺度。

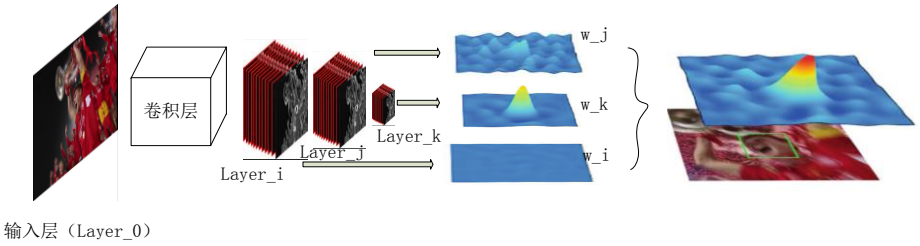



图 2 结合连续卷积算子的自适应加权目标跟踪算法的结构示意图

Fig. 2 The structural schematic diagram of the proposed adaptive weighted object tracking algorithm

### 2.1 连续卷积算子

本文首先利用三次样条插值函数将时域离散特征图转换为时域连续特征图，然后利用连续卷积算子将学习到的连续相关滤波器和连续特征图进行相关滤波，使得目标位置估计更加准确。

假设训练样本 $x$ 在第 $i$ 个特征层有 $D_i$ 个特征通道，$x^d$ 表示训练样本 $x$ 的第 $i$ 卷积特征层的第 $d$ 个特征图，$N_d$ 表示第 $d$ 个特征图的空间像素样本的数量，$T$ 表示插值之后特征图的大小，连续区间 $t\in[0,T)$，对于第 $d$ 个特征图，连续插值运算定义如式（1）所示。

$$J\{x^d\}(t)=\sum_{n=0}^{N_d-1}x^d[n]b\left(t-\frac{T}{N_d}n\right) \quad (1)$$

其中，$J\{x^d\}$表示插值后的特征图，$b$ 是三次样条插值函数，在每两个像素位置间进行插值，插值函数表达式如式（2）所示，$a$ 是固定系数。

$$b(t)=\begin{cases}(a+2)|t|^3-(a+3)t^3+1 & |t|\le 1\\ a|t|^3-5at^2+8a|t|-4a & 1<|t|\le 2\\ 0 & |t|>2\end{cases} \quad (2)$$

假设对应第 $i$ 个卷积特征层学习到的一组连续滤波器为 $f=(f^1,\ldots,f^{D_i})$，对插值得到的连续特征图进行卷积运算得到连续的卷积响应，如式（3）所示。

$$S\{x\}=\sum_{d=1}^{D_i}f^d*J\{x^d\} \quad (3)$$

每一个特征通道首先用式（1）进行插值，然后再与相对应的滤波器进行卷积，最终这一层的连续卷积响应 $S\{x\}$ 由所有通道滤波器卷积响应的和组成。

### 2.2 基于响应图的自适应特征融合

特征提取是目标跟踪的基础和前提，而高效鲁棒的特征是跟踪的关键。良好的特征可以最大程度的区分目标和背景，从而很好的提升算法的跟踪性能。

从 VGG-M[22]卷积神经网络的输入层（为方便陈述，也称为第零层）和 5 层卷积层提取的特征图尺寸大小分别为 224×224 像素，109×109 像素，26×26 像素，13×13 像素，13×13 像素，13×13 像素，第五卷积层的特征图大小大约是第一卷积层的 0.12 倍，是第二卷积层的 0.5 倍。不同的卷积

层得到的特征图大小差别较大，所包含的特征信息也具有不同的特点，对于跟踪的作用也可能不相同，本文基于卷积响应图来对不同层的特征进行自适应加权，从而融合不同层特征用于目标跟踪，旨在提高跟踪效果的稳定性。随着跟踪的进行，当某一层特征跟踪效果变差时，可以自适应地降低该层特征的权重并提高其他层特征的权重，使优势特征占据主导地位，从而实现跟踪器稳定跟踪目标。

本文使用损失差作为衡量跟踪效果的度量，第 $z$ 帧的第 $i$ 层损失差计算如式（4）所示。

$$l_z^i = sum(S_z^i - y_z^i)^2 \qquad (4)$$

式中，$sum()$ 表示矩阵对应项的和，$S_z^i$ 表示当前帧利用公式(3)得到的响应值，$y_z^i = e^{-\frac{1}{2\sigma^2}(t-u_z)}$，$u_z$ 表示当前帧的预测目标位置。根据损失差得到的跟踪效果进行第 $i$ 层特征图的权重自适应计算，如式（5）所示。

$$w_{z+1}^i = \frac{(\sum_i l_z^i) - l_z^i}{(n-1)\sum_i l_z^i} \qquad (5)$$

式中，$n$ 表示用于跟踪的卷积特征层的总层数，$w_{z+1}^i$ 为下一帧第 $i$ 层特征的自适应权重。当某一特征层在当前帧的跟踪损失增大时，它在下一帧跟踪时的权重就会自适应地减少。

得到不同特征层的自适应权重后，就可以融合各特征层相关滤波的结果，如式（6）所示。

$$S_f\{x\} = \sum_i W_i \sum_{d_i=1}^{D_i} f_i^{d_i} * J_i\{x_i^{d_i}\} \qquad (6)$$

式（5）中 $W_i$ 表示第 $i$ 卷积层的特征权重，自适应加权求和多个响应图得到最终的响应图。通过采用文献[32]中的多尺度搜索策略，搜索不同尺度最终响应图的最大值位置即可以确定跟踪目标位置和尺度。

### 2.3 连续相关滤波器的学习

为了学习得到对应于各卷积特征层的连续滤波器，使用了如式（7）所示的损失函数用于优化训练。利用空间惩罚函数 $\beta$ 调节相关滤波器 $f$，对相关滤波器参数添加权重约束，正则化惩罚函数 $\beta_i(p,q) = \tau + \zeta\{(p/P_i)^2 + (q/Q_i)^2\}$，其中 $\tau$ 和 $\zeta$ 是固定参数，$P_i \times Q_i$ 表示第 $i$ 层特征图的大小，使得越靠近边缘位置的空间权重越大，越靠近目标中心位置的空间权重越小。给定 $m$ 个训练样本 $\{(x_z, y_z)\}_1^m$，当第一帧时 $m=1$，当第二帧时 $m=2$，以此类推，每跟踪一帧得到一个新的训练样本，通过最小化如式（7）所示的损失函数训练得到滤波器。

$$\arg\min_f E(f) = \arg\min_f [\sum_{z=1}^{m} \alpha_z \left\| \sum_i W_i \sum_{d_i=1}^{D_i} f_i^{d_i} * J_i\{(x_z)_i^{d_i}\} - y_z \right\|^2 + \sum_i \sum_{d_i=1}^{D_i} \left\| \beta_i f_i^{d_i} \right\|^2] \qquad (7)$$

式中，$\alpha_{z-1} = (1-\lambda)\alpha_z$，且 $\sum_z \alpha_z = 1$，$\alpha_z \geq 0$，$\lambda$ 是固定参数，$\alpha_z$ 决定样本 $x_z$ 对滤波器参数 $f$ 的影响；$\|\bullet\|$ 表示 2 范数；利用共轭梯度法迭代求解上式，第一帧迭代 100 次求解 $f$，在后续视频帧序列中每帧迭代 5 次求解 $f$。

### 2.4 算法流程

结合连续卷积算子的自适应加权目标跟踪算法在每跟踪完成一帧图像后都要更新滤波器参数和权重，更新滤波器是为了能适应目标状态的变化。同时自适应更新计算下一帧多层特征的权重，使优势特征占主导地位。本文算法的具体步骤描述如下：

输入：视频序列和第一帧图像的目标位置和尺寸大小。

输出：视频序列后续帧中的目标位置和尺寸大小。

Begin

IF 第一帧

手动划定需跟踪的目标，提取目标区域的多层卷积特征，初始化各层特征的权重，通过式（7）优化训练得到初始滤波器 ；

Else

Step1:提取预测目标区域的多层卷积特征；

Step2:利用公式（5）计算得到下一帧自适应权重；

Step3:利用公式（6）计算得到的响应图计算当前帧的目标位置和尺度；

Step4:通过公式（7）更新滤波器 ；

Step5:如果不是最后一帧，返回 Step1；

End

## 3 实验及结果分析

为了验证本文算法的性能，使用了 OTB-2013 评估基准数据集[33]的 50 组完全标注的视频序列进行测试，并与 HCFT[13]、DeepSRDCF[34]和 CCOT[15]等近年来比较流行的基于深度学习的跟踪算法进行对比。HCFT 和 CCOT 都是多层卷积特征融合算法，而 DeepSRDCF 是使用单层卷积特征的跟踪算法。

### 3.1 实验环境及参数设置

实验硬件环境是 Intel(R) Core(TM) CPU i5-7300HQ @ 2.50GHz,内存 8GB，显卡 NVIDIA GeForce GTX1050ti，操作系统为 64 位 WINDOWS 10，仿真软件为 MATLAB R2017a。使用 MatConvNet 工具箱的版本是 matconvnet-1.0-beta23。算法参数设置如下：固定系数 $a$ 设置为-0.75，最大保存样本 $m$ 设置为 400,学习率 $\lambda$ 设置为 0.0075,最小正则化惩罚权重 $\beta=\tau=0.0001$，正则化影响因子 $\zeta=0.01$。

### 3.2 跟踪效果比较

本小节实验在 OTB-2013 评估基准数据集[33]上分析比较了本文算法与 HCFT[13]、DeepSRDCF[34]和 CCOT[15]的平均跟踪成功率，即跟踪成功的帧数除以总帧数。当跟踪结果区域与目标真实位置区域的交集除以两者之间的并集，也就是跟踪重合率大于 0.5 时，判定当前帧跟踪成功，否则判定跟踪失败。本实验中，本文算法选用了自适应加权融合第零层（输入层）、第一卷积层和第五卷积层特征进行跟踪，它们的权重全部初始化为 1/3。

本文提出的算法与 HCFT[13]、DeepSRDCF[34]和 CCOT[15]在 OTB-2013 评估基准数据集[33]的 50 个视频序列上的平均跟踪成功率如表 1 所示。从表 1 的实验结果可以看出，本文提出的算法有最好的跟踪成功率，且高出次好的 CCOT 算法 1.7%。这表明本文算法采用的自适应权重融合方法能更好的表达特征，使得跟踪器能够更加准确的跟踪目标。

**表 1 各算法跟踪成功率**

**Table 1 The comparisions of tracking accuracy**

/%

| | HCFT | DeepSRDCF | CCOT | 本文算法 |
|---|---|---|---|---|
| OTB-2013 | 74.0 | 79.4 | *83.7* | **85.4** |

注：每一列的粗字体表示最大值，下滑线斜体字表示次大值

为了进一步比较分析跟踪算法在具有不同复杂情况的视频上的跟踪性能，表 2 分别列出了本文提出的算法与 HCFT 算法[13]、DeepSRDCF 算法[34]和 CCOT 算法[15]在 OTB-2013 评估基准数据集[33]中 11 种不同复杂状况的视频序列上的跟踪成功率。表 2 中用字母缩写分别表示不同的复杂状况，IV 表示光照变化(Illumination Variation, IV)，SV 表示尺度变化(Scale Variation, SV)，OCC 表示遮挡(Occlusion, OCC)，DEF 表示目标形变（Deformation，DEF)，MB 表示运动模糊(Motion Blur, MB)，FM 表示快速运动(Fast Motion, FM)，IPR 表示平面内旋转(In-Plane Rotation, IPR)，OPR 表示平面外旋转(Out-Plane Rotation, OPR)，OV 表示超出视野(Out-of-View, OV)，BC 表示背景杂乱(Background Clutters, BC)，LR 表示低分辨

率(Low Resolution, LR)。在表 2 中，每种状况缩写下方的小括号内的数字表示此类复杂状况包括的视频序列个数。

表 2 不同状况下算法的跟踪成功率比较

**Table 2 The comparisions of Tracking accuracy in 11 different situations**

/%

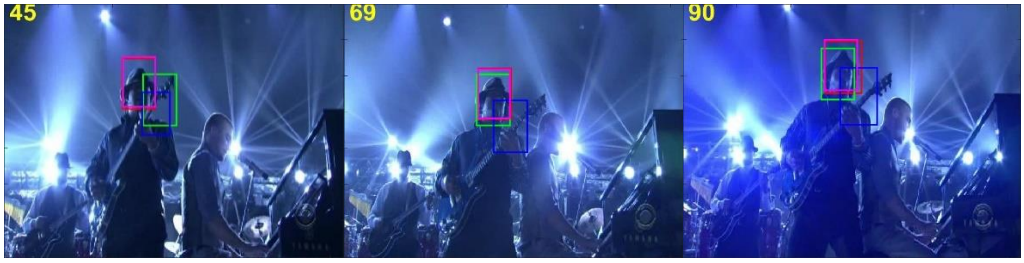
| | IV (26) | SV (30) | OCC (29) | DEF (18) | MB (12) | FM (18) | IPR (34) | OPR (39) | OV (6) | BC (21) | LR (4) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 本文算法 | **81.3** | **81.7** | *86.7* | 82.9 | *81.2* | *81.3* | **79.7** | **83.9** | *87.3* | **79.9** | *71.8* |
| CCOT | *77.3* | *78.0* | **89.8** | **88.2** | **85.8** | **83.9** | 73.6 | *81.4* | **93.9** | 72.1 | **73.8** |
| DeepSRDCF | 71.1 | 76.8 | 73.8 | *84.3* | 79.5 | 76.6 | *75.6* | 77.5 | 70.0 | 69.7 | 44.2 |
| HCFT | 66.5 | 59.4 | 79.3 | 83.0 | 74.9 | 71.3 | 70.0 | 74.1 | 76.5 | *78.8* | 65.5 |

注：每一列的粗字体表示该列的最大值，下滑线斜体字表示该列的次大值

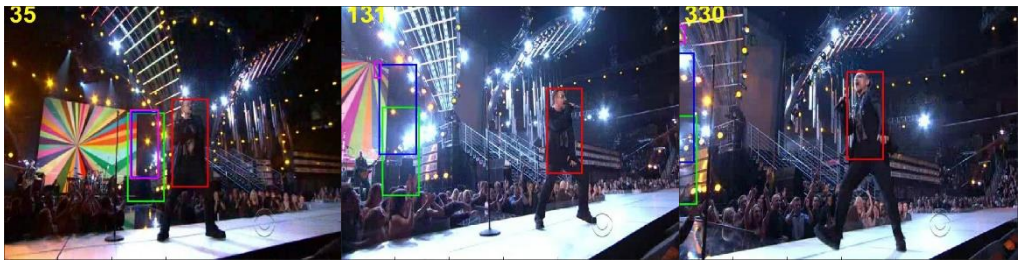
从表 2 的实验结果可以看出，在 11 种不同复杂状况下，除了目标形变外，本文算法的跟踪成功率均为最大值或者次大值。由此表明，本文算法在各种复杂环境条件下都具有较好的跟踪准确性。

图 3 给出了本文算法与 HCFT[13]、DeepSRDCF[34] 和 CCOT[15]跟踪算法的部分跟踪结果图，图中不同的跟踪算法用不同颜色的矩形框表示，红色矩形框表示本文算法，绿色矩形框表示 HCFT 算法，蓝色矩形框表示 CCOT 算法，紫色矩形框表示 DeepSRDCF 算法，在上角的数字为当前帧数。


(a)football1




（b）shaking




(c)singer2

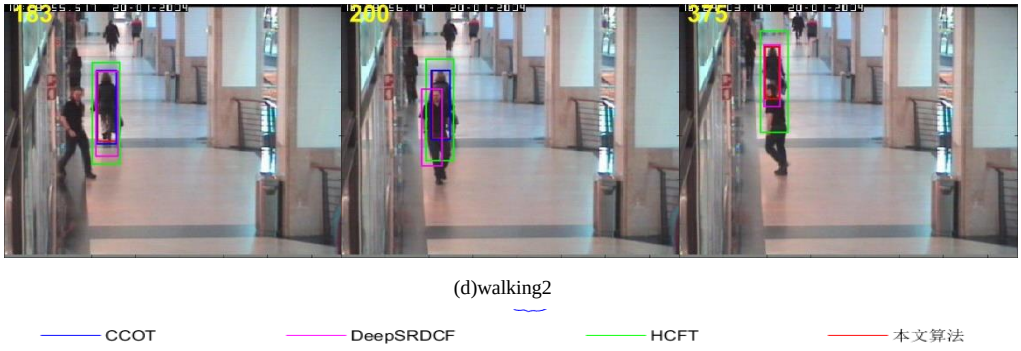



图 3 4 种算法的跟踪效效果对比

Fig.3 Comparison of tracking result of four algorithms: (a) football1; (b)shaking; (c)singer2; (d)walking2.

从图 3（a）可以看出，在包含平面内旋转和运动模糊的 football1 序列中，本文算法在有与目标非常相似的干扰物（第 10 帧），存在运动模糊（第 40 帧）和平面内旋转（第 60 帧）时，都能较为准确的跟踪目标，而 CCOT 算法和 DeepSRDCF 算法在第 40 帧和第 60 帧目标发生运动模糊和平面内旋转的情况下，都发生了跟踪漂移。图 3（b）shaking 序列是一个包含大量杂乱背景的序列，本文算法都能准确跟踪目标，而 CCOT 算法在第 45 帧、第 60 帧和第 90 帧都发生目标跟踪错误。这表明自适应融合特征能够利用优势特征准确表达跟踪目标。

图 3（c）singer2 序列是一个包含背景杂波和大量光照变化的序列，本文提出的算法在整个序列中跟踪定位较为准确；而其他 3 个算法在第 35 帧直到跟踪结束，都发生了跟踪目标丢失现象，这表明背景杂波对他们影响较大。图 3（d）walking2 序列是一个包含尺度变换、遮挡和低分辨率的的序列，本文提出的算法都能准确地定位目标位置。这表明连续卷积算子将离散的位置估计转换成连续位置估计能够更加准确定位目标。DeepSRDCF 算法在发生目标遮挡（第 200 帧）时，发生了跟踪错误，而 HCFT 算法在整个跟踪序列中都不能准确估计目标的尺度。

为了清楚的说明每一帧的跟踪稳定性，在具有代表性的 singer2 序列（光照变化，目标形变，旋转，背景杂乱）上 ，将 4 种算法的中心位置误差（跟踪框中心位置与目标中心位置间的欧式距离的平均值）绘制如图 4 所示。在 singer2 序列上，如图 4 所示，本文算法的中心位置误差远小于其他三个算法的中心位置误差，而中心位置误差越小，表明算法的跟踪稳定性越好，故本文算法的稳定性较好。

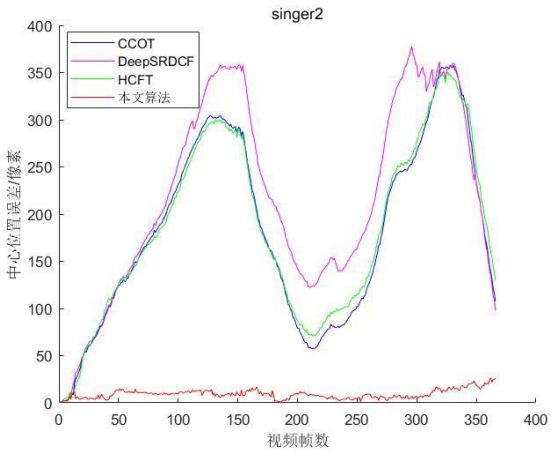



图 4 4 种算法在 singer2 序列的跟踪稳定性对比

Fig.4 Comparison of tracking stability of four algorithms on “singer2” sequence

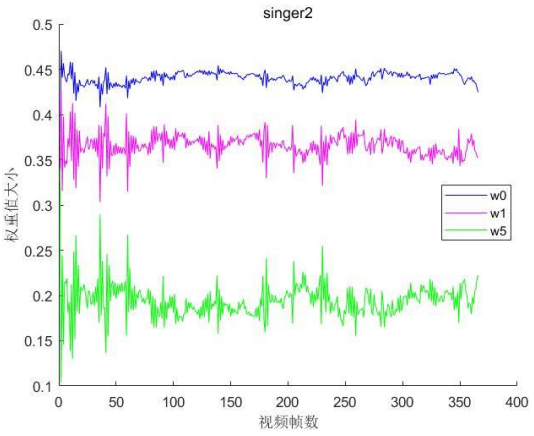


图 5 singer2 序列跟踪过程中的特征权值变化

Fig.5 Variations of feature weightes during tracking on "singer2" sequence

在具有代表性的 singer2 序列上三个卷积层特征的权值在跟踪过程中的变化情况如图 5 所示。从图 5 可以看出，浅层特征权重明显大于高层特征权重，可能是由于 singer2 序列中没有相似背景物体干扰，所以空间位置信息就显的比较重要，故权重较大。

### 3.3 本文算法不同融合方案分析

为了进一步分析自适应加权连续融合不同卷积层特征对本文算法跟踪性能的影响，本实验通过组合不同卷积层特征进行跟踪对比分析，实验结果如表 3 所示。从表 3 可以看出，当只使用二层或者使用四层特征进行跟踪时，算法的成功率均比使用三层低，这可能是太少层特征会造成信息不足，而过多层特征又会引起信息重叠。同时也从表 3 实验结果观察到使用了连续卷积层特征的性能，如融合第 0、1、2、5 层和第 0、1、4、5 层，比融合间隔卷积层特征，如第 0、1、3、5 层，性能要更差，这可能是从间隔卷积层提取的特征互补性更强。

表 3 不同卷积层组合下的跟踪成功率比较

**Table 3 Comparisons of tracking accuracy of different combinations of convolution layers**

| | 初始权重 | 成功率/% |
|---|---|---|
| layer 0 5 | [0.5;0.5] | 70.4 |
| layer 1 5 | [0.5;0.5] | 83.6 |
| layer 0 1 5 | [1/3; 1/3; 1/3] | 85.4 |
| layer 0 1 2 5 | [0.25;0.25;0.25;0.25] | 84.2 |
| layer 0 1 3 5 | [0.25;0.25;0.25;0.25] | 84.5 |
| layer 0 1 4 5 | [0.25;0.25;0.25;0.25] | 83.7 |

## 4 结论

本文提出了一种结合连续卷积算子的自适应加权目标跟踪算法，该算法利用连续卷积算子创建时域连续的相关滤波器进行更准确的定位，同时利用相关滤波算法自适应融合多层卷积特征，达到削弱背景干扰和增强特征表达的效果。与多种主流的相关滤波跟踪算法的实验对比结果表明, 本文提出的算法对常见跟踪难度，如光照变化、尺度变化、平面内旋转、平面外旋转和背景杂波具有较好的适应性。下一步的工作将从优化网络结构，提取更加有效的特征（如深度运动特征）等方面进行分析和研究。

### 参考文献（References）


[1] Huang X Y, Cheng X J, Geng Q C, et al. The apolloScape dataset for autonomous driving[C]//Proceedings of the Computer Vision and Pattern Recognition.Salt Lake City,USA,2018.

[2] Mei X, Ling H B. Robust visual tracking using l(1) minimization[C]//Proceedings of 2009 IEEE International Conference on Computer Vision.Kyoto,Japan:IEEE,2009:1436-1443.[DOI:10.1109/ICCV.2009.5459292]

[3] Liu B Y, Liu Y, Huang J Z, et al.Robust and fast collaborative tracking with two stage sparse optimization[C]//Proceedings of 2010 European Conferenceon Computer Vi-sion.Crete, Greece: Berlin:Springer,2010:624-637.[DOI:10.1007/978-3-642-15561-1_45]

[4] Bao C L, Wu Y, Ling H B, et al. Real time robust L1 tracker using accelerated proximal gradient approach [C] // Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence,RI,USA:IEEE,2012:1830-1837.[DOI:10.1109/CVPR.2012.6247881]

[5] 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]

[6] Zhang W, Kang B S. Recent advances in correlation filter-based object tracking: a review[J]. Journal of Image and Graphics, 2017, 22( 8) :1017-1033] [张微, 康宝生. 相关滤波目标跟踪进展综述[J]. 中国图象图形学报, 2017, 22(8): 1017-1033.] [DOI:10.11834/jip.170092]

[7] Li H X, Li Y, Porikli F. DeepTrack: Learning discriminative feature representations by convolutional neural networks for visual trackin-

g[C]//Proceedings of 2014 British Machine Vision Conference.Nottingham,United kingdom: BMVA Press,2014:1-12.[DOI:10.5244/C.28.56]

[8] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition.Washington,USA:IEEE,2016:4293-4302.[DOI:10.1109/CVPR.2016.465]

[9] Yao R, Shi Q F, Shen C, et al. Part-Based visual tracking with online latent structural learning[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Rec-ognition.Portland,OR,USA:IEEE,2013:2363-2370.[DOI:10.1109/CVPR.2013.306]

[10] Ning J, Yang J, Jiang S, et al. Object tracking via dual linear structured SVM and explicit feature map[C] //Proceedings of IEEE the Computer Vision and Pattern Recognition. Lasvegas,NV,United States:IEEE,2016:4266-4274.[DOI:10.1109/CVPR.2016.462]

[11] Bolmw D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.San Francisco,USA:IEEE,2010.2544-2550 [DOI: 10.1109/CVPR.2010.5539960]

[12] Danelljan M, Häger G, Khan F S, et al. Accurate scale estimation for robust visual tracking[C]//Proceedings of 2014 British Machine Vision Conference.Nottingham:BMVA Pr-ess,2014:65.1-65.11.[DOI:10.5244/C.28.65]

[13] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking[C] //Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago,Chile: IEEE,2015:3074-3082.[DOI:10.1109/ICCV.2015.352]

[14] Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks[C]// Proceedings of 2016 IEEE International Conference on Computer Vision.Santiago,Chile:IEEE,2016:3119-3127.[DOI:10.1109/ICCV.2015.357]

[15] Danelljan M, Robinson A, Khan F S, et al. Beyond Correlation Filters:learning continuous convolution operators for visual tracking[C] // Proceedings of 2016 IEEE Conferenceon European Conference on Computer Vision. Berlin, Germany:IEEE,2016:472-488.[DOI:10.1007/978-3-319-46454-12_9]

[16] He Z, Fan Y, Zhuang J, et al. Correlation filters with weighted convolution responses[C]// Proceedings of 2017 IEEE International Conference on Computer Vision.Venice,Italy: IEEE, 2017:1992-2000.[DOI:10.1109/ICCVW.2017.233]

[17] Qi Y, Zhang S, Qin L, et al. Hedged deep tracking[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,Nevada,USA:IEEE,2016:4303-4311.[DOI:10.1109/CVPR.2016.466]

[18] Tao R, Gacces E, Smeulders A W M.Siamese instance search for tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas,Nevada, USA:IEEE,2016:1420-1429.[DOI:10.1109/CVPR.2016.158]

[19] Held D, Thrun S, Savarese S. Learning to track at 100 fps with deep regression networks [C]//Proceedings of 2016 IEEE Conferenceon European Conference on Computer Vision.Ber-lin,Germany:IEEE,2016:749-765.[DOI:10.1007/978-3-319-46448-0_45]

[20] Cui Z, Xiao S, Feng J, et al. Recurrently target-attending tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas,Nevada,USA:IEEE,2016:1449-1458[DOI:10.1109/CVPR.2016.161]

[21] Fan H, Ling H. SANet: Structure-aware network for visual tracking[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA:IEEE,2017:2217-2224.[DOI:10.1109/CVPRW.2017.275]

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition.Columbus,USA:IEEE,2014.

[23] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Proceedings of 2014 IEEE Conferenceon European Conference on Computer Vision.Zurich, Switzerland.Springer,2014:818-833. [DOI:10.1007/978-3-319-10590-1_53]

[24] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature,2015,521(7553): 436-444.[DOI: 10.1038/nature14539]

[25] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J].Communications of the Acm,2012,60(2):1097-1105.[DOI: 10.1145/3065386]

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas,Nevada, USA:IEEE,2016:770-778.[DOI:10.1109/CVPR.2016.90]

[27] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision,2015,115(3):211-252.[DOI:10.1007/s11263-015-0816-y]

[28] Lu H C,Li P X,Wang D. Visual object tracking: a survey[J]. Pattern Recognition and Artificial Intelligence, 2018,31(1): 61-76. [卢湖川, 李佩霞, 王栋. 目标跟踪算法综述[J]. 模式识别与人工智能,2018,31(1): 61-76.] [DOI:10.16451/j.cnki.issn1003-6059.201801006]
[29] Li F, Tian C, Zuo W, et al. Learning spatial-temporal regularized correlation filters for visual tracking[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition.Munich,Germany:IEEE,2018.
[30] Song Y B, Ma C, Wu X, et al. VITAL: visual tracking via adversarial learning[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition.
[31] Lukezic A, Vojir T, Cehovin L, et al. Discriminative correlation filter with channel and spatial reliability[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,Hawaii,USA:IEEE,2017:4847-4856[DOI:10.1109/CVPR.2017.515]
[32] Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration [C]// Proceedings of 2014 IEEE Conferenceon European Conference on Computer Visi-on.Zurich, Switzerland.Springer,2014:8926(254-265).[DOI:10.1007/978-3-319-16181-5_18]
[33] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland,OR,USA:IEEE,2013.2411-2418.[DOI:10.1109/CVPR.2013.312]
[34] Danelljan M, Häger G, Khan F S, et al. Convolutional features for correlation filter based visual tracking[C]//Proceedings of 2015 IEEE International Conference on Computer Vision Workshop.Santiago,Chile:IEEE,2015:621-629[DOI:10.1109/ICCVW.2015.84]

## 作者简介

罗会兰，1974 年生，女，博士，主要从事机器学习、模式识别等方面的研究

E-mail：luohuilan@sina.com

石武，男，硕士研究生，主要研究方向为目标跟踪。

E-mail：470162846@qq.com