基于卷积神经网络与马尔可夫随机场的目标检测

吉珊珊[1] 陈传波[2]

[1](东莞职业技术学院计算机工程系 广东 东莞 523808)

[2](华中科技大学软件学院 湖北 武汉 430074)

摘 要 为了提高视频目标检测的边缘准确性,提出一种基于卷积神经网络和马尔可夫随机场的视频目标检测算法。通过视频的置信帧来调节与优化神经网络,解决深度卷积神经网络的标签不一致问题;采用马尔可夫模型将前景目标与背景分割,对超像素做平滑处理;设计前后光流融合的密集光流法,提高目标的运动一致性。基于多组公开视频数据集进行了仿真实验,结果显示该算法在目标检测性能方面具有明显的优势,提取的目标轮廓具有较高的准确性。

关键词 卷积神经网络 马尔可夫随机场 视频检测 光流提取 均衡化处理

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.023

TARGET DETECTION BASED ON CONVOLUTIONAL NEURAL NETWORKS AND MARKOV RANDOM FIELD

Ji Shanshan[1] Chen Chuanbo[2]

[1](*Department of Computer Engineering, Dongguan Polytechnic, Dongguan 523808, Guangdong, China*)

[2](*School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China*)

Abstract To improve the edge accuracy of video target detection, we propose a video target detection algorithm based on convolutional neural networks and Markov random field. The neural networks were adjusted and optimized through the belief frames of videos to resolve the labels inconsistency issue of neural networks; the foreground target and background were segmented by Markov model, and the super pixels were smoothed effectively; a front and rear optical flows combined dense optical flow method was designed to enhance the moving consistency of targets. Simulation experimental results based on multiple public video datasets show that our algorithm has obvious advantages in target detection performance, and the extracted target contour has high accuracy.

Keywords Convolutional neural networks Markov random field Video detection Optical flow extraction Equalization process

0 引 言

目标检测是智能监控、虚拟现实、人机交互、动作分析等领域的关键技术,其性能直接影响了应用领域的效果[1]。目前主流的目标检测与识别工作主要针对某些指定的应用场景,如手势识别[2]、动作识别[3]、面部表情识别[4]等。语义分割[5]融合了传统的视频分割和目标识别两个任务,其目标是将视频分割成若干特定语义的区域,最终获得像素语义标注的视频序列。语义分割的优点在于能够无差别地检测出前景区域,本文利用语义分割的优点,将语义分割思想应用到目标检测领域中。

深度卷积神经网络(Deep Convolutional Neural Networks,DCNN)通过训练数据自动地学习特征,在图像、视频的目标识别领域取得了较好的效果[6]。文献

收稿日期:2019-07-18。广东省教育厅青年创新人才类项目(2017GkQNCX119);东莞市社会科技发展项目(2017507156388);东莞职业技术学院政校行企合作项目(政 2018019)。**吉珊珊**,讲师,主研领域:计算机智能信息处理与控制,计算机仿真,计算机教育。**陈传波**,教授。

[7]将 DCNN 应用于人脸标签识别的问题中,该研究通过 GPU 提高了 DCNN 的处理速度,并且实现了较高的检测准确率。文献[8]提出一种基于 DCNN 与长短期记忆网络(Long-Short Term Memory,LSTM)的维吾尔语文本突发事件识别方法,该算法实现了较高的召回率与查准率。文献[9]将 DCNN 应用于动作识别领域中,该算法获得了极高的查全率与查准率,但是该算法需要输入先验动作集进行预训练。上述 DCNN 模型大多为全监督或者半监督的问题,在弱监督的 DCNN 训练过程中存在明显的标签不一致问题,而视频目标检测问题大多属于弱监督数据。

为了解决 DCNN 的标签不一致问题,本文提出置信帧的概念,算法采用置信帧对预训练的 DCNN 模型进行优化调节,提高模型的性能。DCNN 输出的特征还不足以准确提取出目标,采用马尔可夫模型将前景目标与背景分割。本文采用马尔可夫随机场(Markov Random Field,MRF)优化 DCNN 获得的标签,进一步提高像素标签映射的精度,最终通过密集光流法对分割检测结果进行提取,提高目标边缘的检测准确率。

1　基于 DCNN 的特征提取

本文方法的核心思想是利用置信帧的高置信度调节 DCNN 模型。设 Φ 表示视频帧的索引集,Ω 表示视频的弱标签集。采用预训练的 DCNN 模型 θ 处理各帧 $f \in \Phi$,使用 SOFTMAX 函数计算像素 i 属于类 $x_i \in O$ 的概率 $P(x_i \mid \theta)$,O 表示目标与背景的集合。使用 ARGMAX 函数处理每个像素 i,计算语义标签的映射 S:$S(i) = \text{argmax}_{xi} P(x_i \mid \theta)$。

采用训练集 Γ 训练 DCNN 模型 θ,选出全局 CE 帧与局部 CE 帧,计算标签映射 G^{g} 与 G^{l} 来建立自适应数据集。算法 1 为 DCNN 模型训练的伪代码,首先对 S 的每个标签映射进行连通区域分析(Connected Component Analysis,CCA),产生一个目标的候选区域集,记为 $\mathscr{R}$。然后评估目标的置信度 $C(R_k)$,R_k 表示目标区域,k 为区域的序号。$C(\cdot)$ 算子的输入为一个标签映射,输出为标签映射中像素被设为目标的平均概率。

算法 1　DCNN 模型训练

输入:θ, Ω。　　/* θ 为 DCNN 模型,Ω 为弱标记集合 */
输出:θ'。　　/* θ' 为调节后的 DCNN 模型 */
1. $d = 0$;　　/* 局部最优置信度 */
2. FOREACH $f \in \Phi$ DO
3.　　初始化 $G_f^{\text{g}}, G_f^{\text{l}}$;
4.　　计算 $P(x \mid \theta), S = \text{argmax}_x P(x \mid \theta)$;
5.　　计算 S 中连接元素的集合 R;
6.　　FOREACH $R_k \in \mathscr{R}$ DO
7.　　IF $C(R_k) > t_o$ THEN
8.　　$G_f^{\text{g}}(i) = S(i),\ \forall i \in R_k$;
9.　　IF $(S(i) \in \Omega \,\&\&\, i \notin R_k)$
10.　　设置 $G_f^{\text{g}}(i) = G_f^{\text{l}}(i) = 0, \forall i$ 约束条件为 $P(x_i = bg \mid \theta) > t_{\text{b}}$;
11.　　IF $C(G_f^{\text{g}}) > 0$, THEN $\Gamma \leftarrow \Gamma \cup \{G_f^{\text{g}}\}$;
12.　　IF $C(G_f^{\text{l}}) > d$, THEN
13.　　$t = f, d = C(G_f^{\text{l}})$;
14.　　IF $f \bmod \tau_{\text{b}} = 0$ THEN
15.　　IF $G_t^{\text{g}} \notin T_{\text{l}}$ THEN
16.　　$\Gamma \leftarrow \Gamma \cup \{G_t^{\text{l}}\}$;
17.　　$d = 0$;　　/* d 初始化为 0 */
18. 使用 Γ 将 DCNN 模型 θ 调节为 θ'。

如果某个区域的置信度达到阈值 t_o,则设置该区域的标签,生成标签映射 G_f^{g}。如果像素的概率 $P(x_i = bg \mid \theta)$ 达到阈值 t_{b},则将该像素设为背景标签。在构建 G_f^{g} 的过程中,剩余的像素设为“NULL”标签,模型的损失函数中不考虑 NULL 标签的像素,同时也忽略视频的弱标签。最终将至少有一个置信区域($C(G_f^{\text{g}}) > 0$)的全局置信帧加入数据集 Γ 中。

如果选择的帧在时域内分布不均匀,那么模型可能会被短时间的帧所支配,因此选择每个时段 τ_{b} 内目标置信度最高的局部置信帧。局部 CE 帧与其标签映射 G^{l} 的选择方法为:如果帧 f 的标签 $S(i)$ 属于 Ω,则保留所有像素的标签,生成标签映射。计算 $C(G_f^{\text{l}})$ 来评估帧的置信度,将时间 τ_{b} 内置信度最高的帧作为局部置信帧。如果某个局部置信帧未被选为全局置信帧,则将该帧加入数据集 Γ 中。给定自适应训练集 Γ,基于 Γ 将 DCNN 模型 θ 优化为模型 θ',然后采用 θ' 预测新标签。

图 1 为线上程序的训练过程。对输入视频做预训练,选出长时(long-term)帧集与短时(short-term)帧集,对模型进行优化与调节。分别维护 Γ 的长时帧队列 T_{l} 与短时帧队列 T_{s},T_{l} 保存 τ_{l} 的全局 CE 帧,T_{s} 保存 τ_{s} 的局部 CE 帧,T_{l} 队列的优先级高于 T_{s} 队列。T_{l}、T_{s} 作为自适应数据集,更新模型 θ 的参数。

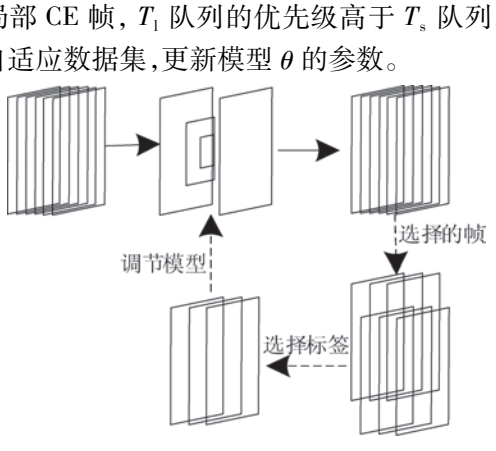

图 1　线上程序的训练过程

2 基于马尔可夫模型的目标分割

DCNN 的输出还不足以准确提取出目标，采用马尔可夫模型[10]将前景目标与背景分割。

2.1 目标分割的上下文模型

视频的前景分割方案大多采用固定的掩膜来提取局部特征，估计出超像素的标签，然后通过 MRF 对标签作平滑处理，所以超像素的分割效果高度依赖超像素的形状与大小。为了解决该问题，考虑多个超像素分割可提高超像素的标签准确率，为此设计了“多假设”MRF 模型。

为局部上下文引入邻接超像素的邻居，同时引入相交超像素的邻居，这两种超像素邻居有助于融合多个超像素的不同描述符。MRF 模型同时描述了超像素内部与超像素外部的上下文信息，内部邻居包含了给定超像素的相邻超像素，外部邻居包含了给定超像素的相交超像素。采用 MRF 模型编码内部邻居与外部邻居的上下文约束条件，以提高超像素标签的一致性。

图 2 为 3 个超像素的 MRF 模型示意图。设 2 个超像素（前景区域与背景区域）的集合为：$SP_m = \{s_i^m\}(m = 1,2)$，两个超像素相交产生第 3 个超像素，设为 $SP_3 = \{s_k^3\}$，定义为：

$$s_k^3 = s_i^1 \cap s_j^2 \neq \varnothing \quad \forall s_i^1 \in SP_1 \quad \forall s_j^2 \in SP_2 \tag{1}$$

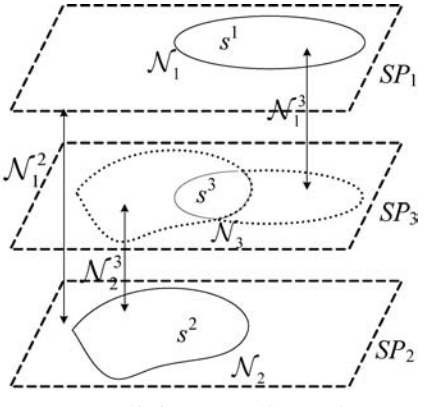

图 2 超像素的 MRF 模型示意图

设 N_m 为内部相邻超像素的上下文，外部超像素的上下文记为 N_n^m，定义为：

$$(s_i^n, s_j^m) \in N_n^m \tag{2}$$

式中：$m,n = \{1,2,3\}$。目标检测是为每个超像素分配一个类标签 c_i^m，检测的结果设为 $c^m = \{c_i^m\}$。将标签问题建模为 MRF 超像素标签集 $c = \{c^1, c^2, c^3\}$ 的能量最小化问题，定义为：

$$J(c) = \sum_{m=1}^{3}\Big(\sum_{s_i^m \in SP_m} D(s_i^m, c_i^m) + \lambda_m \sum_{(s_i^m, s_j^m) \in N_m} E(c_i^m, c_j^m)\Big) + \sum_{(n,m) \in IC} \lambda_n^m \sum_{(s_i^n, s_j^m) \in N_n^m} E(c_i^n, c_j^m) \tag{3}$$

式中：D 与 E 分别为标签分配的数据成本和平滑处理成本；IC 为外部邻居的集合 $IC = \{(1,2),(1,3),(2,3)\}$；$\lambda_m$ 和 λ_n^m 分别为内部上下文和外部上下文的平滑常量。

2.2 基于多假设 MRF 模型的目标分割

MRF 模型的数据成本 $D(s_i, c)$ 定义为超像素 s_i 的类标签为 c 的置信度，平滑成本 $E(c_i, c_j)$ 定义为两个相邻超像素标签分别为 c_i 和 c_j 的概率。DCNN 的输出为视频帧的像素类标签映射，将类标签相同的相邻像素划分为同一个超像素。然后将超像素的强度值设为该超像素中所有像素的平均强度值，基于平均强度定义 MRF 模型的数据项。图 3 为超像素均值化处理的结果图。

(a) 原图像 (b) 超像素均值化处理

图 3 超像素均值化处理的结果

SP_1 与 SP_2 的平滑处理成本依赖标签的共生概率，定义为：

$$E(c_i, c_j) = -\log[(P(c_i \mid c_j)) + P(c_j \mid c_i)/2]\delta \tag{4}$$

式中：$P(c_i \mid c_j)$ 是某个超像素标签为 c_i 同时其邻居标签为 c_j 的条件概率，如果 $c_i = c_j$，δ 则设为 0，否则为 1。

SP_3 由超像素 SP_1 和 SP_2 产生，所以其数据成本和平滑成本定义为关于 SP_1 和 SP_2 的成本函数。SP_3 中包含两种标签 $c^{(1)}$ 和 $c^{(2)}$，分别对应于 SP_1 和 SP_2 的语义分类，最终 $s_k^3 \in SP_3$ 的数据成本定义为：

$$D(s_k^3, c^{(m)}) = f_m(D(s_i^1, c), D(s_j^2, c)) \tag{5}$$

式中：$m = \{1,2\}$；$(s_i^1, s_k^3) \in N_1^3$；$(s_i^2, s_k^3) \in N_2^3$。外部邻居 N_n^m 的平滑处理成本依赖 N_1 和 N_2，定义为：

$$E(c_i^{(1)}, c_j^{(2)}) = g(E(c_i^{(1)}, c_j^{(1)}), E(c_i^{(2)}, c_j^{(2)})) \tag{6}$$

根据式(3)，平滑常量 λ_m 与 λ_n^m 控制 MRF 模型中不同邻居的上下文依赖程度。采用留一法交叉验证[11]处理训练集，将平均像素准确率最大的参数作为平滑常量。采用文献[12]的 α-扩展方法最小化 MRF 能量函数，其输出结果为三个标签集 $c = \{c^1, c^2, c^3\}$，超像素 SP_3 的标签集 c^3 作为图像的最终标签。

2.3 MRF 的数据成本定义

MRF 模型的平滑常量集为 $\{l \times \lambda \mid l \in \{0,1,2\}$；

$5 \leqslant \lambda \leqslant 25, \lambda \in \mathbf{Z}\}$。函数 g 设为 $g(x, y) = 0.5x + 0.5y$,SP_3 的数据成本定义为:

$$\begin{cases} D(s_k^3, c^{(1)}) = \dfrac{w_k}{w_i^1} D(s_i^1, c) \\ D(s_k^3, c^{(2)}) = \dfrac{w_k}{w_i^2} D(s_j^2, c) \end{cases} \tag{7}$$

式中:$w_k = 0.5 \times | s_k^3 | (1/| s_i^1 | + 1/| s_j^2 |)$。该模型为两个假设 SP_1 和 SP_2 的类标签 $c^{(1)}$ 和 $c^{(2)}$ 分配不同的数据成本,基于上下文信息从两个假设中选择其一。超像素权重 w_k 设为与 s_k^3 成正比关系,如果两个超像素的交集较小,说明 s_k^3 的标签可靠度低,为 s_k^3 的数据成本分配低权重。

3　计算密集光流与后处理

计算三个连续帧的光流,首先使用高斯滤波器过滤每帧的噪声,计算当前帧与前一帧之间的光流,记为 OF_1,当前帧与下一帧之间的光流,记为 OF_2,两个光流融合为稠密光流,将 OF_1 与 OF_2 线性组合为每帧的总光流。图 4 为计算密集光流的流程图。

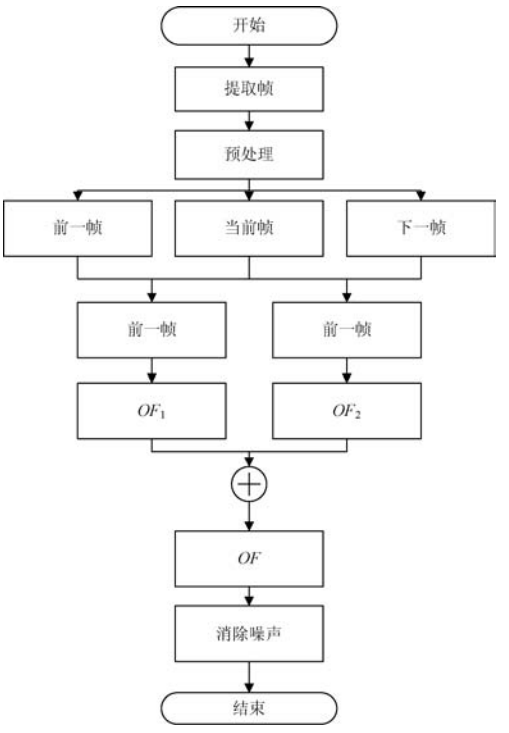

图 4　计算密集光流的流程图

3.1　计算光流

假设亮度恒定,可得:

$$F_{t=i}(x,y) = F_{t+\Delta t}(x + \Delta x, y + \Delta y) \tag{8}$$

式中:(x,y) 为像素的位置;$(x + \Delta x, y + \Delta y)$ 为 Δt 时差的帧坐标;$F_{t=i}$ 与 $F_{t+\Delta t}$ 为时差为 Δt 的两个帧。将式(8)作泰勒级数展开,忽略其高阶项,可得:

$$F_i^x u_i + F_i^y v_i + F_i^t = 0 \tag{9}$$

式中:$u_i = \Delta x/\Delta t$ 和 $v_i = \Delta y/\Delta t$ 分别为水平和垂直方向的速度分量;F_i^x、F_i^y 和 F_i^t 分别为坐标 (x,y)、时间 t 的梯度;u_i 和 v_i 分别为 $F_{t=i}$ 和 $F_{t+\Delta t}$ 之间光流的水平与垂直分量。

式(9)为光流的约束条件,为了获得光流问题的唯一解,需要对 u_i 和 v_i 增加其他光滑约束条件,结合灰度最小化与光滑约束条件估计光流域:

$$E_i = \iint [(F_i^x u_i + F_i^y v_i + F_i^t)^2 + \alpha^2((u_i^x)^2 + (u_i^y)^2 + (v_i^x)^2 + (v_i^y)^2)] \mathrm{d}x\mathrm{d}y \tag{10}$$

式中:参数 α 负责调节光滑度。将 E_i 最小化,可得:

$$u_i = u_i^m - F_i^x \frac{N_i}{D_i} \tag{11}$$

$$v_i = v_i^m - F_i^y \frac{N_i}{D_i} \tag{12}$$

式中:$N_i = F_i^x u_i^m + F_i^y v_i^m + F_i^t$;$D_i = \alpha^2 + (F_i^x)^2 + (F_i^y)^2$;$u_i^m$ 与 v_i^m 分别为像素 u_i 和 v_i 的速度平均值。

选择 3 个连续帧,分别为:当前帧 $F_i(x,y) = F_i^{\mathrm{smooth}}(x,y)$、前一帧 $F_{i-1}(x,y) = F_{i-1}^{\mathrm{smooth}}(x,y)$、下一帧 $F_{i+1}(x,y) = F_{i+1}^{\mathrm{smooth}}(x,y)$。使用式(11)估计帧 $F_{i-1}(x,y)$ 与 $F_i(x,y)$ 的光流 $(u_i^1(x,y), v_i^1(x,y))$,使用式(12)估计帧 $F_i(x,y)$ 和 $F_{i+1}(x,y)$ 的光流 $(u_i^2(x,y), v_i^2(x,y))$。将 2 个光流分别在水平方向和垂直方向求和,计算融合的光流:

$$\cup_i(x,y) = u_i^1(x,y) + u_i^2(x,y) \tag{13}$$

$$\vee_i(x,y) = v_i^1(x,y) + v_i^2(x,y) \tag{14}$$

3.2　基于自适应阈值的降噪处理

因为计算光流的处理中包含不同的处理,所以上述总光流依然含有噪声。前景与背景的分割受噪声的影响较大,采用自适应阈值机制降低噪声的影响。

Otsu 方法[13]是一种全局优化的自适应阈值降噪算法,该方法最小化类内方程、最大化类间方差。帧的像素强度 $F_i(x,y)$ 范围设为 $0 \sim L-1$,设 n_j 是灰度为 j 的像素数量,n 为帧 F_i 的像素总数量。灰度 j 的概率定义为:

$$p_j = \frac{n_j}{n} \tag{15}$$

如果一个帧分为两个类 D_0 和 D_1,D_0 和 D_1 的像素灰度范围分别为 $[0, th-1]$ 和 $[th, L-1]$,其中 th 表示像素的分类阈值。设 $C_0(th)$ 和 $C_1(th)$ 表示累加概率,μ_0 和 μ_1 分别表示 D_0 类和 D_1 类的平均强度。

$$C_0(th) = \sum_{j=0}^{th-1} P_j C_1(th) = \sum_{j=th}^{L-1} P_j = 1 - C_0(th) \tag{16}$$

$$\mu_0 = \sum_{j=0}^{th-1} \frac{jP_j}{C_0(th)}\mu_1 = \sum_{j=th}^{L-1} \frac{jP_j}{C_1(th)} \tag{17}$$

平均灰度值 μ_{th} 计算如下：

$$\mu_{th} = C_0(th)\mu_0 + C_1(th)\mu_1 \tag{18}$$

D_0 与 D_1 两个类之间的方差 σ_e^2 可定义为：

$$\sigma_e^2 = C_0(th)(\mu_0 - \mu_{th})^2 + C_1(th)(\mu_1 - \mu_{th})^2 \tag{19}$$

通过最大化类间方差估计 0 ~ L - 1 范围的最优阈值：

$$TH_i = \arg\left(\max_{0 \leqslant th < L}(\sigma_e^2)\right) \tag{20}$$

假设最优阈值分别为 $TH_i^{\cup}$ 与 $TH_i^{\vee}$，可基于 $\cup_i$ 与 $\vee_i$ 进行降噪处理：

$$\cup_i^{sm}(x,y) = \begin{cases} 0 & |\cup_i(x,y)| \leqslant TH_i^{\cup} \\ \cup_i(x,y) & \text{其他} \end{cases} \tag{21}$$

$$\vee_i^{sm}(x,y) = \begin{cases} 0 & |\vee_i(x,y)| \leqslant TH_i^{\vee} \\ \vee_i(x,y) & \text{其他} \end{cases} \tag{22}$$

3.3 均衡化处理

第 i 帧的密集光流可表示为：

$$OptiFlow_i(x,y) = \sqrt{\cup_i^{sm}(x,y)^2 + \vee_i^{sm}(x,y)^2} \tag{23}$$

图 5(a)和图 5(b)是两个连续的视频帧，图 5(g)是两个连续帧之间的光流图。图 5(c)和图 5(d)是未进行降噪处理和均衡化处理的目标分割结果，图 5(e)和图 5(f)是完成降噪处理和均衡化处理的目标分割结果。

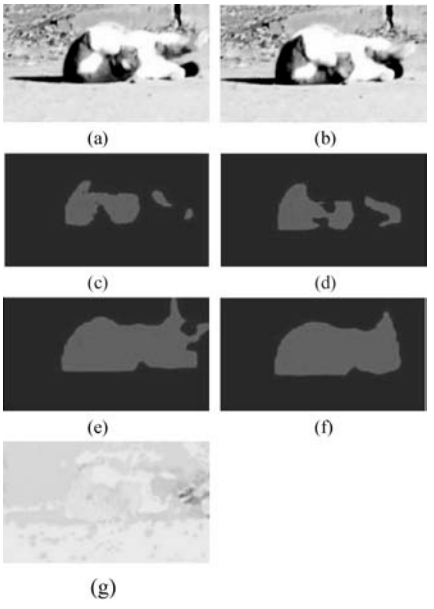

图 5 密集光流法与均衡化处理的结果图

4 实 验

4.1 benchmark 数据集与镜头预处理

采用 traffic 数据集(https://vid.me/videodata)、walking 数据集(https://vid.me/videodata)和 Youtube-Object-Dataset 数据集(https://data.vision.ee.ethz.ch/cvl/youtube-objects/)作为 benchmark 数据集。walking 数据集是一个行人识别的数据集，traffic 数据集是一个交通监控的多目标数据集。Youtube-Object-Dataset 数据集是一个大规模的视频数据集，共有 10 个目标，每个目标包含 9 ~ 24 个视频。Youtube-Object-Dataset 数据集包含正定的前景提取结果，可用作分析检测目标与正定目标的重合程度。

将每个视频分为若干个镜头，每个镜头包含相同的目标与不同的背景。对视频的镜头进行预处理，首先将每个视频帧的长边剪切为 500 像素，然后通过反射处理将视频帧放大至 900 ×900 像素。

4.2 实验环境与性能评价指标

实验环境为 Intel (R) Core (TM) i7 - 4770 CPU @3.40 GHz 处理器，8 GB 内存。基于 Caffe Library[14] 实现 DCNN 模型，基于 MATLAB 编程实现目标检测算法。ODVT[15] 是基于原卷积神经网络的目标检测技术，CSFDV[16] 是一种基于压缩感知的目标检测技术，这两种技术在前景检测的准确率上取得了较大的进步，将本文算法与这两个算法进行横向比较。

算法的参数设为：阈值 $t_o = 0.75$、$t_b = 0.8$，背景值设为略高于前景，留出空间以保留目标周围的像素。局部时间设为 $\tau_b = 30$、$\tau_s = 5$、$\tau_l = 10$。DCNN 的学习率为 0.001，动量为 0.9，权重衰减为 0.000 50。

采用 FPR、TPR、精度和 F-Score 作为目标检测的性能指标，定义如下：

$$FPR = \frac{FP}{FP + TN} \tag{24}$$

$$TPR = \frac{TP}{TP + FN} \tag{25}$$

$$\text{精度} = \frac{TP}{TP + FP} \tag{26}$$

$$F\_Score = \frac{2 \times \text{精度} \times TPR}{\text{精度} + TPR} \tag{27}$$

式中：FP 为假正率；TN 为真负率；TP 为真正率；TPR 为召回率。

Intersection-Over-Union(IOU)定义为系统预测的目标与正定目标的重合程度，计算方法为检测结果与正定值的交集除以两者的并集，该指标能够精细地评估目标检测的准确率。

4.3 实验结果与分析

4.3.1 walking 与 traffic 数据集的实验结果

图 6、图 7 分别为 3 个目标检测算法对于 traffic 和 walking 数据集的实验结果，traffic 和 walking 2 个数据

集均为运动目标的数据集,本文算法对于 3 个数据集均实现了较好的检测准确率和较低的误检率。

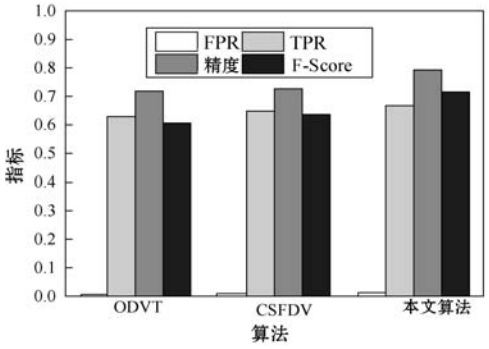

图 6　traffic 数据集的性能结果

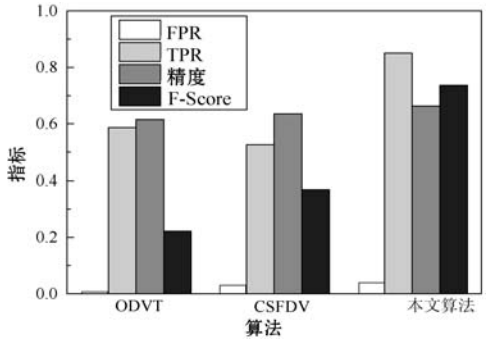

图 7　walking 数据集的性能结果

图 8、图 9 分别为 3 个目标检测算法对于 traffic 和 walking 数据集的前景提取实例。3 个算法虽然均检测出 traffic 数据集中的车辆,但是对车辆的分割结果多有缺失,而本文算法提取的前景目标较为准确,并且保留了较为完好的轮廓。

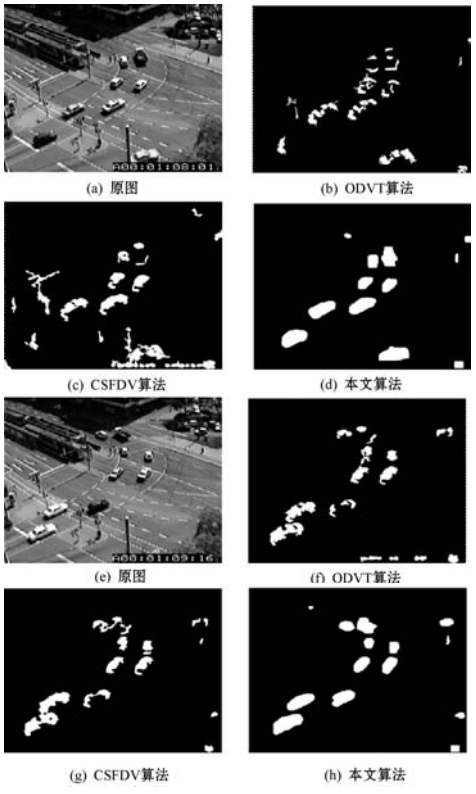

图 8　traffic 数据集的前景提取实例

图 9　walking 数据集的前景提取实例

4.3.2　Youtube-Object-Dataset 的实验结果

为了进一步观察本文算法对于目标提取的细节保留效果,基于 Youtube-Object-Dataset 数据集进行了实验。图 10 为 Youtube-Object-Dataset 数据集的实验结果,本文算法对于 Aero 的检测率低于其他 2 个算法,但其他 9 个目标的效果均明显高于其他 2 个算法,原因是 Aero 目标移动速度较快,本文算法的提取效果较差。

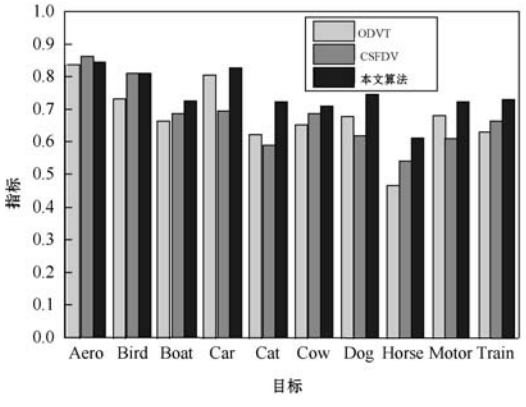

图 10　Youtube-Object-Dataset 数据集的实验结果

图 11 为 Youtube-Object-Dataset 数据集的分割实例图,(a)、(d)、(g)、(j)、(m)为 ODVT 算法的结果,(b)、(e)、(h)、(k)、(n)为 CSFDV 算法的结果,(c)、(f)、(i)、(l)、(o)为本文算法的结果。可看出本文算法对于不同数据集的分割准确率较高,对于目标轮廓的提取更为细致。

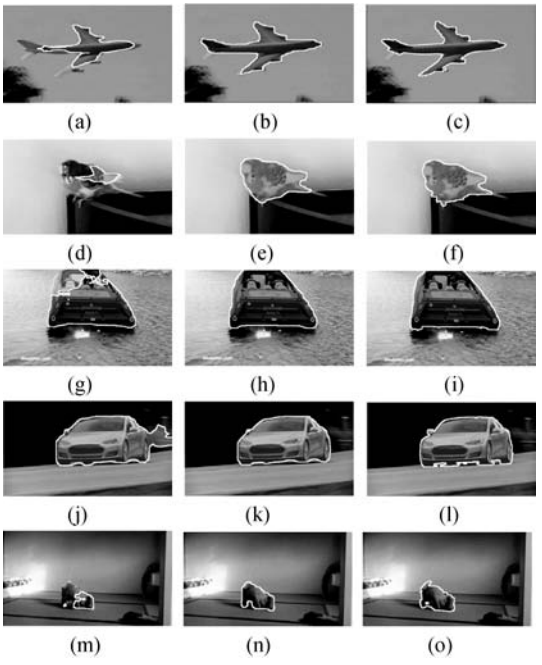

图 11 3 个目标检测算法对 Youtube-Object-Dataset 的分割结果

5 结 语

为了提高视频目标检测的边缘准确性,提出一种基于卷积神经网络和马尔可夫随机场的视频目标检测算法。采用置信帧对预训练的 DCNN 模型进行优化调节,提高模型的性能,采用马尔可夫模型将前景目标与背景分割,采用马尔可夫随机场优化 DCNN 获得的标签,进一步提高像素标签映射的精度。本文算法对视频目标边缘分割的准确率较高,可用于机器人等对精度要求高的领域。本文算法的 DCNN 模型训练的时间复杂度较高,基于 GPU 可实现较快的处理速度,未来将关注于提高算法的时间效率,进一步提高算法的实用性。

参 考 文 献

[1] 尹宏鹏,陈波,柴毅,等. 基于视觉的目标检测与跟踪综述[J]. 自动化学报,2016,42(10):1466 - 1489.

[2] 杜堃,谭台哲. 复杂环境下通用的手势识别方法[J]. 计算机应用,2016,36(7):1965 - 1970.

[3] Wang H,Dan O,Verbeek J,et al. A robust and efficient video representation for action recognition[J]. International Journal of Computer Vision,2016,119(3):219 - 238.

[4] Yin F,Lu X,Li D,et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]//Acm International Conference on Multimodal Interaction,2016.

[5] 李琳辉,钱波,连静,等. 基于卷积神经网络的交通场景语义分割方法研究[J]. 通信学报,2018,39(4):123 - 130.

[6] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229 - 1251.

[7] Mayya V,Pai R M,Pai M M M. Automatic facial expression recognition using DCNN[J]. Procedia Computer Science,2016,93(2):453 - 461.

[8] 黎红,禹龙,田生伟,等. 基于 DCNNs-LSTM 模型的维吾尔语突发事件识别研究[J]. 中文信息学报,2018,32(6):57 - 66.

[9] Patel C I,Garg S,Zaveri T,et al. Human action recognition using fusion of features for unconstrained video sequences[J]. Computers & Electrical Engineering,2018,70(1):284 - 301.

[10] 赵英,韩春昊. 马尔可夫模型在网络流量分类中的应用与研究[J]. 计算机工程,2018,44(5):291 - 295.

[11] 张英堂,马超,李志宁,等. 基于快速留一交叉验证的核极限学习机在线建模[J]. 上海交通大学学报,2014,48(5):641 - 646.

[12] Boykov Y,Kolmogorov V. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(9):1124 - 1137.

[13] Otsu N. A threshold selection method from gray-level histograms[J]. IEEE Transactions on Systems Man & Cybernetics,2007,9(1):62 - 66.

[14] Jia Y,Shelhamer E,Donahue J,et al. Caffe:Convolutional architecture for fast feature embedding[C]//22nd ACM International Conference on Multimedia,2014.

[15] Kang K,Ouyang W,Li H,et al. Object detection from video tubelets with convolutional neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition,2016.

[16] Sukumaran A N,Sankararajan R,Swaminathan M. Compressed sensing based foreground detection vector for object detection in wireless visual sensor networks[J]. AEU-International Journal of Electronics and Communications,2017,72(1):216 - 224.

(上接第 83 页)

[10] Menke J E,Martinez T R. A Bradley-Terry artificial neural network model for individual ratings in group competitions[J]. Neural Computing & Applications,2008,17(2):175 - 186.

[11] 吴霖,陈磊,邓超,等. 基于 TrueSkill 模型的围棋棋手排名方法及评估[J]. 昆明理工大学学报(自然科学版),2013,38(3):47 - 55.

[12] Li M. Efficient mini-batch training for stochastic optimization[C]//ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM,2014.

[13] KGS Go Sever[EB/OL]. [2019-07-03]. http://www.gokgs.com/.

[14] U-go.net[EB/OL]. [2019-07-03]. https://u-go.net/gamerecords/.