文章编号：2095-2163(2023)02-0187-08 中图分类号：TP391.41 文献标志码：A

基于高维 SIFT 改进隐马尔可夫模型的多目标跟踪

刘艺博，奚峥皓，陈健超
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：针对多目标跟踪领域中出现的遮挡、目标身份互换等问题，本文提出了一种基于尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）关键点和隐马尔可夫模型的多目标跟踪算法。首先，在视频序列中逐帧提取每个目标的关键点集，并对其进行条件约束；其次，以得到的关键点集作为目标的状态建立隐马尔可夫模型，根据模型在时序段内传递状态的规律求出模型对应的参数；最后，以当前帧的观测状态和参数求出下一帧的隐性状态，实现对目标位置的预测。为了提升模型的推理速度，建立了表征全部目标的高维观测状态模型。与其他先进的算法在 MOT17、MOT20、KITTI 数据集上进行了仿真实验对比，结果表明本算法在跟踪准确度等指标上表现较优，并对遮挡和身份互换问题有较好的鲁棒性。

关键词：多目标跟踪；尺度不变特征变换；隐马尔可夫模型；高维观测状态

Multi-object tracking based on improved hidden Markov model with high dimensional SIFT

LIU Yibo，XI Zhenghao，CHEN Jianchao
（College of Electrical and Electronic Engineering，Shanghai University of Engineering Science，Shanghai 201620，China）

【Abstract】 Aiming at the problems of occlusion and object identity exchange in the multi-object tracking，this paper proposes a multi-object tracking algorithm based on hidden Markov model with Scale-Invariant Feature Transform（SIFT）features. Firstly，the SIFT features set of each object in each frame are extracted and constrained. Then the constrained SIFT set is taken as the object state，and a hidden Markov model is established with the state. The corresponding parameters of the model are obtained according to the law of the state transmitting in each period. Finally，the hidden state in the next frame is estimated by the observation state and parameters in the current frame so as to realize the prediction of object location. In addition，in order to improve the reasoning speed of the model，a high-dimensional observation state model representing all objects is established. Compared with other algorithms on MOT17、MOT20 and KITTI datasets，the experimental results show that this algorithm performs better in tracking accuracy and other indicators，and has good robustness to occlusion and identity exchange problems.

【Key words】 multi-object tracking；scale-invariant feature transform；hidden Markov model；high-dimensional observation state

0 引 言

自然场景下的多目标跟踪（Multi-object tracking，MOT）是计算机视觉的一个重要问题，其在自动驾驶、军事安全等领域都有广泛的应用。多目标跟踪的目的是在一个视频中根据初始帧一些标定了身份信息的目标，在后续帧中维持这些目标的身份，并形成有效的轨迹[1-2]。然而，在复杂的环境背景下，目标在运动过程中容易被环境障碍物遮挡，并且，当多个目标相互交错时容易引起跟踪目标身份的丢失和互换，导致无法从视频中提取到完整的运动轨迹。

马尔可夫模型在多目标跟踪任务中有显著的优点，为了解决 MOT 中的遮挡问题，Liu 等[3]提出了一种轨迹耦合关联的马尔可夫随机场模型，通过整合密集人群的位置和运动信息，对其中不完整和偏差的位置数据进行校正，提高了跟踪器的精度；Xiang 等[4]将 MOT 表述为马尔可夫决策过程中的决策，以策略学习的方式加强不同帧目标之间的数据关联；Wu 等[5]将马尔可夫决策过程与不同频率的相关滤波器关联，解决了跟踪过程中因为遮挡和尺度变换造成的目标漂移问题；Vojir 等[6]利用隐马

作者简介：刘艺博（1997-），男，硕士研究生，主要研究方向：机器视觉算法；奚峥皓（1981-），男，博士，副教授，主要研究方向：机器视觉、路径规划、智能认知学习与控制等；陈健超（1997-），男，硕士研究生，主要研究方向：机器视觉算法。
通讯作者：奚峥皓 Email：zhenghaoxi@ hotmail.com

收稿日期：2022-04-27

尔可夫模型（Hidden Markov Model，HMM）建立了一种 HMMTxD 方法的跟踪器，将检测结果作为观测值输入到模型中，并输出估计跟踪的结果，对遮挡问题有较好的鲁棒性，然而此时模型受观测状态的影响较大，状态值的选取直接影响了跟踪结果的好坏。

基于马尔可夫在跟踪任务中的优越性，针对 HMM 状态选取的问题，本文以 SIFT 算子作为模型中的状态，建立一种基于 SIFT 的隐马尔可夫模型（SIFT-HMM），并将其应用到多目标跟踪任务中。针对状态关联问题，利用 HMM 设计一种匹配 SIFT 隐性、观测两种状态的方法，从而实现对遮挡目标的状态估计。针对马尔可夫模型复杂度较高的问题，利用 SIFT 关键点设计了一种高维的观测方法，从而增强模型的推理速度。实验结果表明，本文的跟踪器在 MOT17，MOT20、KITTI 公共数据集上实现较好的跟踪性能，高维模型能提高算法的推理速度。

1 基于 SIFT-HMM 的多目标跟踪器

通过 HMM 实现多目标跟踪的过程被描述为根据当前帧的目标状态，获取下一帧目标发生概率最大的状态。本文提出了一种以 SIFT 关键点为状态，建立 HMM 跟踪器模型，通过对状态值进行实时预测实现对多目标的跟踪，SIFT-HMM 跟踪器的工作流程如图 1 所示。

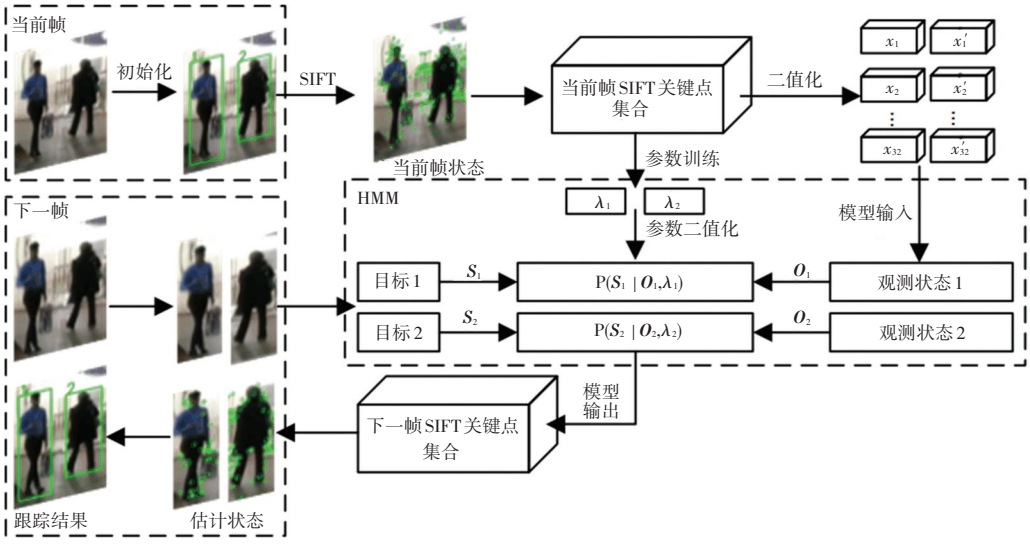


图 1 基于 SIFT-HMM 的跟踪器工作流程

Fig. 1 Workflow of the SIFT-HMM tracker

1.1 基于 SIFT 算法的特征关键点提取

同一目标在相邻帧间的传递过程可用 SIFT 关键点转移来描述。通过 SIFT 算法对单帧图像处理提取得到 $R(R \in \boldsymbol{Z}^+)$ 个 SIFT 特征关键点，$\boldsymbol{X}_r^{'}$ 表示第 r 个关键点的描述子向量，如式(1)所示：

$$\boldsymbol{X}_r^{'} = [x_1^{'}, x_2^{'}, \cdots, x_n^{'}], n \in \boldsymbol{Z}^+ \tag{1}$$

其中，$r \leqslant R$，$x_n^{'}$ 表示向量中每个维度信息的权值。

在尺度空间中，每个关键点可以通过其附近2×2邻域内的梯度信息来描述。同时，每个子域包含 8 个方向的梯度信息，故 $\boldsymbol{X}_r^{'}$ 可描述为一个 2×2×8=32 维的向量，即式(1)可改写为 $\boldsymbol{X}_r^{'} = [x_1^{'}, x_2^{'}, \cdots, x_{32}^{'}]$。

初始化所得目标由独立的回归框框选，SIFT 算法从这些回归框中提取关键点，单一回归框中往往包含多个关键点。令视频序列由 $T(T \in \boldsymbol{Z}^+)$ 帧图像组成，在第 $t(t \leqslant T)$ 帧图像中有 $K(K \in \boldsymbol{Z}^+)$ 个目标，此时第 k 个 $(k \leqslant K)$ 回归框中的 SIFT 状态可表示为 $\boldsymbol{o}_t^{'(k)} = [\boldsymbol{X}_1^{'}, \boldsymbol{X}_2^{'}, \boldsymbol{X}_3^{'} \cdots]$，由 K 个目标组成的 SIFT 状态集合即为第 t 帧时的关键点集合 $\boldsymbol{o}_t^{'} = [\boldsymbol{o}_t^{'(1)}, \boldsymbol{o}_t^{'(2)}, \cdots, \boldsymbol{o}_t^{'(K)}]$。

将每帧由 SIFT 直接提取的关键点集合 $\boldsymbol{o}_t^{'}$ 作为 HMM 输入的观测状态，输入到模型前，需要对关键点集合进行二值化处理，从而满足马尔可夫收敛的约束[7]。

定义 $\boldsymbol{\Omega}, \boldsymbol{\Psi}$ 两个集合如式(2)所示：

$$\begin{aligned} \boldsymbol{\Omega} &= \{x_n^{'} \mid x_n^{'} \geqslant 0.8x_{\max}^{'}, n \in [1,32]\} \\ \boldsymbol{\Psi} &= \{x_n^{'} \mid x_n^{'} < 0.8x_{\max}^{'}, n \in [1,32]\} \end{aligned} \tag{2}$$

其中，$x_{\max}^{'} = \max\{x_1^{'}, \cdots, x_{32}^{'}\}$ 表示 32 维描述符中权值最大的维度，$x_{\max}^{'} \in \{x_1^{'}, \cdots, x_{32}^{'}\}$ 二值化后的权值 x_n，如式(3)所示：

$$x_n = \begin{cases} 1 & x_n^{'} \in \boldsymbol{\Omega} \\ 0 & x_n^{'} \in \boldsymbol{\Psi} \end{cases} \tag{3}$$

二值化后的权值组成的向量 $\boldsymbol{X}_r$ 是一个只包含 0 和 1 的 32 向量，即 $\boldsymbol{X}_r = [x_1, \cdots, x_{32}]$，二值化后的 SIFT 关键点在 t 帧时第 k 个回归框的 SIFT 状态表示为 $\boldsymbol{o}_t^{(k)} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3 \cdots]$，在 t 帧的二值化 SIFT 集合为 $\boldsymbol{o}_t = [\boldsymbol{o}_t^{(1)}, \cdots, \boldsymbol{o}_t^{(K)}]$。

1.2　基于 SIFT 关键点建立的 HMM

HMM 由观测状态序列 $\boldsymbol{O}$、隐性状态序列 $\boldsymbol{S}$ 和参数 λ 组成。观测状态是实时状态可以观测的，用 SIFT 提取的关键点集表示观测状态，可得在 t 帧时的 SIFT 关键点集合为 $\boldsymbol{o}_t$，那么由 T 帧序列组成的视频的观测序列为 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3, \cdots, \boldsymbol{o}_T\}$。隐性状态是不可被直接观测的未知状态，用 $t+1$ 帧的估计 SIFT 关键点集表示 t 帧的隐性状态，那么由 T 帧序列组成的视频的隐性状态序列为 $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \cdots, \boldsymbol{s}_T\}$，HMM 中模型单帧的状态转移结构，如图 2 所示。

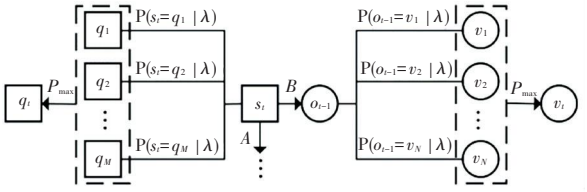


图 2　HMM 中单帧状态转移的结构图

Fig. 2　Structure of one frame state transition in HMM

此外，视频首帧图像的隐性状态与观测状态都表现为检测器检测得到的初始化结果，此时 $\boldsymbol{s}_1$ 不具有实际意义，在数值上 $\boldsymbol{s}_1$ 取值与 $\boldsymbol{o}_1$ 一致；而视频在最后一帧的图像是跟踪器的终止帧，模型在最后一帧时的观测状态 $\boldsymbol{o}_T$ 不具有实际意义，在数值上取值与 $\boldsymbol{s}_T$ 一致，故模型实际输入的两个状态序列为 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_{T-1}\}$ 和 $\boldsymbol{S} = \{\boldsymbol{s}_2, \boldsymbol{s}_3, \cdots, \boldsymbol{s}_T\}$。

跟踪器工作时，当前帧的目标状态是由上一帧的状态决定的，即 t 帧目标的状态 $\boldsymbol{s}_t$ 与 $t-1$ 帧时的观测状态 $\boldsymbol{o}_{t-1}$ 有关，当 $t \in [2, T]$，用 $\boldsymbol{F}_t$ 表示第 t 帧时模型的状态向量，如式(4)所示：

$$\boldsymbol{F}_t = [\boldsymbol{o}_{t-1}, \boldsymbol{s}_t, \lambda] \tag{4}$$

其中，模型的参数 $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi})$，分别表示状态转移概率矩阵、观测概率矩阵和状态概率向量。

t 帧状态 $\boldsymbol{o}_{t-1}$ 所能取到的观测状态数量是有限的，此时的状态数量为 $N(N \in \boldsymbol{Z}^+)$，由于 $\boldsymbol{o}_{t-1}$ 与 $\boldsymbol{s}_t$ 存在矩阵变换关系，故隐性状态 $\boldsymbol{s}_t$ 所能取到的状态数量也是有限的，隐性状态数量为 $M(M \in \boldsymbol{Z}^+)$。用 $\boldsymbol{V} = \{v_1, v_2, \cdots, v_N\}$ 表示所有的观测状态集合，$\boldsymbol{Q} = \{q_1, q_2, \cdots, q_M\}$ 表示所有的隐性状态集合，则 $P(\mid \boldsymbol{o}_{t-1} \mid = v_t)$ 表示观测状态 $\boldsymbol{o}_{t-1}$ 取到 v_t 的概率，如式(5)所示，$P(\mid \boldsymbol{s}_t \mid = q_t)$ 表示 t 帧时隐性状态 $\boldsymbol{s}_t$ 取到 q_t 的概率，如式(6)所示：

$$P(\mid \boldsymbol{o}_{t-1} \mid = v_t) = \max\{P(\mid \boldsymbol{o}_{t-1} \mid = v_1), \cdots, P(\mid \boldsymbol{o}_{t-1} \mid = v_N)\} \tag{5}$$

$$P(\mid \boldsymbol{s}_t \mid = q_t) = \max\{P(\mid \boldsymbol{s}_t \mid = q_1), \cdots, P(\mid \boldsymbol{s}_t \mid = q_M)\} \tag{6}$$

其中，$v_t \in \boldsymbol{V}, q_t \in \boldsymbol{Q}$。

状态在相邻帧之间的转移是通过参数 λ 的变换关系来传递的，且参数包含 3 个子参数，即 $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi})$ 分别表示状态转移概率矩阵，观测概率矩阵，状态概率向量，如式(7)~式(9)所示：

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{bmatrix}_{M \times M} \tag{7}$$

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1N} \\ b_{21} & b_{22} & \cdots & b_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ b_{M1} & b_{M2} & \cdots & b_{MN} \end{bmatrix}_{M \times N} \tag{8}$$

$$\boldsymbol{\Pi} = [\pi_1, \pi_2, \cdots, \pi_M]_{1 \times M} \tag{9}$$

其中，a_{ij} 表示 $t-1$ 帧的状态 q_i 转移到 t 帧的状态 q_j 的概率，$a_{ij} = P(\mid \boldsymbol{s}_t \mid = q_j \mid\mid \boldsymbol{s}_{t-1} \mid = q_i)$；$b_{ij}$ 表示 t 帧隐性状态 q_i 产生观测值 v_j 的概率，$b_{ij} = P(\mid \boldsymbol{o}_{t-1} \mid = v_j \mid\mid \boldsymbol{s}_t \mid = q_i)$；$\pi_i$ 表示初始帧的状态 q_i 出现的概率；$\pi_i = P(\mid \boldsymbol{s}_1 \mid = q_i)$，下角标 i, j 表示可能状态的标号。

确定每一帧状态的取值以及帧与帧之间传递关系后，HMM 可以用 $\boldsymbol{O}, \boldsymbol{S}, \lambda$ 之间的关系来表示，如式(10)所示：

$$P(\boldsymbol{O}, \boldsymbol{S} \mid \lambda) = P(\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_T) = P(\boldsymbol{F}_2, \cdots, \boldsymbol{F}_{T-1}, \boldsymbol{o}_T, \boldsymbol{s}_1 \mid \lambda) = P(\boldsymbol{o}_1, \cdots, \boldsymbol{o}_{T-1}, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T, \boldsymbol{o}_T, \boldsymbol{s}_1 \mid \lambda) = \pi_{s_1} b_{s_1 o_1} a_{s_1 s_2} b_{s_2 o_2} \cdots a_{s_{t-1} s_t} b_{s_t o_t} \tag{10}$$

1.3　SIFT-HMM 的高维观测状态

1.3.1　高维观测状态的描述

基于 HMM 建立的多目标跟踪器一般对每个目标建立独立的 HMM，状态表示为每个目标的全部关键点集，此时的观测维度较低，将其定义为初始的低维状态模型。

在 SIFT-HMM 中，基于 SIFT 关键点得到的观测序列 $\boldsymbol{O}$ 的每一个观测状态值 $\boldsymbol{o}_t$ 包含了全部目标的关键点集，即 $\boldsymbol{o}_t$ 包含了全部 $R = \sum_{k=1}^{K} R_k$ 个关键点的信息。为全部目标建立的 HMM 中，状态表示为全

部目标的全部关键点集,此时的观测维度较高,将其定义为优化后的高维状态模型,优化高维模型的原理如图 3 所示。

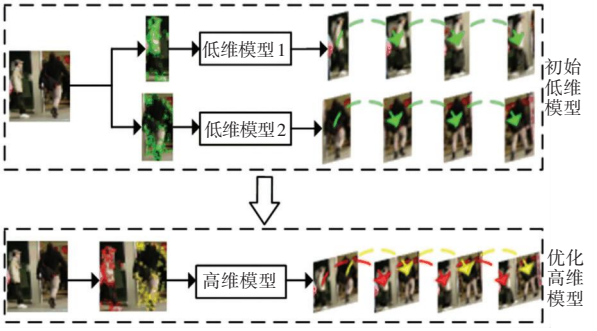


图 3 优化高维模型的原理图

Fig. 3 Schematic diagram of the improved high-dimensional model

高维模型中,$\boldsymbol{o}_t$ 表示 t 帧 K 个目标的观测状态,包含目标数量、关键点数量、关键点维度 3 重信息。t 帧时模型的状态向量 $\boldsymbol{F}_t$ 包含 K 个目标和 R 个关键点,每个关键点在形式上表现为一个 $m \times n$ 的二维矩阵,那么 $\boldsymbol{o}_t$ 在形式上表示为一个 $K \times R \times m \times n$ 的四维数组,将高维数组 $\boldsymbol{o}_t$ 进行逐层分解,则有式(11):

$$\boldsymbol{o}_t = [\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \boldsymbol{o}_t^{(3)}, \cdots, \boldsymbol{o}_t^{(K)}] \tag{11}$$

其中,$\boldsymbol{o}_t^{(k)}$ 表示 t 帧时第 k 个目标的观测值,表现为 $R \times m \times n$ 的三维数组,如式(12)所示:

$$\boldsymbol{o}_t^{(k)} = [\boldsymbol{o}_t^{(k)}|_1, \boldsymbol{o}_t^{(k)}|_2, \boldsymbol{o}_t^{(k)}|_3, \cdots, \boldsymbol{o}_t^{(k)}|_R]^{\mathrm{T}} \tag{12}$$

其中,$\boldsymbol{o}_t^{(k)}|_r$ 表示 t 帧时第 k 个目标在第 r 个关键点的观测值,表现为 $m \times n$ 的二维矩阵,如式(13)所示:

$$\boldsymbol{o}_t^{(k)}|_r = \begin{bmatrix} x_{11}(t) & \cdots & x_{1n}(t) \\ \vdots & \cdots & \vdots \\ x_{m1}(t) & \cdots & x_{mn}(t) \end{bmatrix} \tag{13}$$

关键点由 1×32 的描述子向量表示,即 $m \times n = 1 \times 32$,故 $\boldsymbol{o}_t^{(k)}|_r$ 可改写为式(14):

$$o_t^{(k)}|_r = [x_1(t), x_2(t), \cdots, x_{32}(t)] \tag{14}$$

根据式(11)~式(14),t 帧时 $K \times R \times m \times n$ 的观测状态 o_t 表示为 $K \times R \times n(n=32)$ 的三维数组,如式(15)所示:

$$\boldsymbol{o}_t = \begin{bmatrix} \begin{bmatrix} [x_1^{(1)}(t)|_1 & \cdots & x_n^{(1)}(t)|_1] \\ \vdots & & \\ [x_1^{(1)}(t)|_{R_1} & \cdots & x_n^{(1)}(t)|_{R_1}] \end{bmatrix} \\ \vdots \\ \begin{bmatrix} [x_1^{(K)}(t)|_1 & \cdots & x_n^{(K)}(t)|_1] \\ \vdots & & \\ [x_1^{(K)}(t)|_{R_K} & \cdots & x_n^{(K)}(t)|_{R_K}] \end{bmatrix} \end{bmatrix} \tag{15}$$

1.3.2 高维观测状态模型的复杂度

低维状态模型和高维状态模型表现在状态维度的不同,而状态维度决定了参数的计算量,从而影响跟踪过程的推理速度。低维状态模型中每帧的状态模型针对一个目标,多个目标的场景需要多个模型同时运行,其参数总计算量等于每个模型对应参数的乘积,用 μ_{low} 表示模型的复杂度,如式(16)所示:

$$\mu_{\mathrm{low}} = \prod_{k=1}^{K} \prod_{t=1}^{T} \alpha_k(t) \tag{16}$$

其中,$\alpha_k(t)$ 表示第 k 个目标在 t 帧的复杂度。

高维状态模型中,每帧的状态模型针对其全部目标,多个目标的场景只需要一个模型运行,其参数计算量等于该模型的参数计算量,用 μ_{high} 表示此时模型的复杂度如式(17)所示:

$$\mu_{\mathrm{high}} = \sum_{k=1}^{K} \prod_{t=1}^{T} \alpha_k(t) \tag{17}$$

低维模型的复杂度 μ_{low} 表现为多个目标复杂度的积,高维模型复杂度 μ_{high} 表现为多个目标复杂度的和,则 $\mu_{\mathrm{high}} \leqslant \mu_{\mathrm{low}}$,表明高维状态模型能一定程度的降低模型的复杂度。

1.4 SIFT-HMM 的参数训练和求解

Baum - Welch 方法是 HMM 参数学习的通用方法之一,使用该方法对 SIFT-HMM 的参数训练[8]。若视频序列中一共存在 K 个目标,初始帧图像中有 k_1 个目标,对这 k_1 个目标建立高维观测状态模型。该模型单独训练,得到高维模型的参数。而对于在后续帧中出现的 $K - k_1$ 个目标,初始化每个目标并建立低维观测状态模型,得到独立的 $K - k_1$ 个模型并进行训练。综合上述,1 个高维模型和 $K - k_1$ 个低维模型,最终能够得到整个 SIFT-HMM 的 $K - k_1 + 1$ 个收敛的参数。对于第 θ 个模型,$\theta \in [1, K - k_1 + 1]$,该模型收敛的参数 $\lambda_\eta(\theta)$ 可由式(18)得到:

$$\begin{aligned} &\max \boldsymbol{A}_\eta(\theta) = \\ &\left[\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=1}^{T-1} \lg a_{ij} P(\boldsymbol{O}, \boldsymbol{s}_t = q_i, \boldsymbol{s}_{t+1} = q_j \mid \lambda(\theta))\right] \Big| \sum_{j=1}^{M} a_{ij} = 1 \\ &\max \boldsymbol{B}_\eta(\theta) = \left[\sum_{i=1}^{M}\sum_{t=1}^{T} \lg b_{i o_{t-1}} P(\boldsymbol{O}, \boldsymbol{s}_t = q_i \mid \lambda(\theta))\right] \Big| \sum_{t=1}^{T} b_{i o_{t-1}} \\ &\max \boldsymbol{\Pi}_\eta(\theta) = \left[\sum_{i=1}^{M} \lg \pi_i P(\boldsymbol{O}, \boldsymbol{s}_1 = q_i \mid \lambda(\theta))\right] \Big| \sum_{i=1}^{M} \pi_i = 1 \end{aligned} \tag{18}$$

其中,$\boldsymbol{A}_\eta(\theta), \boldsymbol{B}_\eta(\theta), \boldsymbol{\Pi}_\eta(\theta)$ 表示收敛的 3 个参数;$\max \boldsymbol{A}_\eta(\theta), \max \boldsymbol{B}_\eta(\theta), \max \boldsymbol{\Pi}_\eta(\theta)$ 表示 3 个参数的极大值;η 表示模型的最大迭代次数。

将式(18)得到的 3 个参数以及输入观测序列 $\boldsymbol{O}$ 代入到式(10)的计算关系式中可以得到隐性状态 $\boldsymbol{s}_t$ 的最大概率状态值如式(19)所示：

$$P(\boldsymbol{s}_t \mid \lambda(\theta)) = \pi_{s_1} \prod_{\tau=1}^{t} b_{s_\tau o_\tau} a_{s_{\tau-1} s_\tau} \tag{19}$$

该概率最大的状态即是 t 帧时 $\boldsymbol{s}_t$ 的状态，逐帧计算全部 T 帧的状态，从而得到整个视频的状态序列 $\boldsymbol{S}$，将这些估计的状态(SIFT 关键点集)用一个可视的回归框框选出来，形成关于时序 t 的跟踪轨迹，即得到 SIFT-HMM 的跟踪可视化结果。

2　基于检测器的初始化预处理方法

SIFT-HMM 跟踪算法是根据初始帧选定的目标框，实现对后续帧中目标的跟踪。在建立跟踪器之前，需要对图片序列进行预处理，得到初始化结果。本文的预处理即建立一个检测器，得到图片序列的初始化目标框。选用 MOT Challenge 的 DPM 公共检测器作为检测框架，在输入到跟踪器模型前用该检测器进行目标的初始化定位，DPM 检测器在 MOT17-10-DPM 序列的首帧检测结果如图 4 所示。检测器利用回归框来框选检测到的目标，而在得到待测目标的同时，图 4(a)中出现了许多错检的回归框，这些错检的回归框在验证跟踪器效果时不具有实际意义，因此将第一帧图像中错检的回归框剔除，剔除异常后的检测结果如图 4(b)所示，优化后的检测结果将作为跟踪器首帧的输入。

(a) DPM 检测器检测结果　　(b) 剔除异常后的检测结果

图 4　MOT17-10-DPM 首帧检测结果

Fig. 4　Detected result of MOT17-10-DPM in 1st frame

3　实验结果分析与算法评估

在 MOT17、MOT20 和 KITTI 公共数据集上对 SIFT-HMM 的性能进行评估，并与其他先进的跟踪算法进行比较分析。实验环境为一台具有 Intel Corei5-9600KF CPU 处理器和 16 GB 内存的个人电脑，算法是通过 python3.8 在 64 位 Windows10 操作系统上实现的。

3.1　数据集和评价指标

3.1.1　数据集

选取了 3 个数据集中具有各自特点的视频序列来检验跟踪器的有效性。MOT17 数据集是摄像机在室外环境下拍摄到的行人视频，选取 MOT17-08-DPM 视频序列用于检验在拥挤、遮挡环境下的跟踪性能；选取 MOT17-10-DPM 视频序列用于检验在动态场景下的跟踪性能；MOT20 数据集中目标数量较多，选取 MOT20-01、MOT20-02 数据集用于检验高维模型的优越性；KITTI 数据集是针对自动驾驶的数据集，选取 KITTI-0006 用于检验以刚体为目标的跟踪性能。

3.1.2　评价指标

为了量化跟踪器结果，并且统一与其他对比算法的评价标准，采用 IDS、MOTA、MOTP3 个指标来量化多目标跟踪器的性能。其中，IDS 表示在一条跟踪轨迹跟踪的过程中，目标身份发生交换的次数；MOTA 表示多目标跟踪的准确率，如式(20)所示；MOTP 表示多目标跟踪的精度，反应了预测结果与标准框的匹配度，如式(21)所示：

$$\mathrm{MOTA} = 1 - \frac{FN + FP + IDS}{GT} \tag{20}$$

$$\mathrm{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_{t} c_t} \tag{21}$$

其中，FN、FP、GT 表示跟踪中漏检的目标数、错检的目标数、全部目标的实际总数；$d_{t,i}$ 表示第 t 帧中第 i 个目标对的距离；c_t 表示 t 帧中预测轨迹与 GT 匹配上的数目。

3.2　实验结果分析

3.2.1　可视化结果分析

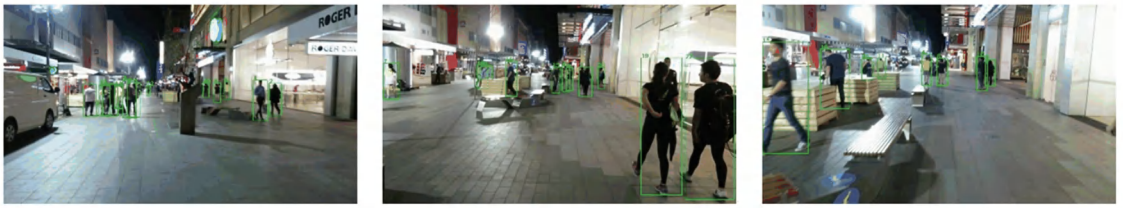
SIFT-HMM 在部分数据集的可视化结果如图 5 所示。可视化结果表明，在 3 种不同的场景下，SIFT-HMM跟踪器都能对目标实现有效的跟踪。

3.2.2　优化的高维状态模型结果分析

SIFT-HMM 在 MOT20-02 第 25 帧的跟踪可视化结果如图 6 所示，该图中目标较多，对初始帧每个目标建立独立的 HMM 会导致此时的跟踪器复杂度较高。

为了验证本文提出的优化高维模型的优越性，控制两个变量来设置消融实验，进而比较优化前的初始低维模型和优化后的高维观测模型的复杂度。在同一序列下截取一段视频，比较在相同视频下两个模型的运行时间，量化运行时间的指标为 HZ，表示跟踪器在一秒中处理帧数的速度；保持其他条件不变的情况下，增加截取视频的帧数，比较在增加帧数之后运行时间的变化。

(a) MOT17-08-DPM(从左到右为:#55,#103,#203)



(b) MOT17-10-DPM(从左到右为:#5,#245,#405)

(c) KITTI-0006(从左到右为:#50,#102,#151)

图 5 SIFT-HMM 在部分数据集的可视化结果

Fig. 5 Visualization results of SIFT-HMM in some sequences

图 6 MOT20-02 第 25 帧的跟踪可视化结果

Fig. 6 Visualization results in MOT20-02, #25 frame

共选取了 3 个视频,分别从每个视频初始帧截取 30 帧的视频段和 120 帧的视频段做两次测试,测试的结果见表 1。实验结果表明,基于高维模型的 SIFT-HMM 跟踪器在 6 个测试的 HZ 上都表现最优。在 MOT20-01 数据集的两次测试中,基于高维模型的跟踪器分别比初始模型在运行速度上提升 94.11%和 68.75%;在 MOT20-02 数据集的两次测试中,基于高维模型的跟踪器分别比初始模型在运行速度上提升 183.33%和 126.92%;在 MOT17-08-DPM 数据集的两次测试中,基于高维模型的跟踪器分别比初始模型在运行速度上提升 19.28%和 4.51%。这些结果表明,高维优化模型能够有效的降低跟踪器的复杂度,提升跟踪器的推理速度。

表 1 SIFT-HMM 高维模型与初始低维模型的消融实验

Tab. 1 Ablation experiments of high-dimensional model and initial model of SIFT-HMM

视频序列	初始帧目标数量	方法	测试 1(30 帧)		测试 2(120 帧)	
			时间/s ↓	HZ ↑	时间/s ↓	HZ ↑
MOT20-01	23	低维	87.32	0.34	377.16	0.32
		高维	45.73	0.66	223.12	0.54
MOT20-02	26	低维	125.78	0.24	469.23	0.26
		高维	44.12	0.68	200.15	0.59
MOT17-08-DPM	8	低维	10.71	2.80	31.86	3.77
		高维	8.99	3.34	30.43	3.94

3.3 与其他跟踪算法的对比分析

为了验证 SIFT-HMM 跟踪器的优越性,本文选择与 MOT-MDP[4], MDP-OF[9], HMOT[10], JDMOT[11]跟踪算法进行对比分析,对比结果见表 2。实验结果表明,对于 IDS 指标,本文在这 3 个视频序列上都取得最好的性能,比第二好的跟踪器在 MOT 数据集上高出 33.3%~70.5%;在 KITTI 数据集上高出 6.7%。这是由于本算法为每个目标建立的 HMM,使每个目标的转移通过独立的马尔可夫链来传递,具有较好的独立性。

表 2　与其他跟踪算法的对比结果

Tab. 2　Comparison results with other trackers

视频序列	算法	FN ↓	FP ↓	IDS ↓	MOTA ↑/%	MOTP ↑/%
MOT17-08-DPM	MOT-MDP[4]	12 505	1 224	195	37.7	79.8
	MDP-OF[9]	13 988	543	234	33.9	71.1
	HMOT[10]	9 091	749	460	53.9	67.9
	JDMOT[11]	11 230	370	75	47.7	72.1
	SIFT-HMM	9 444	294	44	56.2	75.7
MOT17-10-DPM	MOT-MDP[4]	4 200	574	121	42.1	71.1
	MDP-OF[9]	3 452	411	144	52.5	62.2
	HMOT[10]	2 999	586	182	55.4	61.3
	JDMOT[11]	3 890	616	36	46.3	73.7
	SIFT-HMM	3 113	367	27	58.5	69.8
KITTI-0006	MOT-MDP[4]	92	23	19	84.6	89.1
	MDP-OF[9]	122	18	24	81.2	84.5
	HMOT[10]	28	32	21	90.7	88.4
	JDMOT[11]	112	8	16	84.4	92.2
	SIFT-HMM	76	25	15	86.7	85.4

MOTA 受 *FN*、*FP*、*IDS* 3 个指标影响。HMOT 算法在个体关联上引入了细化运动模式的方法，减少了算法的误阳性数据，在 *FN* 指标上有较好的效果，但在目标轨迹交叉时容易产生误差，*IDS* 指标较差。本文的 *IDS* 指标在 MOT 数据集上平均比 HMOT 高 7 倍，在 KITTI 数据集上平均比 HMOT 高出 42.7%。因此，本文在 MOT17 的两个序列上取得最好的性能，而在 KITTI-0006 序列上取得第二好的性能，比最好的 HMOT 算法低出 4.6%。

本文的 MOTP 指标在 MOT17-08-DPM 取得第二好的性能，而在 KITII 数据集上取得第四的性能，这是因为以 SIFT 关键点建立 HMM 时，模型容易受背景特征点的干扰，导致框选目标时多框选了背景的特征点造成目标框与 GT 的不匹配。

4　结束语

本文提出了一种 SIFT-HMM 的多目标跟踪算法，该算法能够有效解决多目标跟踪过程的遮挡，以及目标身份互换的问题，在 MOT20 数据集上通过实验检验高维 SIFT 观测状态优越性，实验结果表明在视频序列首帧图像采用高维观测状态建立的 SIFT-HMM，能够降低模型的复杂度。在 MOT17 和 KITTI 数据集上与其他跟踪算法进行了对比分析，对比结果表明，本文算法的 *IDS* 指标在两个数据集上表现最优，比第二优的算法高出 34.3% 和 76.2%，表明本文的算法对目标身份互换问题有较好的鲁棒性；MOTA 指标在 2 个数据集上分别取到 56.84% 和 86.7%的性能，在 MOT17 数据集上比第二优的算法高出 4.7%，表明本文的算法具有较好的跟踪准确度，并对遮挡问题有较好的鲁棒性。然而，本文在 MOTP 指标上未能表现最优，因此约束 SIFT 关键点来提高跟踪的精度会是未来研究的方向。

参考文献

[1] YIN J, WANG W, MENG Q, et al. A unified object motion and affinity model for online multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6768-6777.

[2] CIAPARRONE G, SÁNCHEZ F L, TABIK S, et al. Deep learning in video multi-object tracking: A survey[J]. Neurocomputing, 2020, 381: 61-88.

[3] LIU P, LI X, WANG Y, et al. Multiple object tracking for dense pedestrians by Markov random field model with improvement on potentials[J]. Sensors, 2020, 20(3): 628.

[4] XIANG Y, ALAHI A, SAVARESE S. Learning to track: Online multi-object tracking by decision making[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4705-4713.

[5] WU C, SUN H, WANG H, et al. Online multi-object tracking via combining discriminative correlation filters with making decision[J]. IEEE Access, 2018, 6: 43499-43512.

[6] VOJIR T, MATAS J, NOSKOVA J. Online adaptive hidden markov model for multi-tracker fusion[J]. Computer Vision and Image Understanding, 2016, 153(12): 109-119.

（下转第 199 页）