doi:10.11887/j.cn.201402019 http://journal.nudt.edu.cn

基于检测的人体跟踪算法*

吴建宅,陈芳林,胡德文
(国防科技大学 机电工程与自动化学院,湖南 长沙 410073)

摘 要:传统的目标跟踪算法需要人为标定跟踪区域,且受到漂移问题的影响。为了解决这些困难,针对人体跟踪问题,提出了一种新的基于检测的跟踪算法。为了减少漏跟踪,使用了多个检测算子,用来定位多个身体部位,将其检测结果映射到一个相同的身体区域。为了适应快速运动的目标,使用KLT跟踪和凝聚聚类将检测窗口连接起来形成人体轨迹。实验结果表明:使用多个检测算子明显地提高了跟踪性能;KLT跟踪对于快速运动目标具有适应能力。该算法基本满足实时性。

关键词:目标跟踪;人体检测;KLT跟踪;凝聚聚类;实时性
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2014)02-0113-05

A detection-based person tracking algorithm

WU Jianzhai, *CHEN Fanglin*, *HU Dewen*
(College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China)

Abstract: The traditional object tracking algorithms require manually annotated tracking area, and suffer from the problem of drift. To address these difficulties, the problem of person tracking was focused on, and a new detection-based tracking algorithm was proposed. To reduce failure in tracking, multiple detectors to locate multiple body parts were employed, and then their detection results were mapped to a common body area. To adapt for the quickly moving objects, the KLT tracker and agglomerative clustering for linking the detection windows to form person body trajectories was employed. The experimental results reveal that using multiple detectors improves the tracking performance significantly, and the KLT tracker is adaptable for quickly moving objects. This algorithm is nearly real-time.

Key words: object tracking; person detection; KLT tracker; agglomerative clustering; real-time

视频目标跟踪是计算机视觉领域内的一个比较活跃的研究课题。传统的目标跟踪算法[1-2]主要存在两点不足:(1)需要在第一帧人为标出待跟踪的目标区域;(2)由于目标表象和光照等各种变化因素,长时间跟踪时漂移会被无限放大。

为了解决以上问题,很多工作[3-6]提出了基于检测的跟踪方法。该方法主要包含两个步骤:首先,使用预先训练好的检测算子[7-8]在各帧内进行目标检测和定位;然后,对检测窗口进行链接(linking)形成目标轨迹。但是,这些方法具有以下两个主要的缺点:(1)跟踪性能不理想,丢失了很多目标;(2)速度较慢,不能满足实时性要求。

在本文,我们主要针对人体跟踪问题。为了提高跟踪性能,采用了多个人体检测算子,用来检测不同的人体部分(例如上半身或整个人体);并将不同算子得到的检测窗口映射到一个共同的参考窗口。通过这种方法,可以适应更多不同的姿态以及遮挡等情况。然后,我们利用KLT跟踪方法[9]和凝聚聚类算法(agglomerative clustering)将窗口连接起来形成轨迹;其中KLT跟踪对于快速运动目标具有很好的适应能力。为了提高速度,我们利用文献[10]提出的快速检测算子,但是该原始算子只能检测整个人体。因此,我们利用一种新的学习算法[8]对其重新进行训练,以获得多个部分检测算子。

我们在TVHI视频数据库[12]上对该方法进行测试。该数据库中的视频一般包含多个人体,且存在明显的姿态变化以及遮挡等情况。实验结果表明:使用多个检测算子能够十分明显地提高跟踪性能;KLT跟踪对于快速运动目标具有很好的适应能力。另外,我们的方法基本能够满足实时性要求,并且我们方法的跟踪性能优于文献中的相关算法。

* **收稿日期**:2013-08-01
基金项目:国家重点基础研究发展计划项目(2013CB329401);国家自然科学基金资助项目(61203263)
作者简介:吴建宅(1983—),男,河北保定人,博士研究生,E-mail:wjz_gfkd@163.com;
胡德文(通信作者),男,教授,博士,博士生导师,E-mail:dwhu@nudt.edu.cn

1 人体检测

1.1 使用多个检测算子

为了在视频中实现人体或目标检测，Patron-Perez 等[3]利用标准的 HOG 算子[7]检测人体的上半部分，而其他一些方法[4-6]采用文献[8]提出的算子检测整个人体。这些方法具有一个共同的缺点，即只检测一个人体部分。由于各种阻挡以及姿态变化等因素的影响，只检测一个人体部分并不能获得满意的性能。另外，HOG 算子[7]和文献[8]中的算子检测速度较慢，不能满足实时性要求。

本文中，我们使用文献[10]提出的快速检测算子。该算子利用了可以实现快速计算的积分特征以及尺度近似技术，从而大幅度地提高了检测速度。但是，该原始算子只能检测整个人体。为了获得多个部分检测算子，利用文献[8]提出的隐含 SVM 学习算法和 Pascal VOC2007 图像数据库[13]重新进行训练；最后得到了 3 个检测算子，分别对应于不同的人体部分：上半身(UB1)、上半身(稍大，UB2)和整个人体(FB)。为了减少漏检的情况，将检测算子的检测阈值设得很小。值得一提的是，文献[10]的方法并不是当前最快的，例如文献[1,11]所提出的方法据报道能够达到 100 帧/秒，但却需要双摄像头来构建深度图，因此本文不采用该算法。

1.2 窗口映射

这些算子得到的检测窗口对应于不同的人体区域，因此不能直接连接到一起。可以利用以下方法将它们映射到共同的参考框架。对于每一个检测算子，使用一组线性回归参数 $\alpha=(\alpha_1,\alpha_2,\alpha_3)$，将原始检测窗口 $w=(x,y,W,H)$(其中(x,y)代表矩形窗的左上角位置，W 和 H 分别代表宽度和高度)映射到一个新的方形窗口：

$$R(w,\alpha)=(x-W\alpha_1,y-H\alpha_2,W\alpha_3,W\alpha_3) \tag{1}$$

其中，参数 α 是在训练数据上学习得到的。给定 n 个检测窗口(w^i)以及对应的手动标识的真实方形窗口 h^i，可以通过优化以下目标函数[14]得到参数 α 的值：

$$\alpha^*=\arg\max_{\alpha}\sum_{i=1}^{n}\mathrm{IoU}[h^i,R(w^i,\alpha)] \tag{2}$$

其中，$\mathrm{IoU}(a,b)=|a\cap b|/|a\cup b|$代表窗口 a 与 b 间交集与并集之比。我们采用 ETHZ Pascal Stickmen 数据库[15]对 α 进行训练，该图像数据库中包含手动标示的人体窗口。

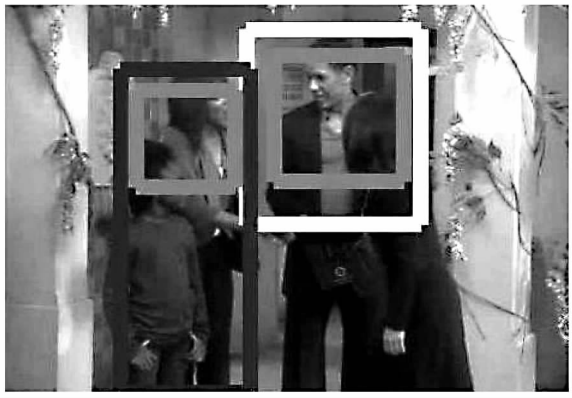


图1 使用多个检测算子

Fig. 1 Using multiple detectors

图1中显示了 FB(黑色框)和 UB1(白色框)返回的检测窗口，以及映射到共同的参考窗口之后的结果(灰色方框)。选定的共同参考窗口为：从头的上半部分到躯干中部的方形区域。

为了进一步提高速度，我们发现只需要在一部分帧内进行检测即可，因为相机抖动以及人体空间位置变化的频率一般都不会太快。

2 生成人体轨迹

为了对窗口进行连接从而生成人体轨迹，需要解决以下两个问题：首先，如何计算各帧窗口之间的连接程度；其次，如何对窗口进行连接。

2.1 计算连接度

首先进行以下设定：(1)相同帧内不同检测窗口之间连接度为0；(2)距离太远(间隔大于 t_d 帧)的两帧内所有窗口之间连接度为0。

然后，给定一对窗口 a 和 b，其连接度 s_{ab} 可以定义为：

$$s_{ab}=\begin{cases}s'_{ab}\cdot\lambda^{(|t_a-t_b|-1)} & \text{当 } 0<|t_a-t_b|\leqslant t_d\\ 0 & \text{其他}\end{cases} \tag{3}$$

其中，t_a 和 t_b 分别代表检测窗口 a 和 b 所在的帧，参数 λ 满足：$0<\lambda<1$，代表对不相邻帧进行惩罚。我们设定 $t_d\geqslant 2$，从而可以适应暂时的遮挡及检测失败等情况。s'_{ab} 可以定义为窗口 a 和 b 交集与并集之比：

$$s'_{ab}=\begin{cases}|a\cap b|/|a\cup b| & \text{当 } |a\cup b|>l\\ 0 & \text{其他}\end{cases} \tag{4}$$

其中，$|a\cap b|$代表同时穿过 a 和 b 的点轨迹数目，$|a\cup b|$代表至少穿过 a、b 中任意一个的点轨迹数目。当 $|a\cup b|$很小时会导致结果不准确，因此我们设定：当 $|a\cup b|\leqslant l$ 时，s'_{ab} 为0，其中 $l\geqslant 1$。我们

利用 KLT[9] 点跟踪方法来计算点轨迹。与一般使用的基于窗口之间面积重叠比例的方法相比,使用点跟踪方法能够适应高速运动的目标。

2.2　连接窗口

本文采用凝聚聚类算法(agglomerative clustering)对各帧内的窗口进行连接。文献[6]采用网络流方法计算全局最优的人体轨迹,但却设定只能采用窗口重叠率来计算连接度。文献[3]也采用了凝聚聚类方法,与之不同的是,我们采用了最大化抑制方法来去除冗余轨迹,因此能够更好地适应多检测算子的情况。

该聚类算法中,每一簇(cluster)构成一条人体轨迹。首先进行初始化:每一个窗口构成一簇。然后,在以下两个步骤之间进行迭代:(1)搜索具有最大连接度的簇对,并将其连到一起;(2)通过以下方法对簇对之间的链接度进行更新。给定两个簇$(a_1,\cdots,a_{n_a})$与$(b_1,\cdots,b_{n_b})$(其中 n_a 与 n_b 代表簇内窗口数目),如果二者间没有任意共同穿过的帧,则链接度设为 $\max_{i,j} s_{a_i,b_j}$,否则为 0。该聚类算法直到找不到任意链接度大于 $\delta(0\leqslant\delta<1)$ 的簇对为止。

对于所得到的每一条人体轨迹,计算所有相连窗口之间的链接度之和,用来表示该轨迹的显著性,并且按照显著性分数对所有轨迹进行排序。然后,采用非最大化抑制方法(non-max suppression)去除那些属于同一目标的多余轨迹。在帧下采样的情况下,通过近邻插补方法填充那些未采样到的帧。

3　实验分析与比较

3.1　数据库与实验设计

TVHI 数据库[12] 共有 300 个视频,是从 23 个电视节目片段中提取而来。该数据库主要包含 4 种行为:握手、击掌、拥抱和亲吻,如图 2 所示。视频的长度范围从 30 到 600 帧,并且存在大量的变化,如在每个场景中的人体个数、尺度、大小和摄像机角度变化(可能存在突然的视点变化)。

该数据库中的每幅视频都人为地标注了人体位置(用矩形框出上半身)。根据这些人体位置,可以得到真实的人体轨迹。然后,基于这些真实的人体轨迹对我们的跟踪算法进行测试。

以前,很多基于检测的跟踪方法[3,6] 通过每帧中目标检测的精度来度量其跟踪算法的性能。这种度量方法可以用来表征目标检测的精度,但并不能很好表征整条轨迹的好坏。为了解决这个问题,我们提出了一种新的直接基于整条轨迹的度量方法。假设 V_a 是一条手动标识的目标轨迹,具有 N_a 个人体框;另外 V_b 为一条计算得到的目标轨迹。如果 V_a 中至少 $\alpha\cdot N_a$ 个人体框被 V_b 以交合比(IoU)大于 β 的方式覆盖,则我们定义 V_a 被 V_b 检测到。其中,设定 $0<\alpha,\beta<1$。

图 2　TVHI 视频数据库

Fig. 2　TVHI video dataset

3.2　检测算子对性能的影响

为了检验多个检测算子的影响,我们分别测试了以下 4 种设置:UB1、UB2、FB 和 UB1 + UB2 + FB,并在 图 3 中显示了这些设置的精度 - 检测率曲线。我们设定窗口检测率 $\alpha=0.5$,并采用了几种不同的重叠比例 $\beta=0.1/0.3/0.5$。从这些曲线中可以观察到以下一些有趣的结果。

首先,对于不同的 β 值, UB1 + UB2 + FB 的性能比单独使用任何一个检测算子(UB1/UB2/FB)都要好。例如,对于 $\beta=0.1/0.3/0.5$,UB1 + UB2 + FB 的平均精度(AP)比使用单个检测算子的最好结果分别高出 0.055/0.115/0.059。

第二,广泛使用的行人检测算子 FB 的性能远逊于 UB1 和 UB2。例如,当 $\beta=0.1$ 时,UB1 和 UB2 获得的 AP 值分别为 0.777 和 0.748,而 FB 获得的 AP 值为 0.329。

第三,即使使用多个检测算子,广泛使用的重叠比例 $\beta=0.5$ 也导致了很小的 AP 值 0.193 。造成该结果的原因可以总结如下:(1)这三个人体检测算子都是预先在静止图像数据库[13] 中训练得到的,而在自然视频中人体外观变化往往比图像数据库中更为显著;(2)共同的参考窗口(从头的上半部分到躯干中部的方形区域)与该数据库所标注的人体上半身窗口并不完全匹配。因此,重叠比例很小的检测窗口也可能代表正确的结果,只要它们可以被稳定地在连续帧内跟踪。从图 3 可以观察到,使用较低的重叠比例

$\beta = 0.1(0.3)$，可能得到相对较大的 AP 值 0.832 (0.633)。

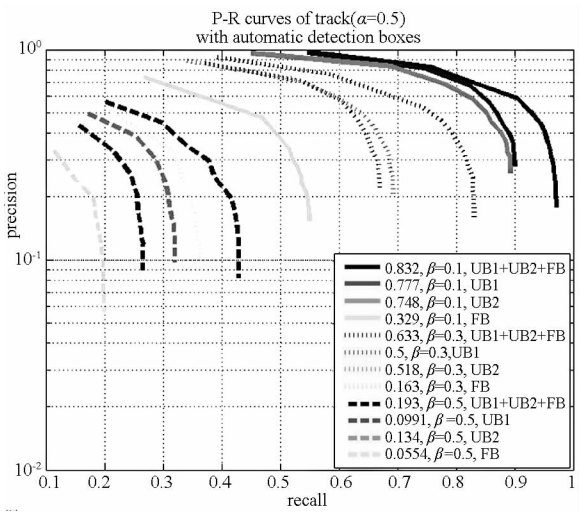


图 3 不同的检测算子的影响

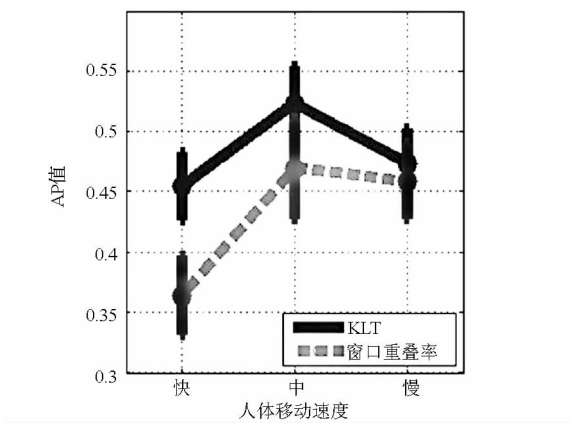
Fig. 3 The influence of using different detectors

3.3 KLT 跟踪的影响

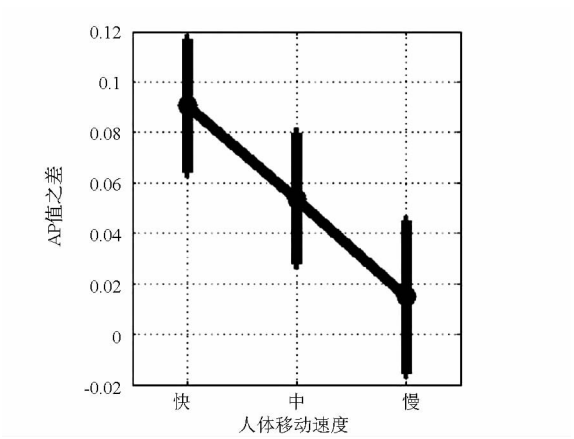
本文利用 KLT 点跟踪算法来计算窗口之间的链接度。为了测试 KLT 跟踪对性能的影响，我们将其结果与窗口面积重叠比例法进行对比。设定 $\alpha = 0.5$、$\beta = 0.3$，检测算子为 UB1 + UB2 + FB。在整个数据库上，KLT 跟踪的 AP 值为 0.633，而窗口重叠率的结果为 0.585。可见，KLT 跟踪的性能优于窗口重叠率方法。

为了进一步分析二者的性能，我们将以下实验过程重复 50 次：首先，从 300 幅视频中随机提取 100 幅；然后，将这 100 幅视频中所有的人体轨迹按照运动速度（利用相邻窗口重叠率的均值进行度量）平均分成快、中、慢 3 部分；最后，计算 3 种运动速度、2 种连接度计算方法的 AP 结果。图 4 展示了这 6 种情况的 AP 值结果。如图 4(a)所示，这 6 种情况的 AP 值结果都相对较低(0.37 ~ 0.53)；这是因为对于真实轨迹只使用了约 1/3，而对于自动跟踪的轨迹则使用了整个集合。从图 4(b)中可以观察到，随着人体运动速度的增加，KLT 跟踪的优势越来越明显，说明了 KLT 跟踪方法对于快速运动目标的适应能力。

图 5 中显示了一段视频中的跟踪结果。图中显示了 4 条最显著的人体轨迹，如白色序号所示。左下角的数字表示图像所在的帧。其中，轨迹 1 和 2 可以看作正确的跟踪结果。在轨迹 2 中，第 21 帧内没有检测窗口，而在前后帧中都有检测窗口，这在一定程度上说明了我们的跟踪方法能够适应暂时的检测失败。



(a) 对于快、中、慢 3 种轨迹的跟踪性能



(b) KLT 跟踪与窗口重叠率性能之差

图 4 KLT 跟踪与窗口重叠率的性能比较

Fig. 4 Comparing the performance of KLT tracker and bounding boxes overlap

3.4 计算速度

为了提高检测速度，我们进行了时间下采样，可能会对跟踪性能有所影响。然而，经过试验发现：即使采用了 5 倍的下采样，最终的跟踪性能也不会明显下降(AP 值下降不超过 0.01)。

文献[10]的报道以及我们的实验都表明了使用单个检测算子的速度约为 6 帧(640 × 480 像素)/秒。本文一共使用了 3 个检测算子，同时进行了 5 倍的时间下采样，因此基本上能够满足实时性要求。另外，如果使用多核计算机的话，3 个检测算子可以并行计算，这样速度会更快。

其次，我们使用 KLT[9] 跟踪方法来计算窗口间的链接度。如果每帧提取 800 个跟踪点的话，计算速度约为 10 帧/秒。在现实应用中，如果目标运动不是太快的话，可以利用窗口之间面积重叠比例来直接计算链接度，所耗的时间几乎可以忽略，但却不能适应快速运动的目标。

3.5 与以前方法的比较

以前一些工作[3-6]也采用了基于检测的人体

跟踪技术。这些方法都只使用了一个检测算子。对于这些方法,我们都一致采用 UB1,并设定 $\alpha = 0.5$ 和 $\beta = 0.3$。对于文献[3],我们采用本文所提的窗口链接算法;对于文献[4-5],采用窗口重叠率和凝聚聚类;而对于文献[6]采用窗口重叠率和该文所提的基于网络流的全局最优算法来计算人体轨迹。

表 1 中列出了这些方法取得的 AP 值结果。由表可见,基于全局最优的文献[6]的结果甚至比基于局部优化的文献[4-5]的结果还要差,这可能是因为该数据库中存在大的突然的视角变化等状况。与文献[3]、文献[4-5]和文献[6]的方法相比,我们方法的跟踪性能分别高出 0.133、0.174 和 0.201,说明了本文所提出的算法的有效性。

表 1　与以前方法的 AP 值比较

Tab. 1　Comparing with the previous methods

	文献[3]	文献[4-5]	文献[6]	本文
AP 值	0.500	0.459	0.432	0.633

图 5　跟踪结果

Fig. 5　Tracking results

4　结论

本文提出了一种新的基于检测的全自动人体跟踪算法。该算法包含两个步骤:首先在视频内进行人体检测和定位;然后对各帧内的检测窗口进行连接,形成多条人体轨迹。为了提高跟踪性能,我们使用了多个检测算子,用来检测不同的人体部分,例如整个人体和上半身;因此能够适应各种不同的人体姿势和遮挡情况。为了适应快速运动的目标,我们采用 KLT 跟踪和凝聚聚类技术将窗口链接起来。实验结果表明:使用多个检测算子能够显著地提高跟踪性能。另外,该算法基本能够达到实时性要求。

参考文献(References)

[1] Jia X, Lu H C, Yang M H. Visual tracking via adaptive structural local sparse appearance model[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012:1822-1829.

[2] 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.

[3] Patron-Perez A, Marszalek M, Reid I, et al. Structured learning of human interactions in TV shows[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(12): 2441-2453.

[4] Jiang Y G, Li Z G, Chang S F. Modeling scene and object contexts for human action retrieval with few examples[J]. IEEE Transactions on Circuits System Video Technology, 2011, 21(5): 674-681.

[5] Ikizler-Cinbis N, Sclaroff S. Object, scene and actions: combining multiple features for human action recognition[C]//Proceedings of the 11th European Conference on Computer Vision, 2010:494-507.

[6] Pirsiavash H, Ramanan D, Fowlkes C C. Globally-optimal greedy algorithms for tracking a variable number of objects[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011:1201-1208.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, 1:886-893.

[8] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[9] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]//Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981, 2:674-679.

[10] Dollar P, Belongie S, Perona P. The fastest pedestrian detector in the west[C]//Proceedings of the British Machine Vision Conference, 2010:1-11.

[11] Benenson R, Mathias M, Timofte R, et al. Pedestrian detection at 100 frames per second[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012:2903-2910.

[12] Patron-Perez A, Marszalek M, Zisserman A, et al. High five: Recognizing human interactions in tv show[C]//Proceedings of the British Machine Vision Conference, Aberystwyth, 2010.

[13] Everingham M, Gool L, Williams C K, et al. The Pascal visual object classes challenge 2007 (VOC) results[J]. International Journal of Computer Vision, 2010, 88(2):303-338.

[14] Prest A, Schmid C, Ferrari V. Weakly supervised learning of interactions between humans and objects[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 601-614.

[15] Eichner M, Ferrari V. Better appearance models for pictorial structures[C]//Proceedings of the British Machine Vision Conference, Rama Chellappa, 2009.