Team Phoenix in

Pedestrian detection

So, what skills will dominate
Team Phoenix recruitment in 2019?

https://github.com/MahmoudElbaz74/Pedestrian-detection/blob/main/README.md

Data set

https://www.kaggle.com/datasets/karthika95/pedestrian-detection

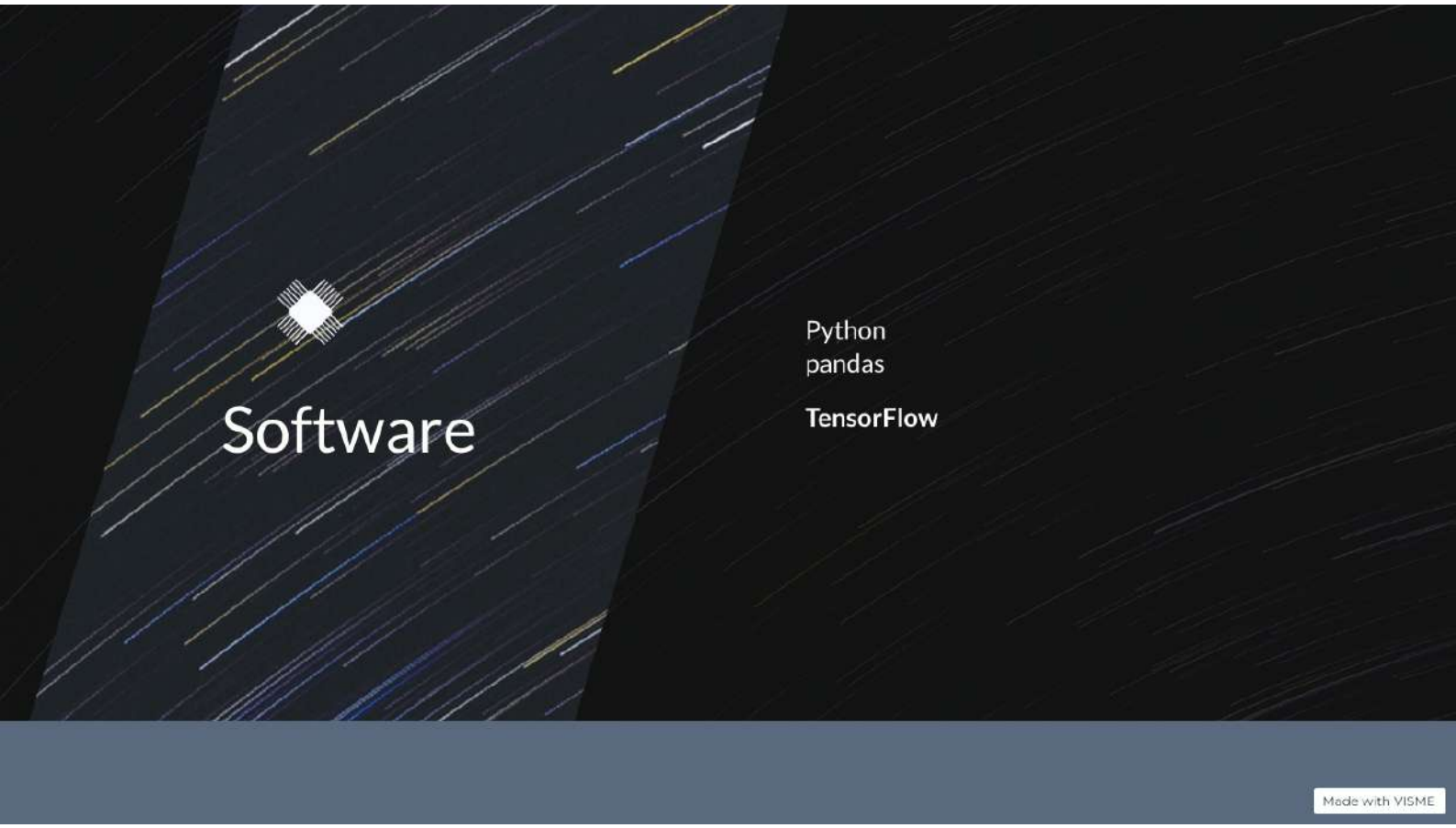
INTRODUCTION

Pedestrian detection is an essential and significant task in any intelligent video surveillance system, as it provides the fundamental information for semantic understanding of the video footages. It has an obvious extension to automotive applications due to the potential for improving safety systems. Many car manufacturers (Volvo, Ford, GM, Nissan) offer this as an ADAS option in 2017

Ways to detect pedestrians

Holistic detection
Part-based detection
Patch-based detection
Motion-based detection
Detection using multiple cameras

https://en.wikipedia.org/wiki/Pedestrian_detection



Detection Driven Adaptive Multi-cue Integration for Multiple Human Tracking

Ming Yang, Fengjun Lv, Wei Xu, Yihong Gong
NEC Laboratories America, Inc.
10080 North Wolfe Road, SW-350, Cupertino, CA 95014
{myang,flv,xw,ygong}@sv.nec-labs.com

Abstract

In video surveillance scenarios, appearances of both human and their nearby scenes may experience large variations due to scale and view angle changes, partial occlusions, or interactions of a crowd. These challenges may weaken the effectiveness of a dedicated target observation model even based on multiple cues, which demands for an agile framework to adjust target observation models dynamically to maintain their discriminative power. Towards this end, we propose a new adaptive way to integrate multi-cue in tracking multiple human driven by human detections. Given a human detection can be reliably associated with an existing trajectory, we adapt the way how to combine specifically devised models based on different cues in this tracker so as to enhance the discriminative power of the integrated observation model in its local neighborhood. This is achieved by solving a regression problem efficiently. Specifically, we employ 3 observation models for a single person tracker based on color models of part of torso regions, an elliptical head model, and bags of local features, respectively. Extensive experiments on 3 challenging surveillance datasets demonstrate long-term reliable tracking performance of this method.

1. Introduction

Tracking multiple human is critical to many applications, ranging from video-based surveillance to human behavior analysis. Reliable human trackers have been intensively studied for several decades with significant progresses [11, 6, 23, 12, 28, 22, 25]. Nevertheless, it is still not uncommon for trackers to be challenged by enormous variations of targets and scenes, *e.g.* cluttered backgrounds, scale and view angle changes, unpredictable occlusions, and complicated interactions among multiple human. These difficulties stem from the fundamental challenge: how to design and maintain observation models of targets that are robust to numerous variabilities and capable of distinguishing themselves from their nearby background constantly.

In general, an observation model of targets based on a single cue may be robust to certain distractions but vulnerable to some others, *e.g.* color-based cue is robust to object deformations but sensitive to lighting changes, while, shape-based cue is insensitive to lighting changes but could be distracted by cluttered background. Therefore, it is appealing to fuse multiple cues into one observation model. For the sake of simplicity, most existing approaches assume different cues are conditionally independent or the dependence is fixed all the time. However, in reality discriminative capabilities and dependence of different cues are unknown and may change dynamically. Therefore, to maintain discriminative observation models for targets with dynamic appearances, it is desirable to adapt the way to integrate multiple cues on-the-fly during tracking.

Online adaptation of observation models without any supervision is risky, since adaptation errors may be accumulated gradually and lead to tracking drift. Consequently, for long-term robust tracking, certain supervision is indispensable to initialize a tracker, guide the adaptation, and help the tracker recover from tracking failures. Object detectors are ideal means to provide such supervision for a fully automatic tracking system, which has been an active research topic for decades itself. It is extremely hard to design a perfect object detector with both high detection rate and precision rate. Nonetheless, it is feasible to obtain a detector with high precision only to provide limited supervision. Therefore, we incorporate such a human detector with high precision and acceptable detection rate into a tracking system to guide the adaptation of multi-cue integration for individual human trackers.

In real-world sequences, appearances of both targets and their nearby scenes are dynamic in general. In view of these facts, we propose tracking multiple human where multi-cues are adaptively integrated driven by human detections. For a single target tracker, multiple cues based on color models, shape matching, and bags of local features are combined to infer the MAP estimation as tracking results in the Bayesian filtering framework. When a human detection can be associated with a trajectory reliably, we regard it as the

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

true target location and adapt the combination of different cues to enhance the discriminative power of the integrated observation model for this target. By formulating this adaption as a regression problem, we analytically and efficiently solve the optimal combination of multiple cues in terms of the integrated model's discriminative capability against its local vicinity.

The proposed method effectively unites the strengths of detection and tracking. The observation model of each cue which encodes the domain knowledge and is specifically designed for the target remains unchanged during tracking, so as to largely alleviate the risk of model drift. Instead, the integration of multiple cues is adapted on-the-fly which is supervised by reliable detections. This enables handling targets with dynamic appearances in non-stationary cluttered background. Thus, off-line designed target models gradually evolve to customized models for different targets online. We incorporate this adaptive cue-integration algorithm into a multiple human tracking system, which employs a human head detector based on a Convolutional Neural Network (CNN) [16], and 3 cues based on color models of part of torso regions, an elliptical head model, and bags of local features, respectively. This fully automatic system has been evaluated extensively on the CAVIAR dateset [7] and 24 hours of real surveillance videos in retail and airport scenarios, and demonstrates prominent long-term tracking performance in these challenging unconstrained environments.

2. Related Work

Literature review about visual tracking is beyond the scope of this paper. As the proposed method mixes the ideas of multi-cue integration, online observation model adaptation, and detection driven tracking, we mainly discuss within these contexts in visual tracking.

For multi-cue integration, the simplest case is that different cues are assumed to be independent, which can be fused optimally using the best linear unbiased estimator (BLUE). For example, [4] assumed two complementary cues are independent with equal variance, so that the matching of intensity gradients around the objects boundary and the color histogram of the objects interior were combined with equal weights in a header tracker. Considering the cue dependence, [26] formulated cue integration as a co-inference problem where multiple modalities interact and guide the updates of each other. [20] represented each cue as a different Bayesian filter and assumed sequentially conditional dependent among them, thus, the cue dependence is considered in the re-sampling stage in the particle filtering [11].

Online learning of target observation model or dynamic feature selection [13, 5, 18, 2, 27, 9] are effective and popular approaches to coping with targets with dynamic appearances. Typically, the tracking result at current frame is used to update the observation model directly or to collect training samples for online learning. However, in either case, such unsupervised adaptation is prone to clutters and partial occlusions. In addition, model errors may be accumulated gradually. Therefore, model drift is not rare in practice. In contrast, our approach differs from conventional online model adaptation in that the adaptation is not performed blindly but driven by detection results. Moreover, the way to integrate multiple cues is adapted rather the observation model of individual cues. Since generally these models of different cues are specifically devised for a target and they remain unchanged during tracking, the risk of model drift is alleviated. On the other hand, the combination of these models in the integrated observation model is updated to make the target distinguishable from its neighborhood. Thus, we adopt online regression in the adaptation and need not explicitly determine positive and negative training samples as in previous online learning [2, 27, 9].

There were a few attempts to combine the strengths of detection and tracking recently. [17] optimized the space-time trajectories of pedestrian detections, then fed back these trajectories to guide removing false positives from detections. [1] incorporated pedestrian detectors and human pose estimator to obtain short tracks, then linked them by the Viterbi algorithm. [10] proposed to associate human detections by a hierarchical of matching schemes utilizing dynamics and scene knowledge. The fundamental difference from our method is that these work all assumed detectors with high recall rates can provide sufficient detection results though with some false positives. In contrast, we assume detectors with high precision merely output reliable detection results occasionally. Additionally, the observation models to establish the associations among detections in consecutive frames are fixed in these methods.

3. Overview of our approach

The key idea of our approach is to utilize object detectors to provide supervision to the adaption of multi-cue integration for single object trackers. Thus, a generic object tracker can gradually evolve to a specific object tracker in order to be distinguishable in its neighborhood. Specifically, for each input frame, we employ the Bayesian filtering framework to infer the tracking results of single target trackers which combine multiple cues in their observation models. In the meantime, we run a human detector with high precision and acceptable recall rate. If a detection can be associated with a tracked trajectory reliably, we utilize this detection to adapt the way of multi-cue integration for this tracker. The updated observation models are used to track targets and associate with detections in the following frames. A detection that is not associated with an existing trajectory is used to initialize a new tracker. The system block diagram is summarized in Fig. 1.

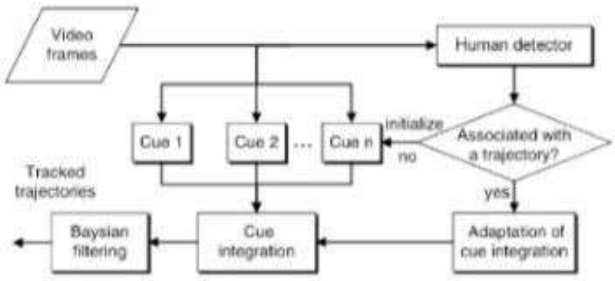


Figure 1. The system block diagram.

3.1. Single target tracking

We formulate single target tracking in the Bayesian filtering framework. Denote the motion parameters of the target by $\mathbf{x} = \{u, v, s\}$ where (u, v) is the translation and s is the scale, and the corresponding image observation by $\mathbf{z}$. The posterior is recursively estimated based on the likelihood or observation model $P(\mathbf{z}_t|\mathbf{x}_t)$ and the dynamic model $P(\mathbf{x}_t|\mathbf{x}_{t-1})$, as

$$P(\mathbf{x}_t|\mathbf{z}_t) \propto P(\mathbf{z}_t|\mathbf{x}_t) \int P(\mathbf{x}_t|\mathbf{x}_{t-1})P(\mathbf{x}_{t-1}|\mathbf{z}_{t-1})d\mathbf{x}_{t-1}. \tag{1}$$

The tracking result is the MAP estimation $\mathbf{x}_t^* = \arg\max_{\mathbf{x}_t} P(\mathbf{x}_t|\mathbf{z}_t)$.

3.2. Association of detections with trajectories

A detection is assumed to associate with a trajectory reliably if they are consistent in terms of both appearance matching and motion dynamics. Given the detection responses $\mathbf{y}^i$ (the location and scale of an object) and the tracking results $\mathbf{x}_t^j$ at t, we associate them by solving an optimal assignment problem. For each pair of $\mathbf{y}^i$ and $\mathbf{x}_t^j$, their association likelihood $P(\mathbf{y}^i, \mathbf{x}_t^j)$ is defined by plugging $\mathbf{y}^i$ into Eq. 1, *i.e.*,

$$P(\mathbf{y}^i, \mathbf{x}_t^j) = P(\mathbf{y}^i|\mathbf{x}_t^j)P(\mathbf{y}^i|\mathbf{x}_{t-1}^j). \tag{2}$$

Note here a detection result $\mathbf{y}^i$ is regarded as an observation. We use a Gaussian constant velocity model for the dynamic model $P(\mathbf{y}^i|\mathbf{x}_{t-1}^j)$, where the velocity and its variance are calculated using the history trajectory. The observation model based on multiple cues is discussed in details in Sec. 4 and Sec. 5. Thus we construct an assignment matrix $\mathbf{C}$ where each element $C_{ij} = \log P(\mathbf{y}^i, \mathbf{x}_t^j)$. The optimal maximizing assignments are computed using the well-known Hungarian algorithm [14]. If the matching of an assignment $(\mathbf{y}^i, \mathbf{x}_t^j)$ is too low or $\mathbf{y}^i$ can not find a match at all, $\mathbf{y}^i$ will be initialized as the start of a new trajectory. Otherwise, we substitute $\mathbf{x}_t^j$ by $\mathbf{y}^i$ as the tracking result $\mathbf{x}_t^{*j}$ for this tracker to guide the adaptation of cue integration.

4. Multi-cue Integration and Adaptation

For a single target tracker, given the associated detection result $\mathbf{x}_t^* = \mathbf{y}^i$, we adapt the integration of multiple cues to enhance its discriminability with respect to its close neighborhood. Denote the observations of N different cues at time t by $\mathbf{z}_t = \{\mathbf{z}_t^1, \cdots \mathbf{z}_t^N\}$. The key of multi-cue integration is how to model the joint likelihood $P(\mathbf{z}_t^1, \cdots \mathbf{z}_t^N|\mathbf{x}_t)$. If multiple cues are assumed conditionally independent, then $P(\mathbf{z}_t^1, \cdots \mathbf{z}_t^N|\mathbf{x}_t) = \prod_{n=0}^N P(\mathbf{z}_t^n|\mathbf{x}_t)$. Without confusion, we drop the subscript t in this section.

The joint likelihood can be modeled using a joint dissimilarity function $d(\cdot)$,

$$P(\mathbf{z}^1, \cdots \mathbf{z}^N|\mathbf{x}) = \frac{1}{Z}\exp(-d(\mathbf{z}^1(\mathbf{x}), \cdots, \mathbf{z}^N(\mathbf{x})), \tag{3}$$

where $\mathbf{z}^n(\mathbf{x})$ is the image observation of the nth cue given the motion parameter $\mathbf{x}$, and Z is a normalization term. Not using the assumption of conditional independence, we model the joint dissimilarity function as a linear combination of the dissimilarity functions of individual cues:

$$d(\mathbf{z}^1(\mathbf{x}), \cdots, \mathbf{z}^N(\mathbf{x})) = \sum_{n=1}^N w_n d(\mathbf{z}^n(\mathbf{x})) = \mathbf{w}^T\mathbf{d}(\mathbf{x}), \tag{4}$$

where $\mathbf{w} = \{w_1, \cdots, w_N\}$ are non-negative weights and $\mathbf{d}(\mathbf{x}) = \{d(\mathbf{z}^1(\mathbf{x})), \cdots, d(\mathbf{z}^N(\mathbf{x}))\}$ concatenates the dissimilarity function of each cue. These dissimilarity functions should give 0 for perfect matching, and their ranges should be consistent. For example, in our implementation, for the color-based cue, the dissimilarity is one minus the Bhattacharyya coefficient *w.r.t* the stored color model; for the shape-based cue, it is one minus the sum of matching score along the head shape model. Their values range in $[0, 1]$. w_n is initialized to 1 when a new tracker starts.

Denote $d(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{d}(\mathbf{x})$. Given $\mathbf{x}^*$, we strive to enhance the discriminative power of this integrated dissimilarity function $d(\mathbf{x}; \mathbf{w})$ in its close neighborhood $N(\mathbf{x}^*)$. Namely, the farer one motion parameter $\mathbf{x}$ away from $\mathbf{x}^*$, the larger of the difference between $d(\mathbf{x}; \mathbf{w})$ and $d(\mathbf{x}^*; \mathbf{w})$, which indicates a more discriminative joint dissimilarity function spatially. To explicitly model this property of $d(\mathbf{x}; \mathbf{w})$, we introduce a monotonic function $f(\mathbf{x}; \mathbf{x}^*)$ *w.r.t* the distance of $\mathbf{x}$ to $\mathbf{x}^*$ to represent the discriminative capability. Thus, the adaptation of cue integration can be well formulated as a regression problem:

$$d(\mathbf{x}_m; \mathbf{w}) - d(\mathbf{x}^*; \mathbf{w}) = f(\mathbf{x}_m; \mathbf{x}^*) - \xi_m, \forall \mathbf{x}_m \in N(\mathbf{x}^*), \tag{5}$$

where ξ_m are slack variables. Then, if $d(\mathbf{x}_m; \mathbf{w})$ satisfies this equation with all $\xi_m \leq 0, \forall \mathbf{x}_m \in N(\mathbf{x}^*)$, this indicates that $d(\mathbf{x}; \mathbf{w})$ is more discriminative than $f(\mathbf{x}; \mathbf{x}^*)$. Sample functions of f based on the normalized distance $||(u, v) - (u^*, v^*)||/s^*$ are shown in Fig. 2.

Team Phoenix

Teammate

- Mahmoud Elbaz - |
- Mahmoud Mohamed - ||
- Mahmoud Marry - |||
- Marwan Amr - |V
- Medhat Wahed - V

Teammate

- | - Bliud the model
- || -
- ||| -
- |V -
- V -