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.

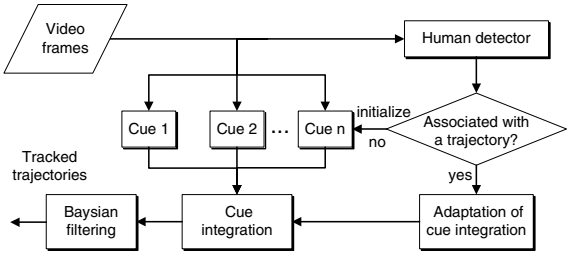

Figure 1. The system block diagram.

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.

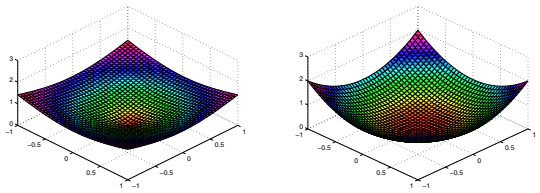

Figure 2. Sample $f(\mathbf{x};\mathbf{x}^*) = (||(u,v)-(u^*,v^*)||/s^*)^p$. $p=1$ (left); $p=2$ (right).

To find the optimal $\mathbf{w} = \{w_1, \cdots, w_n\}$ that maximizes the discriminative power of $d(\mathbf{x}_m;\mathbf{w})$, we need to minimize $\sum_{m=0}^{M} c(\xi_m)$, where M is the number of motion parameters in the neighborhood $N(\mathbf{x}^*)$, and $c(\xi_m)$ is a cost function to penalize $\xi_m > 0$. In the meantime, we need to minimize a regularizer $||\mathbf{w}||^2$ to favor a *flat* function $d(\mathbf{x};\mathbf{w})$ (Please refer to the support vector regression (SVR) [24] for the discussion of *flatness*). Thus, this regression problem can be solved by an optimization problem with a regularization term λ, given as

$$\begin{aligned}\min L(\mathbf{w}) &= \frac{1}{M}\sum_{m=0}^{M} c(\xi_m) + \frac{\lambda}{2}||\mathbf{w}||^2 \\ &= \frac{1}{M}\sum_{m=0}^{M} c(f(\mathbf{x}_m;\mathbf{x}^*) - \mathbf{w}^T(\mathbf{d}(\mathbf{x}_m) - \mathbf{d}(\mathbf{x}^*))) \\ &\quad + \frac{\lambda}{2}||\mathbf{w}||^2. \end{aligned} \tag{6}$$

This optimization problem can be solved either by the standard primal-dual approach employed in SVR [24], or by direct gradient descent in the primal space. We use the gradient descent method since the dimensionality of $\mathbf{w}$ is low. The gradient $\nabla L(\mathbf{w})$ and the update rules for $\mathbf{w}$ given a learning rate Λ are derived by

$$\nabla L(\mathbf{w}) = \frac{1}{M}\sum_{m=0}^{M} -c'(\xi_m)(\mathbf{d}(\mathbf{x}_m) - \mathbf{d}(\mathbf{x}^*)) + \lambda\mathbf{w}. \tag{7}$$

$$\mathbf{w} \leftarrow \mathbf{w} - \Lambda\nabla L(\mathbf{w}),$$

$$\mathbf{w} \leftarrow (1-\Lambda\lambda)\mathbf{w} + \frac{\Lambda}{M}\sum_{m=0}^{M} c'(\xi_m)(\mathbf{d}(\mathbf{x}_m) - \mathbf{d}(\mathbf{x}^*)) \tag{8}$$

The cost function $c(\xi)$ should only penalize $\xi > 0$, we choose $c(\xi) = |\xi|_+ = \max(0,\xi)$, so $c'(\xi) = 1$ if $\xi > 0$ and $c'(\xi) = 0$ if $\xi \leq 0$. Arbitrary function f can be used. In our implementation we use $f(\mathbf{x};\mathbf{x}^*) = ||(u,v)-(u^*,v^*)||/s^*$. λ is set equal to 1. The initial Λ is set to 1 and decreases by a half in each iteration with up to 5 steps in Eq. 8. Note we do not enforce the sum of w_n equals to one, instead, the function $f(\mathbf{x};\mathbf{x}^*)$ implicitly enforces the similar constraint.

We select the neighborhood $N(\mathbf{x}^*)$ from those hypotheses whose dissimilarity functions $d(\mathbf{z}^n(\mathbf{x}))$ have been calculated in tracking. Therefore, this cue-integration adaptation module does not induce much extra computation. The computational complexity is almost the same as integrating multiple cues without adaptation.

5. Observation Models of Individual Cues

The proposed cue-integration method is flexible to accommodate arbitrary cues, so long as the ranges of their dissimilarity functions are consistent. Color, shape and texture are 3 basic characteristics of an object in monocular videos. Therefore, for a single person tracker, we employ 3 observation models based on color models of human head and part of torso region, an elliptical head model, and bags of local features on torso, respectively.

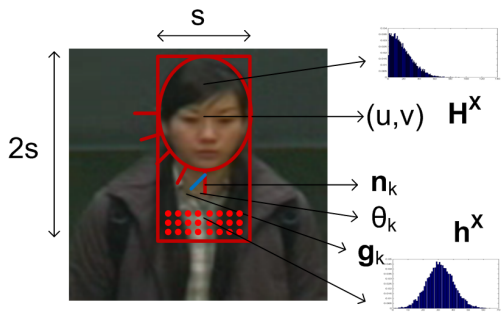

Figure 3. Illustration of the observation models of 3 different cues in the single person tracker.

5.1. Cue 1 based on color models

The kernel-weighted color histogram [6] of the head and upper torso region, drawn as the red rectangle in Fig. 3, is stored for each target. The histogram consists of $8\times8\times8 = 512$ bins in the RGB space. The Bhattacharyya coefficient (*BC*) is used to measure the similarity between a color histogram of a hypothetic target $\mathbf{x}$ and the stored model, *i.e.*

$$BC(\mathbf{H^x},\mathbf{H^m}) = \sum_{q}^{512}\sqrt{H_q^{\mathbf{x}} H_q^{\mathbf{m}}}, \tag{9}$$

where $\mathbf{H^x}$ and $\mathbf{H^m}$ are the color histograms of $\mathbf{x}$ and the stored model respectively, and q is the bin index. $1 - BC(\mathbf{H^x},\mathbf{H^m}) = 0$ indicates a perfect match.

To make this model more robust, for each trajectory, two color histograms are stored. $\widetilde{\mathbf{H}}^{\mathbf{m}}$ is the histogram of the last tracked instance on the trajectory and $\overline{\mathbf{H}}^{\mathbf{m}}$ is the running average of the latest a few instances. The first one accounts for the short-term memory of the target. When a person has been correctly tracked in consecutive frames, $1 - BC(\mathbf{H^x},\widetilde{\mathbf{H}}^{\mathbf{m}})$ shall give a very small value. In case

of losing tracking or occlusions, long-term memory modeled by $\overline{\mathbf{H}}^{\mathbf{m}}$ is helpful so that when the person re-appears, he/she can still be remembered and tracking can be correctly recovered. Therefore, the dissimilarity function of this cue is written as

$$d(\mathbf{z}^1(\mathbf{x})) = \min(1 - BC(\mathbf{H}^{\mathbf{x}}, \widetilde{\mathbf{H}}^{\mathbf{m}}), 1 - BC(\mathbf{H}^{\mathbf{x}}, \overline{\mathbf{H}}^{\mathbf{m}})). \tag{10}$$

5.2. Cue 2 based on an elliptical head model

The elliptical shape of a head is an evident indication of the presence of a human [4]. Given a hypothesis $\mathbf{x}$, we calculate the intensity gradients on $K = 36$ normal vectors $\vec{\mathbf{n}}_k$ along the ellipse centered at $\mathbf{x}$, as drawn as the red ellipse in Fig. 3. The angle θ_k between the largest gradient $\vec{\mathbf{g}}_k$ on $\vec{\mathbf{n}}_k$ and $\vec{\mathbf{n}}_k$ measures the discrepancy. The larger the sum of all cosine of θ_k indicates the better matching with this elliptical model. The dissimilarity function is defined by

$$d(\mathbf{z}^2(\mathbf{x})) = 1 - \frac{1}{K}\sum_{k=1}^{K} |\cos(\theta_k)|. \tag{11}$$

$d(\mathbf{z}^2(\mathbf{x})) = 0$ indicates a perfect match.

5.3. Cue 3 based on bags of local features

We employ a bag of local features to capture texture characteristics of human torso regions. We compute fast dense SIFT-like features [19, 8] on every grid (with size of 4×4 pixels) within the red rectangular region in Fig. 3. By assigning each feature vector into one of 256 clusters, and then counting the frequency of assignments for every cluster, we obtain a histogram of these local features. Inspired by recent success of the spatial pyramid matching (SPM) method [15] in object classification, we choose SPM to incorporate the spatial layout information of local features. Two levels with 1×1 and 2×2 cells are used. So, for each hypothesis region the final feature representation has $256 \times 5 = 1280$ dimensions.

Again, the Bhattacharyya coefficient is used to to measure the difference between two SPM histograms, *i.e.*

$$d(\mathbf{z}^3(\mathbf{x})) = 1 - BC(\mathbf{h}^{\mathbf{x}}, \mathbf{h}^{\mathbf{m}}), \tag{12}$$

where $\mathbf{h}^{\mathbf{x}}$ and $\mathbf{h}^{\mathbf{m}}$ are the SPM histograms of $\mathbf{x}$ and the stored model respectively.

6. Experimental results

6.1. Settings

We test the proposed tracking algorithm on a variety of real-world surveillance video datasets including 26 sequences in the CAVIAR dataset [7], 4 hours videos in a retail scenario, and 20 hours videos in an airport scenario used in the TRECVID 2008 event detection [21]. The proposed approach is compared with methods using the individual color-based, shape-based and texture-based cues in Sec. 5, and the method fusing them with the assumption of conditional independence. Up to 300 samples and 3 scale factors $\{0.95, 1, 1.05\}$ are used in a single target tracker. The computational complexity of the algorithm depends on the number of targets and image resolution. Given the detection results, the system runs at about 1-5 fps on a Core2Duo 3.16GHz desktop.

We focus on evaluating average tracking performance over a long time period rather than performance for a few tough cases. Evaluation of multiple human tracking requests annotating every target, which is quite costly even for not so crowded scenes. Given these annotations with correspondences, we measure the tracking performance by the multiple object tracking accuracy (MOTA) criterion [3], which penalizes missed targets, false positives, and identity mismatch errors, *i.e.*,

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}, \tag{13}$$

where m_t, fp_t, and mme_t are the numbers of misses, of false positives, and of mismatch errors at time t, respectively, and g_t is the number of true targets. If the correspondences among annotations are not available, we measure the performance by the F1 score, *i.e*, the harmonic mean of precision and recall rates.

6.2. Human detection

Partial occlusions of human body occur frequently in crowded scenes, therefore, we train a multi-view human head detector based on a Convolutional Neural Network (CNN) [16]. The performance is tuned to have fairly high precision (around 80%) with acceptable recall rates (around $40\% - 50\%$). The detection performance (denoted by *Det.*) on each dataset is reported in the first row of Tabs.1-3.

6.3. The CAVIAR dataset

The CAVIAR dataset [7] includes 26 videos at resolution 384×288 captured in a corridor, and its ground truth contains 235 trajectories. We estimate head locations from the ground truth to compare with the detection and tracking results. The proposed method (denoted by *MC-Adapt*) is compared with the method that integrates multiple cues assuming conditional independence (denoted by *MC-Ind*), and the methods using a single cue based on color, shape, and texture information in our detection driven framework (which are denoted by *Cue 1 (C)*, *Cue 2 (S)*, and *Cue 3 (T)*, respectively). The average precision rate (*Prec.*), recall rate (*Rec.*), F1 score, and the MOTA score over 26 sequences of these methods are listed in Tab. 1.

From Tab. 1, we can see color information is a good cue for this dataset since people mainly walk towards or away

Method	Prec.	Rec.	F1	MOTA
Det.	0.873	0.457	0.599	-
Cue 1 (C)	0.788	0.699	0.738	0.511
Cue 2 (S)	0.857	0.579	0.685	0.483
Cue 3 (T)	0.863	0.513	0.644	0.434
MC-Ind	0.804	0.706	0.749	0.535
MC-Adapt	**0.819**	**0.754**	**0.783**	**0.593**

Table 1. Performance comparison on the CAVIAR dataset.

Method	Prec.	Rec.	F1	MOTA
Det.	0.909	0.468	0.618	-
Cue 1 (C)	0.827	0.515	0.634	0.437
Cue 2 (S)	0.812	0.532	0.642	0.442
Cue 3 (T)	0.889	0.484	0.627	0.425
MC-Ind	0.836	0.581	0.678	0.486
MC-Adapt	**0.855**	**0.621**	**0.719**	**0.525**

Table 2. Performance comparison for the retail scenario.

Method	Prec.	Rec.	F1
Det.	0.867	0.456	0.598
MC-Ind	0.770	0.610	0.681
MC-Adapt	**0.831**	**0.632**	**0.718**

Table 3. Performance comparison for the airport scenario.

from the camera and seldom turn around. Nevertheless, the other two cues can help when partial occlusions occur. The *MC-Ind* method obtains the MOTA score 0.535. By adaptively integrating multiple cues, our method further improves the tracking accuracy and achieves the MOTA score 0.593. In comparison, recent work [25] and [10] reported MOTA=0.537 and MOTA=0.540. Note [10] showed that employing high-level scene knowledge, *e*.g., the locations of exits and pillars, and iteratively associating short trajectories off-line can further improve the performance. While, in our method, we do not use such prior knowledge and make decision of tracking results online.

The detection and tracking results are illustrated in Fig. 4. The human detection results are drawn as yellow rectangles in the first row, where the pseudo-color dots indicate the confidence of the detections. Our tracking results are displayed in the second row, where the colors indicate different identities. At frame 277, when the man in the left of the frame largely occludes the coming woman (pointed by a yellow arrow in Fig. 4), the integrated joint dissimilarity function $d(\mathbf{x}; \mathbf{w})$ in the *MC-Adapt* method is more discriminative than that of combining them with equal weighs in the *MC-Ind* method, as shown in Fig. 5. The trackers follow all the targets quite well though detections are absent frequently, *e.g* at frame 404.

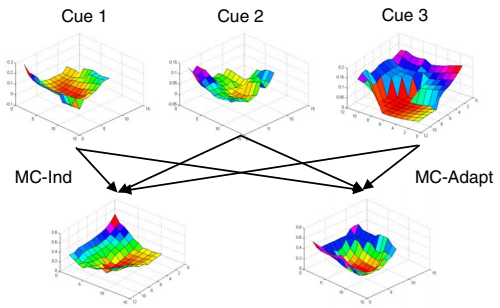

Figure 5. The surfaces of dissimilarity functions in a 11×11 neighborhood for each cue, and the MC-Ind and MC-Adapt methods.

6.4. Results in a retail scenario

We recorded 4 hours of surveillance videos at resolution 640×480 in a retail shop and annotated all human heads every 3 frames. This scenario is challenging since the background is very cluttered and people in suits or white shirts exhibit almost identical appearances, as shown in Fig. 6. The color-based cue is not as effective as in the CAVIAR dataset. By adaptively combining the 3 cues, the *MC-Adapt* method achieves MOTA score 0.525 in this retail scenario.

6.5. Results in an airport scenario

We further evaluate the proposed method in an extremely crowded scenario recorded in London Gatwick International Airport, which is used in the TRECVID 2008 event detection evaluation [21]. We annotated all human heads every 750 frames for 20 hours of 720×576 videos from 2 camera views. Since the exact correspondences for the labelled persons are hard to obtain (someone leaves the scene in tens of seconds), we use the F1 score to measure the performance. Although on average there are 6.97 persons per frame, there may be around 20 people in the scene, as shown in Fig. 7. The persons may undergo large scale and view angle changes and very complex partial occlusions. So no single cue works well enough in this scenario. Our system is fairly robust to deal with these tough cases by combining detection and tracking, as demonstrated in Fig. 7 and 8. As shown in Tab. 3, the proposed method improves the recall rate by about 18% over the human detection with only 3.6% loss in precision, thus the F1 score increases by 0.12.

7. Conclusion

In this paper, we propose to adaptively integrate multiple cues driven by detections and present a complete multiple human tracking system. Multiple cues are combined to explicitly enhance the discriminative capability of the joint observation model with respect to its close neighborhood. Thus, a generic object tracker can gradually adapt to its environments. Without much extra computations, this adaptive method effectively improves the performance compared with integrating multiple cues with fixed equal

Figure 4. Frame #185, 277, 319, 404, and 794 of sequence [OneStopMoveEnter1cor]. The human head detections (top row) and our tracking results (bottom row).

Figure 6. Frame #256, 266, 331, 505, and 936 of sequence [shopping mall]. The human head detections (top row) and our tracking results (bottom row).

weights. The proposed approach has been extensively evaluated on over 24 hours of challenging surveillance videos in retail and airport scenarios and achieves prominent performance. Our further work is to take into consideration of the interactions of multiple trajectories in the system.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR'08*, Anchorage, AK, June 24-26, 2008. 2

[2] S. Avidan. Ensemble tracking. In *CVPR'05*, volume 2, pages 494–501, San Diego, CA, June 20-25, 2005. 2

[3] K. Bernardin and R. Stiefelhagen. Evaluating multipe object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008. 5

[4] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR'98*, pages 232–237, Santa Barbara, CA, June 23-25, 1998. 2, 5

[5] R. T. Collins and Y. Liu. On-line selection of discriminative tracking features. In *ICCV'03*, volume 1, pages 346–352, Nice, France, Oct. 13-16, 2003. 2

[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR'00*, volume 2, pages 142–149, Hilton Head Island, SC, June 13-15, 2000. 1, 4

[7] EC funnded CAVIAR project / IST 2001 37540. http://homepages.inf.ed.ac.uk/rbf/caviar/. 2, 5

[8] B. Fulkerson, A. Vedaldi, and S. Soatto. Localizing objects with smart dictionaries. In *ECCV'08*, Marseille, France, Oct. 12-18, 2008. 5

[9] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC'06*, volume 1, pages 47–56, Edinburgh, 4-7, 2006. 2

[10] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV'08*, volume 2, pages 788–801, Marseille, France, Oct. 12-18, 2008. 2, 6

[11] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV'96*, pages 343–356, Cambridge, UK, 1996. 1, 2

[12] M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. In *ICCV'01*, volume 2, pages 34–41, Vancouver, Canada, July 7-14, 2001. 1

[13] A. D. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. In *CVPR'01*, volume 1, pages 415–422, Kauai, Hawaii, Dec. 9-14, 2001. 2

[14] H. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, (2):83 – 97, 1955. 3

[15] S. Lazebnik, C. Achmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural

Figure 7. Frame #246, 317, 1114, 1150, and 1250 of sequence [`Airport CAM1`]. The human head detections (top row) and our tracking results (bottom row).

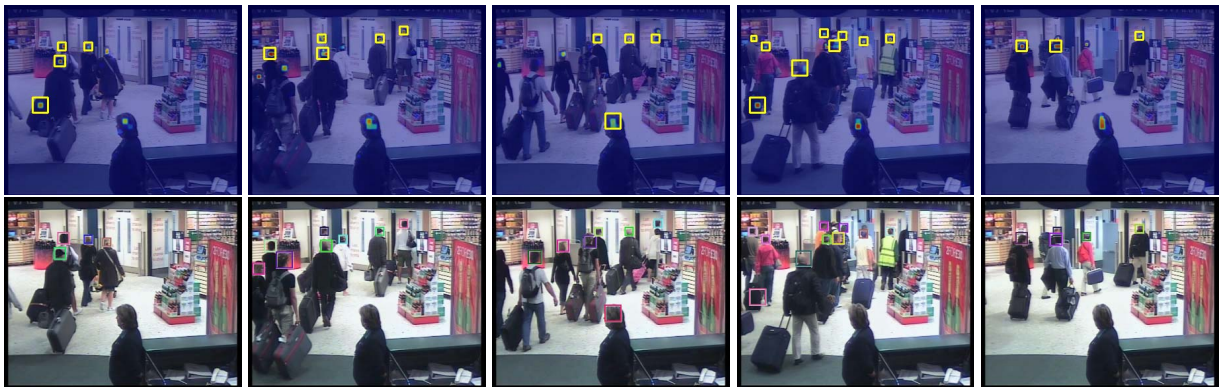

Figure 8. Frame #87, 204, 285, 774, and 980 of sequence [`Airport CAM2`]. The human head detections (top row) and our tracking results (bottom row).

scene categories. In *CVPR'06*, volume 2, pages 2169–2178, New York City, June 17-22, 2006. 5

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based leraning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, Nov. 1998. 2, 5

[17] B. Leibe, K. Schindler, and L. V. Cool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV'07*, Rio de Janeiro, Brazil, Oct. 14-20, 2007. 2

[18] J. Lim, D. Ross, R.-S. Lin, and M.-H. Yang. Incremental learning for visual tracking. In *Advances in Neural Information Processing Systems 17 (NIPS'04)*, pages 801–808, Vancouver, Canada, Dec. 13-18, 2004. 2

[19] D. G. Lowe. Distinctive image features from scale invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004. 5

[20] F. Moreno-Noguer, A. Sanfeliu, and D. Samaras. Dependent multiple cue integration for robust tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(4):670–685, Apr. 2001. 2

[21] National Institute of Standards and Technology (NIST): TRECVID 2008 Evaluation for Surveillance Event Detection. http://www.nist.gov/speech/tests/trecvid/2008/ and http://www.nist.gov/speech/tests/trecvid/2008/doc/eventdet08-evalplan-v07.htm#tasks, 2008. 5, 6

[22] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV'04*, pages 28 – 39, Prague, Czech Republic, May 11-14, 2004. 1

[23] C. Rasmussen and G. D. Hager. Probabilistic data association methods for tracking complex visual objects. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(6):560 – 576, June 2001. 1

[24] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995. 4

[25] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *Int'l Journal of Computer Vision*, 75(2):247–266, Nov. 2007. 1, 6

[26] Y. Wu and T. S. Huang. Robust visual tracking by integrating multiple cues based on co-inference learning. *Int'l Journal of Computer Vision*, 58(1):55–71, June 2004. 2

[27] M. Yang and Y. Wu. Tracking non-stationary appearances and dynamic feature selection. In *CVPR'05*, volume 2, pages 1059–1066, San Diego, CA, June 20-26, 2005. 2

[28] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR'04*, volume 2, pages 1063 – 1069, Washington, DC, Jun.27-Jul.2 2004. 1