This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

Online-Learning-Based Human Tracking Across Non-Overlapping Cameras

Young-Gun Lee, *Student Member, IEEE,* Zheng Tang, *Student Member, IEEE,* Jenq-Neng Hwang, *Fellow, IEEE*

***Abstract*—Due to the expanding scale of camera networks, multiple camera tracking of human has received higher attention in recent years. In this paper, we present a novel approach to track each human within a single camera and across multiple disjoint cameras. Our framework includes a multi-object tracking and segmentation system, a two-phase feature extractor, and an online-learning-based camera link model estimation. For tracking within a single camera, we apply tracking by segmentation and local object detection with multi-kernel feedback to adaptively improve robustness of the algorithm. In inter-camera tracking, we introduce an effective integration of appearance and context features. Automatically couples are detected, and the couple feature is also integrated with existing features. The proposed algorithm is scalable by a fully unsupervised online learning framework. In our experiments, the proposed method outperforms all the state-of-the-art in the benchmark NLPR_MCT dataset.**

***Index Terms*—Adaptive segmentation, multi-object tracking, visual surveillance, multiple camera tracking, NLPR_MCT dataset.**

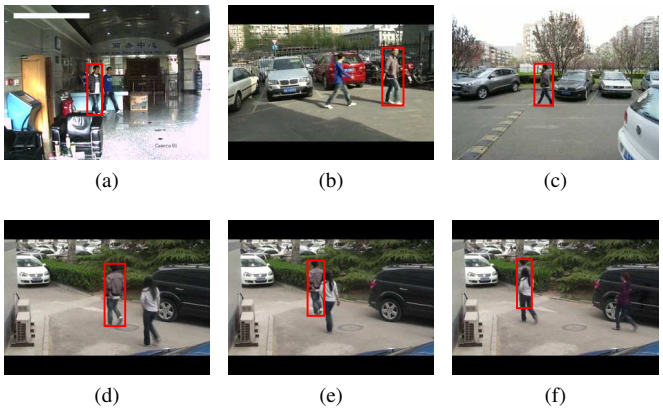


Fig. 1: Examples of MCT in NLPR_MCT Dataset4 (the same identity is marked with red bounding box).

I. Introduction

FOR security and safety purpose, the demands for surveillance cameras rapidly increase in the world in recent years. Because of limitation of Field Of Views (FOVs) and cost-efficiency, in most cases multiple cameras are installed in wide area with no overlap. One of the most important things in intelligent surveillance and monitoring system is automated object tracking for huge amount of recorded and live streaming video data. The goal of automated object tracking in the camera network is to keep the unique identity of each object within each single camera and across multiple cameras without human intervention. In other words, tasks involved in Multiple Camera Tracking (MCT) include multi-object tracking in each single camera and delivering detected identities to disjoint cameras.

Many MCT approaches exploit a two-step framework, Single-Camera Tracking (SCT) is first performed in each camera to create trajectories of multiple targets, and then Inter-Camera Tracking (ICT) is carried out to associate the tracks belonging to the same identity. There are several difficulties in ICT. First, people may exhibit dramatic changes on account of varied illuminations, viewing angles, poses and camera responses, under different cameras. Figures 1(a)-(c) show examples of ICT. One approach to solve this problem is person re-identification (re-id), which is to identify the same person in more than two cameras with only some human image-pairs. Many researchers on person re-id focus on extracting discriminative visual features to characterize the appearance of individual human without taking advantage of context information, e.g., group behavior between a pair of accompanying persons. Second, to achieve a good performance in ICT, a robust SCT, which detects accurate object positions and keeps the same identity on each object, needs to be guaranteed. Since ICT algorithms derive the appearance and context features by using SCT results, i.e., detected/tracked objects and segmentation masks, ICT performance highly depends on SCT.

The research in SCT in recent years has shown significant progress and enhances ICT performance as well. Figures 1(d)-(f) show examples of SCT, where (e) shows partial occlusion and (f) shows severe occlusion. The task of single-target tracking has been well addressed by following the cues of appearance or silhouette of selected target. However, for multi-target tracking in a video, the situation is more complex due to the data association problem and the interactions among objects. Furthermore, because of variations in number of targets, we need to employ effective schemes to initialize the object locations in every frame, which are usually derived from appearance-based object detection [1], [2] or background subtraction [3]. The former category is called tracking by detection, and the latter tracking by segmentation. In [4], most of the state-of-the-art methods are in the tracking-by-detection school, in which they exploit the continuity in space and time, but the information of object appearance is seldom considered

Y.-G. Lee, Z. Tang and J.-N. Hwang are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98105, USA e-mail: {lygstj,zhtang,hwang}@uw.edu

Manuscript received December 31, 2016.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

to facilitate tracking.

In this paper, we propose a robust MCT system based on a two-step framework. For SCT, we present multi-object tracking within a single camera that can adaptively refine the segmentation results based on multi-kernel feedback from preliminary tracking to handle the problems of object merging and shadowing. Meanwhile, ICT can benefit from the optimal segmented foreground blobs of each object as well. Besides, detection in local object region is incorporated to address initial occlusion when people appear in groups. Additionally, we follow our previous work [5] to rely on Constrained Multi-Kernel (CMK) tracking to deal with occlusion. The presented SCT method has been partially described in [6]. In addition to giving a more detailed explanation of our proposed SCT method, we introduce how to combine the algorithm with advanced change detection to improve segmentation performance, add the detection module for enhancing robustness against initial occlusion, and conduct new experiments on the benchmark dataset of MCT.

For ICT, we present a fully unsupervised online learning approach which integrates discriminative visual features and context feature efficiently, and systematically builds camera link model without any human intervention. With a two-phase feature extractor, which consists of TWO-Way Gaussian Mixture Model Fitting (2WGMMF) and couple features in phase I, followed by the holistic color, regional color/texture features in phase II, the proposed method effectively and robustly identifies the same person across cameras. To be more specific, illumination variation is dealt with by a fully unsupervised color transfer method [7], and changes of poses and camera viewpoints are overcome with pose-invariant features, 2WGMMF and regional color/texture features. The context information is represented as couple feature, which describes a pair of person traveling together through the scene. Those features are integrated with fusion feature weights belonging to camera link model.

We evaluate our approach on the benchmark NLPR_MCT dataset [8] that is recorded with multiple disjoint cameras. Our proposed method outperforms all the state-of-the-art methods in both SCT and ICT.

The contribution of this paper mainly includes:

1) a robust two-step MCT method based on SCT by segmentation and local object detection, and a two-phase online feature learning ICT framework;
2) combination of the advanced change detection algorithm and multi-kernel feedback to preserve precise foreground segmentation;
3) effective integration of three pose-invariant color features;
4) incorporation of context-based couple feature with appearance cues; and
5) validated superior performance on benchmark NLPR_MCT dataset.

The rest of the paper is organized as follows. Related papers are reviewed in Section II and the system of SCT and object segmentation is introduced in Section III. The proposed ICT method is detailed in Section IV, and camera link model estimation in Section V. Comparative experimental results of our scheme with the state-of-the-art methods and discussions are shown in Section VI. Finally, we conclude this paper in Section VII.

II. Related Works

A. Single-Camera Object Tracking (SCT)

Among many SCT techniques for tracking a single target, kernel-based object tracking such as mean shift tracker [9] that searches for similar candidate model around local neighboring regions, has gained lots of popularity, because of its fast convergence and low computation. To improve kernel-based tracking, Chu $et\ al.$ [5] propose to handle occlusion based on adaptive multiple kernels with constraints on their spatial relation, i.e., CMK tracking, and the accuracy is comparable to the state-of-the-art trackers. In [10], they embed CMK tracking into a Kalman filtering tracking system to further increase computational efficiency. To extend the application to multi-object tracking, it is necessary to find a way to automatically define the locations of targets. Most of the top-ranked methods in SCT depend on object detection for target initialization [4], but none of them considers to combine the information from segmentation to jointly improve performance.

Robust object segmentation is essential to feature extraction in ICT and supporting intra-camera tracking by segmentation. Many recent works in this field emphasize the concept of adaptation. In [11], a regularized background adaptation for automatically controlling the learning rate of Gaussian Mixture Model (GMM) is presented. Hoffmann $et\ al.$ propose the Pixel-Based Adaptive Segmenter (PBAS) in [12], which utilizes two dynamic controllers to adaptively adjust the decision threshold and learning rate. Self-Balanced SENsitivity SEgmenter (SuBSENSE), as introduced by St-Charles $et\ al.$ [13], further improves PBAS by adding adaptation to local sensitivity and update rate, which allows the technique to rank among the top of the benchmark dataset of change detection, CDnet [14]. Nevertheless, none of the algorithms are designed specifically for supporting tracking, as they can easily fail when target(s) enter into area with similar background color (i.e., the problem of object merging), or encounter strong shadowing effect, when the subsequent SCT will be negatively influenced as well.

B. Inter-Camera Object Tracking (ICT)

Human tracking across multiple cameras has been one of the most active research topic in computer vision [15] and many approaches have been proposed to address this problem. Most of these approaches utilize appearance cues, which include color [16]–[18], texture [19], [20], and shape [21], [22] of targets, to describe human and match correct correspondence with spatio-temporal reasoning. However, the color appearance is easily influenced by illumination and viewpoint changes across cameras. To solve the problem, Brightness Transfer Functions (BTFs) [23], [24], which map color information between a pair of cameras, and appearance relationship [25]–[27] are modeled from training data. For spatio-temporal feature, transition time distribution, which is the probability of an object entering a camera view with a certain travel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

time given the location and velocity of its exit from the other camera view, is estimated [18], [28], [29]. However, it requires training data whose correspondences are pre-labeled. Since methods relying on human operators are ineffective and lacking in scalability, these supervised learning approaches are less feasible in practice.

For this reason, recently unsupervised [19], graph modeling [30], [31], and online learning [22], [32] methods are exploited. More specifically, Chu *et al.* [19] estimate camera link model as an optimization problem to build the relationship between directly connected camera pairs based on an unsupervised learning scheme. However, they need separate training data for training stage, and transition time distribution included in camera link model is less reliable in case of longer transition time, because the variance of traveling time between connected cameras increases. Chen *et al.* [30] treat multi-camera object tracking as a global tracklet association, which is formulated as a global Maximum A Posteriori (MAP) problem with Piecewise Major Color Spectrum Histogram Representation (PMCSHR) and minimum uncertainty gap measurements. Tracking performance is enhanced with an improved similarity metric, which equalizes the inter-camera similarities in [31]. However, the disappearing points need to be manually selected in each enter/exit area. Chen *et al.* [32] employ an adaptive learning method, which uses the spatio-temporal information and Markov chain Monte Carlo sampling, to learn both spatio-temporal and appearance relationships among cameras. Kuo *et al.* [22] collect the online training samples by observing the spatio-temporal constraints in a time sliding window and use the Multiple Instance Learning (MIL) algorithm to learn a discriminative appearance model online. However, their method is limited to utilize low-level appearance features.

To further improve the tracking performance, group information is also exploited as complementary features in recent approaches [21], [33], [34]. Cai *et al.* [21] propose a relative appearance context model of groups to mitigate ambiguities in individual appearance matching. However, their relaxed definition of the group-named neighboring set has no social connection, therefore their assumption that the same set of people will reappear in the neighboring camera is not always valid. Wei *et al.* [33] propose a subject-centric group feature to reduce the re-id ambiguity. However, the group feature is limited to improve the accuracy of individual re-id when there is an outlier, who is seen in just one camera without being seen in the other, and sensitive to noise of persons positions and velocities. Chen *et al.* [34] integrate social grouping behavior of an elementary group [35] and an online learned target-specific appearance model by using AdaBoost. The tracking problem is formulated using an online learned Conditional Random Field (CRF) model that minimizes a global energy cost. However, the effectiveness of grouping information is not guaranteed when object detection is not sufficiently robust.

III. Single-Camera Tracking and Object Segmentation

Since both accurate segmentation and SCT results are necessary for supporting ICT, we develop a robust tracking and segmentation system to achieve the goal. The proposed system is coined as MAST, short for "Multi-kernel Adaptive Segmentation and Tracking", because we make use of multi-kernel feedback to adaptively control the thresholding parameters in segmentation for preserving more foreground around object region. Figure 2 shows the overview flow diagram of the MAST architecture. Note that this framework is extendable for the use of any object segmentation method, tracking-by-segmentation method, and object detection method.

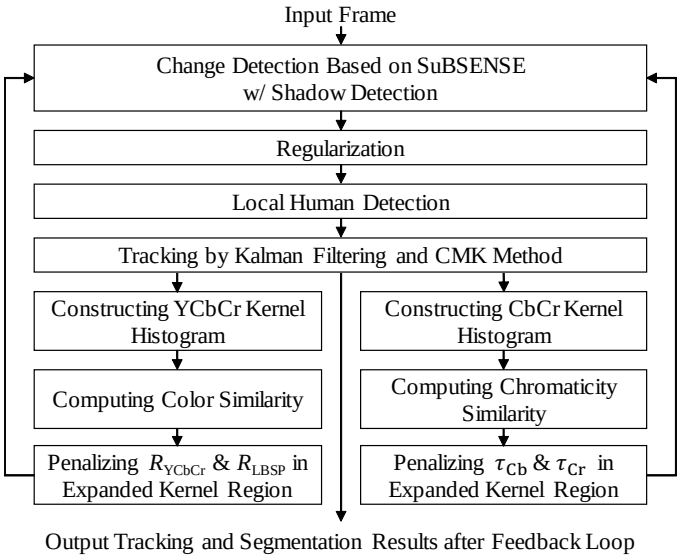


Fig. 2: Flow diagram of MAST for SCT and segmentation. The role of each block is detailed in Section III.

To begin with, the state-of-the-art change detection scheme, SuBSENSE [13], is adopted for object segmentation. Each pixel in the input frame is represented by color (here we choose to use YCbCr space instead of RGB space, since it will facilitate the process of shadow detection) and Local Binary Similarity Patterns (LBSP) feature [36]. The background model is constructed by a set of background samples $B_n(x,y)$ at each pixel location (x,y), which is updated according to an automatically adjusted learning rate. When each new pixel arrives for background/foreground classification, it will be compared with all background samples at the corresponding location. The comparisons are based on two distance thresholds, R_{YCbCr} and R_{LBSP}, in the color space and feature space, respectively. If the number of matching samples (with sufficiently short distance to the input pixel) is smaller than a specific minimum, the pixel is labeled as foreground. To further enhance robustness of SuBSENSE, we add a shadow detection block based on YCbCr color space that starts to function after a pixel is classified as foreground,

$$Q_t(x,y) = \begin{cases} 1, & \#\{\left(\alpha_{\text{Y}} \le I_t^{\text{Y}}(x,y)/B_n^{\text{Y}}(x,y) \le \beta_{\text{Y}}\right) \\ & \wedge \left(\left|I_t^{\text{Cb}}(x,y) - B_n^{\text{Cb}}(x,y)\right| \le \tau_{\text{Cb}}\right) \\ & \wedge \left(\left|I_t^{\text{Cr}}(x,y) - B_n^{\text{Cr}}(x,y)\right| \le \tau_{\text{Cr}}\right), \forall n\} \\ & > N_{\min}^{Q} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $Q_t(x,y)$ indicates shadow when the value is 1, $I_t(x,y)$ is a pixel from current frame t, the superscripts of $I_t(x,y)$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

$B_n(x, y)$ indicate the YCbCr channels, $N^Q_{\max}$ is the maximum number of matches required for shadow detection, and $\alpha_{\rm Y}$, $\beta_{\rm Y}$, $\tau_{\rm Cb}$ and $\tau_{\rm Cr}$ are the thresholds for their corresponding color channels. If a pixel is detected as shadow, it is discarded from foreground, and will be used for updating the background model. After segmentation, morphological operations, e.g., closing, opening, and flood-filling, are further applied on the derived foreground mask for shape refinement.

In the segmented foreground, each object blob may contain more than one target, i.e., the problem of initial occlusion. Thus, a Histogram of Oriented Gradient (HOG) human detector [1] is run on the cropped frame image within each object bounding box. If multiple targets are detected and their overlapping area with each other is small enough, they will be initialized separately for SCT. Different from traditional tracking by detection that needs to process each entire frame image, the computation complexity is much reduced since only the local region around each foreground blob is considered. Other than HOG human detector, we have also tested Deformable Part Model (DPM) human detector [2] and C^4 pedestrian detector [37]. HOG is chosen for its simplicity and efficiency.

Based on the initialization of object positions from segmentation and local object detection, we can start tracking each target. The preliminary tracking results are generated by the method proposed by Chu *et al.* that combines Kalman filtering and CMK tracking [10]. Kalman filter prediction is first conducted on all the objects tracked in the previous frame. Then we detect whether there is abnormality in size change of each foreground blob, which can be caused by occlusion or failure in segmentation. The abnormal targets and those initialized by object detection are tracked by the CMK method, which relies on multiple inter-related kernels to represent different parts of human, so that we can add weights of trust on different kernels depending on their severity of occlusion. Multiple measurements are produced from CMK tracking that are handled by probabilistic data association. On the other hand, the normal foreground blob with single object is directly selected as the measurement for Kalman filtering.

From preliminary tracking results, we follow the concept of multiple kernels to measure similarity between current frame and background in object regions. In our experiments, each human target is described by two kernels that cover half of his/her body on the top and bottom respectively, as people usually wear differently in these two body parts. Moreover, since the bounding box of each object may include background area, we can use kernel histogram to emphasize on the central region that the object occupies. Two kernel histograms are constructed within each kernel region for both current frame and background model: one of them is built in the YCbCr color space, and the other only uses the Cb and Cr channels to represent the chromaticity information. Note that the kernel histograms for background use all background samples and is normalized for comparison. To emphasize the object region that usually covers the central area of each kernel, a Gaussian kernel function is added for constructing kernel histograms,

$$w_{\rm ker} = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{(x-x_m)}{2\sigma_x^2} - \frac{(y-y_m)}{2\sigma_y^2}\right], \qquad (2)$$

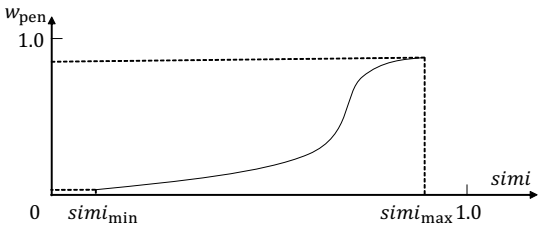


Fig. 3: The shape of fuzzy Gaussian penalty weighting function for adaptation of thresholding parameters in object segmentation.

(a) (b)

Fig. 4: Comparison of segmentation performance. (a) Segmentation from the preliminary result of SuBSENSE with shadow detection. (b) Segmentation after the application of multi-kernel feedback loops (foreground in red, and detected shadow in blue).

in which σ_x and σ_y are set as half of the width and height of the kernel bounding box respectively, while x_m and y_m locate the mean point of the foreground shape within the kernel.

Afterwards, the color similarity and chromaticity similarity are computed as the reciprocals of Bhattacharyya distances [38] between corresponding kernel histograms, i.e.,

$$simi_{\rm color} = 1 \Big/ \sum_c \sqrt{h^I_{\rm YCbCr}(c) \cdot h^{BG}_{\rm YCbCr}(c)}, \qquad (3)$$

$$simi_{\rm chrom} = 1 \Big/ \sum_c \sqrt{h^I_{\rm CbCr}(c) \cdot h^{BG}_{\rm CbCr}(c)}, \qquad (4)$$

where superscripts I and BG denote the kernel histograms in current frame and background respectively, and c is the index of channel bin. We have also tested other measurements such as correlation, Kullback-Leibler (KL) distance [39], dual KL distance [40], schemes in Automatic Reference Color Selection (ARCS) [41], etc. The Bhattacharyya distance is selected for its superior performance in our scenarios. The higher the color similarity of object region with background, the more likely the object will mistakenly merge into background during segmentation. Likewise, if the object region shares high similarity in chromaticity with the background, e.g., a human wearing black pants is walking on a grey ground plane, it is easy for his/her body parts to be wrongly recognized as shadow and removed from foreground. Next, a second segmentation using thresholding parameters penalized by $simi_{\rm color}$ and $simi_{\rm chrom}$ is performed in order to preserve more foreground in the local region around tracking targets. Under the consideration of smoothness of segmentation, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

penalty weights on segmentation thresholds are computed by a fuzzy Gaussian penalty weighting function as shown in (5),

$$w_{\text{pen}} = \begin{cases} \exp\left[-\frac{9\cdot(1.0-simi)^2}{4\cdot(1.0-simi_{\text{min}})^2}\right], & simi_{\text{min}} \le simi \le simi_{\text{max}} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

in which $simi$ is the color or chromaticity similarity computed from (3) or (4), respectively, while $simi_{\text{min}}$ and $simi_{\text{max}}$ indicate the region of $simi$ value to perform re-segmentation. As shown in the fuzzy Gaussian curve in Fig. 3, when $simi$ is smaller than the lower bound $simi_{\text{min}}$, the preliminary segmentation is considered successful, and there is no need for further adaptation, which is based on the concept of fuzzy set. On the contrary, if $simi$ is too large, it is highly likely to be caused by tracking error, where CMK tracking wrongly shifts to a background area. Hence, to prevent propagation of errors, an upper bound $simi_{\text{max}}$ is necessary for the similarity between current frame and background. The w_{pen} computed based on $simi_{\text{color}}$ is used to penalize R_{YCbCr} and R_{LBSP} in SuBSENSE, while the one for $simi_{\text{chrom}}$ is applied on τ_{Cb} and τ_{Cr} in shadow detection, where the penalization is defined by multiplying $(1 - w_{\text{pen}})$. Meanwhile, since the preliminary foreground blob may fail to cover the entire object body, the kernel region to conduct re-segmentation is expanded by a factor of $w_{\text{pen}}/2$. In summary, the adaptive segmentation is operated in a larger kernel region with lower thresholds for background subtraction and less shadow detected, therefore, the segmented foreground area is expanded to maintain continuity of tracking by segmentation. The final foreground mask is created by a union combination of the first segmentation across the entire frame and local adaptive segmentation in selected kernel regions.

Lastly, the tracking module is called again to generate the final tracking results from the updated foreground mask. Note that Kalman filter update is not performed until after re-segmentation. The optimized segmentation results will also be used in ICT for feature extraction. The superiority of adding multi-kernel feedback loops to adaptively control the segmentation thresholds can be seen from Fig. 4(b), in which more foreground belonging to the target is retained, compared to Fig. 4(a), when the chromaticity of her clothing is similar to background.

IV. Inter-Camera Object Tracking

In this section, we present main components of our proposed ICT methodology. An overview of the proposed approach is shown in Fig. 5. First, the SCT and segmentation results are acquired from each disjoint surveillance camera as input to ICT. The features are extracted on image domain, so more precise masks lead to better ICT results (see Table V). Examples are shown in Fig. 6, our proposed method gives more accurate segmentation masks compared to SuBSENSE method. To mitigate variations of illumination and color response among cameras, we transfer color characteristics of a source image to a target image before extracting features (Section IV-A). For extracting appearance features, an object is divided into three parts, head, torso and legs (Section IV-B). Our feature extractor consists of two phases, and the phase change occurs after having at least two good matches to build

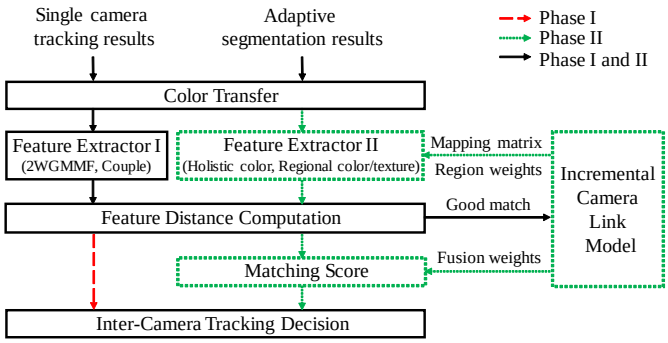


Fig. 5: An overview of our inter-camera multiple target tracking approach.

camera link model, which includes region mapping matrix, region matching weights and feature fusion weights. In phase I, ICT relies on 2WGMMF (Section IV-D) and couple features (Section IV-F). Subsequently, holistic color (Section IV-C) and regional color/texture features (Section IV-E) are further incorporated with feature fusion weights in phase II after the camera link model is systematically and continuously learned and updated (Section IV-G).

A. Color Transfer

The appearance of the same person may appear differently under two cameras because of illumination changes and different cameras color responses. In [7], [42], an algorithm corrects/transfers one images color characteristics to the other by de-correlating color space and statistical computation. More specifically, the RGB color space is transformed to the $l\alpha\beta$ color space and the data points composing the color transformed image are scaled by factors determined by the respective standard deviations in each channel as follows:

$$\begin{aligned} l_s^{'} &= \frac{\sigma_t^l}{\sigma_s^l}(l_s - \mu_s^l) + \mu_t^l, \\ \alpha_s^{'} &= \frac{\sigma_t^\alpha}{\sigma_s^\alpha}(\alpha_s - \mu_s^\alpha) + \mu_t^\alpha, \\ \beta_s^{'} &= \frac{\sigma_t^\beta}{\sigma_s^\beta}(\beta_s - \mu_s^\beta) + \mu_t^\beta, \end{aligned} \quad (6)$$

where μ and σ denote mean and standard deviation, and subscripts s and t denote source and target images, respectively.

We apply color characteristics transfer method between bounding boxes of two objects. Fig. 6(g) shows such an example of color transfer.

B. Body Partition

Since a pedestrian is commonly acquired at very low resolution in surveillance cameras, it is reasonable to notice that the most distinguishable body parts are three: head, torso and legs [43], [44]. Two boundary lines are systematically located and used to separate head-torso and torso-legs parts, respectively, as shown by the red lines in Fig. 7(b) and 7(c). From the top to the bottom of the rectangular foreground

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 6: (a) Source frame. (b) Global ID 6 in CAM4. (c) Masked image of (b) with SuBSENSE segmentation. (d) Masked image of (b) with the proposed segmentation. (e) Target frame. (f) Global ID 6 in CAM5. (g) Color transferred result of (f). (h) Masked image of (g) with SuBSENSE segmentation. (i) Masked image of (g) with the proposed segmentation.

bounding box, we calculate the histogram distance line-by-line between two block regions, i.e., the blue and green blocks in Fig. 7(a). Each block is of height δ_h and width W from a line T_i. Intuitively, we expect that color similarity between two different body parts to be low. Therefore, a boundary line is located at height T_i computed by solving the following problem, for both head-torso and torso-legs regions, respectively:

$$\max_{T_i \in \{S,E\}} d(\mathbf{h}_{[T_i, T_i+\delta_h]}, \mathbf{h}_{[T_i-\delta_h, T_i]}), \tag{7}$$

where $d(\cdot)$ denotes Euclidean distance and $\mathbf{h}_{[a,b]}$ denotes the color histogram derived from the region from a to b. Moreover, the boundary line is empirically assumed to be located within $\{S, E\}$, i.e., $\{0.18H, 0.25H\}$ for head-torso and $\{0.48H, 0.70H\}$ for torso-legs. In our experiments, 8-bin RGB histogram is employed and the height δ_h value is empirically set as 5 pixels.

C. Holistic Color Feature

Color histogram is widely used for representing color distributions [15]. In case two camera viewpoints are similar, the color histogram is effective to match the same person. So we utilize it as the holistic color feature to describe person clothing from head-torso boundary to the bottom. The total cost function for the holistic color feature is

$$d_{\text{holistic color}}(A, B) = d(\mathbf{h}^A, \mathbf{h}^B), \tag{8}$$

in which $\mathbf{h} \in \mathbb{R}^n$ denotes the holistic color histogram of the observation concatenating all color channels. In this paper, we also use 8-bin histogram for each RGB channel.

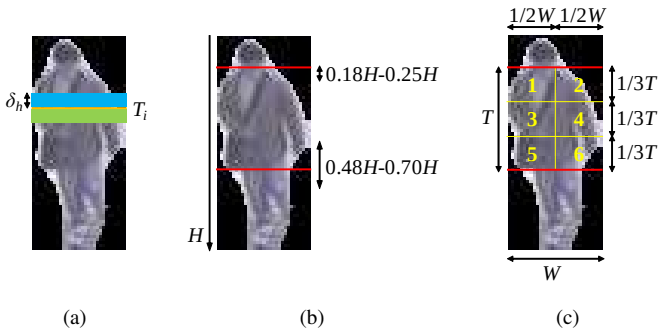


Fig. 7: An example of body partition. (a) Two blocks on masked image to find boundary lines using (7). (b) Two boundary lines on masked image. (c) Seven body regions for regional features.

D. 2WGMMF Feature

Since some parts of a human body in an image frame can be occluded by other parts of a human body, and the unseen body parts in an image can be visible in another image when their poses or camera viewpoints are changed. To handle the variation of poses and viewpoints, 2WGMMF feature [44] is thus employed.

The main idea of this feature is that main color modes of the same identity in color histogram domain should be consistent regardless the changes of poses and viewpoints. So 2WGMMF feature represents main color modes of a query person and candidates as GMMs, and computes the two-way distances (i.e., query-to-target and target-to-query) between the color histograms and GMMs. In detail, the feature distance between color histogram of person A and the GMM of person B can be computed by Negative Loglikelihood (NL) as follows:

$$\begin{aligned} d_{NL}(\mathbf{h}_i^A, G(\mathbf{h}_i^B)) &= -\ln p(\mathbf{h}_i^A | \theta_1^B, ..., \theta_K^B) \\ &= -\ln \left(\sum_{k=1}^{K} \pi_k^B \mathcal{N}\left(\mathbf{h}_i^A | \boldsymbol{\mu}_k^B, \boldsymbol{\Sigma}_k^B\right) \right), \end{aligned} \tag{9}$$

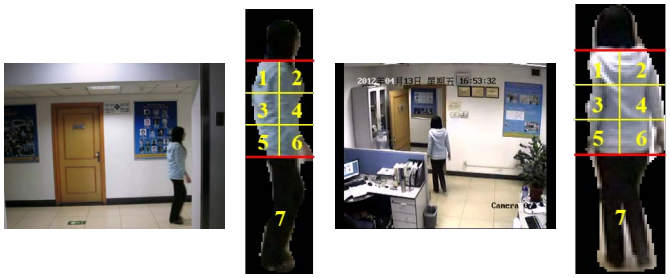
where $\mathbf{h} \in \mathbb{R}^{m^c}$ denotes joint color histogram of m-bin and c-channel, which is obtained from either of the body part $i = \{\text{torso}, \text{legs}\}$, $G(\cdot)$ denotes GMM from given color histogram and $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ indicates the set of parameters for component k. Moreover, π_k denotes the mixing proportion, $\boldsymbol{\mu}_k \in \mathbb{R}^c$ denotes the mean vector, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{c \times c}$ denotes the covariance matrix and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. K is the number of Gaussian components, i.e., the number of dominant color modes. Equation (9) computes the likelihood function of the GMM of B in response to the histogram of A.

The result from (9) is regarded as one-way distance of i-part, and a small value resulting from (9) indicates that they are likely to belong to the same identity. The 2WGMMF feature distance is represented as follows:

$$\begin{aligned} &d_{\text{2WGMMF}}(A, B) = \\ &d_{NL}\left(\mathbf{h}^A_{\text{torso}}, G(\mathbf{h}^B_{\text{torso}})\right) + d_{NL}\left(\mathbf{h}^A_{\text{legs}}, G(\mathbf{h}^B_{\text{legs}})\right) \\ &+ d_{NL}\left(\mathbf{h}^B_{\text{torso}}, G(\mathbf{h}^A_{\text{torso}})\right) + d_{NL}\left(\mathbf{h}^B_{\text{legs}}, G(\mathbf{h}^A_{\text{legs}})\right). \end{aligned} \tag{10}$$

Here, we use 32 bins for each channel in RGB color space.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology



(a) (b) (c) (d)

Fig. 8: (a) Frame 241 in CAM3 of Dataset3. (b) Seven body regions of Global ID 4 in CAM3. (c) Frame 67 in CAM4 of Dataset3. (d) Seven body regions of Global ID 4 in CAM4.

E. Regional Color and Texture Features

To enhance the ability of ICT through a more detailed comparison, we divide a human torso into multiple regions, since torso part usually carries richest and the most discriminant appearance. After body partition (Section IV-B), the torso part is further divided into six equal-size regions (r_1, r_2,..., r_6) as shown in Fig. 7(c). Because the region of legs (r_7) usually changes little under different perspectives, we do not further divide region of legs. Since each specific region normally covers different area of the human torso due to different viewpoints (see Fig. 8(b) and 8(d)), and as observed in Chu *et al.* [19] that a walking human is usually captured at the similar viewing perspective of body by a fixed camera on either the exit/entrance point, so the histogram extracted from one region of human torso can be modeled as a linear combination of the histograms extracted from multiple regions of human torso in the other camera,

$$\mathbf{h}^A_{map_k} = \left[\mathbf{h}^A_{r_1} \dots \mathbf{h}^A_{r_6}\right] \mathbf{w}_k, \tag{11}$$

where $\mathbf{h}^A_{r_k} \in \mathbb{R}^n$ denotes the regional color histogram extracted from the region k of the observation A and $\mathbf{w}_k \in \mathbb{R}^6$ is the mapping matrix of region k for linear combination.

Furthermore, because some regions may be visible only under one cameras view, they should have small weights in the feature distance computation. The distance of regional color feature is the weighted sum of the distances from all seven regions derived from torso and legs as

$$\begin{aligned} d_{\text{region color}}(A,B) = &\sum_{k=1}^{6} q_k \times d\left(\mathbf{h}^A_{map_k}, \mathbf{h}^B_{r_k}\right) \\ &+ q_7 \times d\left(\mathbf{h}^A_{r_7}, \mathbf{h}^B_{r_7}\right), \end{aligned} \tag{12}$$

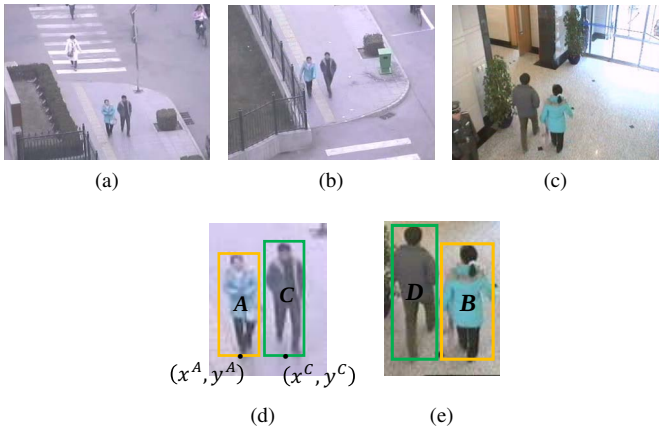
where $\mathbf{q} = [q_1 \dots q_7]^T$ denote the weights for all seven region distances. The computation of region matching weights is discussed in Section V-D. Note that all the seven regions are included in the feature distance computation, however only the torso regions are considered for the region mapping by using the mapping matrix $\mathbf{W}_{map} = [\mathbf{w}_1 \dots \mathbf{w}_6]$.

The texture feature distance can be computed similar to that of color feature. The Local Binary Pattern (LBP) [45] is exploited as the texture feature and is represented as l-dimensional LBP histograms of observation A, $\mathbf{h}^A_{r\text{LBP}_k} \in \mathbb{R}^l$, in which k is from 1 to 7. Hence, the distance of regional texture feature is

$$\begin{aligned} d_{\text{region texture}}(A,B) = &\sum_{k=1}^{6} q_k \times d\left(\mathbf{h}^A_{map\text{LBP}_k}, \mathbf{h}^B_{r\text{LBP}_k}\right) \\ &+ q_7 \times d\left(\mathbf{h}^A_{r\text{LBP}_7}, \mathbf{h}^B_{r\text{LBP}_7}\right), \end{aligned} \tag{13}$$

where $\mathbf{h}^A_{map\text{LBP}_k} \in \mathbb{R}^l$ is the linear combination of torso region LBP histograms with the same weights $\mathbf{w}_k$ defined in (11) as follows:

$$\mathbf{h}^A_{map\text{LBP}_k} = \left[\mathbf{h}^A_{r\text{LBP}_1} \dots \mathbf{h}^A_{r\text{LBP}_6}\right] \mathbf{w}_k. \tag{14}$$



(a) (b) (c) (d) (e)

Fig. 9: Examples of couple across multi-cameras. (a) Couple in CAM1. (b) Couple in CAM2. (c) Couple in CAM3. (d)(e) Cropped and enlarged couples in CAM1 and 3 (A-B and C-D denote the same person, respectively).

F. Couple Feature

We present a simple and effective group feature to improve the accuracy of ICT. Figures 9(a)-(c) show examples of a couple on three different cameras. In this paper, a couple is defined as a pair of persons traveling together through the scene and formulated as

$$\left|x^A - x^C\right| < \delta_x, \quad \left|y^A - y^C\right| < \delta_y, \tag{15}$$

$$\left|t^A_{\text{exit}} - t^C_{\text{exit}}\right| < \delta_t, \quad \left|t^A_{\text{entry}} - t^C_{\text{entry}}\right| < \delta_t, \tag{16}$$

where x and y denote the 2-D coordinate of center of bottom line of the bounding box (see Fig. 9(d)), and t_{exit} and t_{entry} denote time stamps when the person exits and enters FOV, respectively. With these spatio-temporal conditions, couples are detected in each camera.

To identify the same couple across cameras, 2WGMMF feature is again utilized as

$$\begin{aligned} &d_{\text{couple identifier}}(AC, BD) \\ &= \min\left(d_{\text{2WGMMF}}(A,B), d_{\text{2WGMMF}}(A,D)\right) \\ &+ \min\left(d_{\text{2WGMMF}}(C,B), d_{\text{2WGMMF}}(C,D)\right). \end{aligned} \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

In phase I, it is the negative of 2WGMMF feature distance between one and couple person of target that is used as follows,

$$\begin{aligned} d^{\mathrm{I}}_{\text{couple}}(A,B) &= -d_{\text{2WGMMF}}(A, B_{\text{couple}}) \\ &= -d_{\text{2WGMMF}}(A, D), \end{aligned} \tag{18}$$

and the couple feature distance exploits other feature distances to match person-to-person in a couple in phase II. The combination of four feature distances with feature fusion weights in phase II is shown as follows,

$$\begin{aligned} &d^{\mathrm{II}}_{\text{couple}}(A,B) = \\ &-\alpha_1 d^{Norm}_{\text{2WGMMF}}(A,D) - \alpha_2 d^{Norm}_{\text{holistic color}}(A,D) \\ &-\alpha_3 d^{Norm}_{\text{region color}}(A,D) - \alpha_4 d^{Norm}_{\text{region texture}}(A,D), \end{aligned} \tag{19}$$

where α_j denote feature fusion weights (see Section V-E). Note that (19) only shows a scenario of Comb1 (see Table II for different feature combinations). Moreover, to normalize feature distance, min-max normalization is used:

$$d^{Norm}_j(A,D) = \frac{d_j(A,D) - \min d_j}{\max d_j - \min d_j}, \tag{20}$$

where $\min d_j$ and $\max d_j$ represent the smallest and largest values of each feature j's distance, respectively. They are obtained from computing feature fusion weights with training data. In our experiments, δ_x, δ_y and δ_t are set to 15 empirically.

G. Final Score

In order to further improve the discriminative power, we utilize a combination of features for distance measures. Since the value range of each feature distance is different, we use normalization and fusion methods to get the final score. Feature fusion weights are derived by exploiting d-prime metric [46] (see Section V-E). In phase I, final score is combination of 2WGMMF and couple features as follows:

$$d^{\mathrm{I}}_{\text{Final}}(A,B) = d_{\text{2WGMMF}}(A,B) + d^{\mathrm{I}}_{\text{couple}}(A,B). \tag{21}$$

In phase II, final score is a combination of normalized feature distances of 2WGMMF, holistic color, regional color/texture and couple features with feature fusion weights. The following is formulation of the first category of combinations, Comb1:

$$\begin{aligned} &d^{\mathrm{II}}_{\text{Final}}(A,B) = \alpha_1 d^{Norm}_{\text{2WGMMF}}(A,B) \\ &+ \alpha_2 d^{Norm}_{\text{holistic color}}(A,B) + \alpha_3 d^{Norm}_{\text{region color}}(A,B) \\ &+ \alpha_4 d^{Norm}_{\text{region texture}}(A,B) + d^{\mathrm{II}}_{\text{couple}}(A,B). \end{aligned} \tag{22}$$

V. Camera Link Model Estimation

After collecting some video samples online (Section V-A), camera link models including transition time distributions for time window (Section V-B), region mapping matrix (Section V-C), region matching weights (Section V-D), and feature fusion weights (Section V-E) are estimated, and phase change occurs.

A. Online Sample Collection

In an FOV of a surveillance camera, a pedestrians appearance is usually captured in dozens of frames. So one good matching pair in ICT is equal to dozens of positive samples. If we have two good matching pairs, e.g., A-C and B-D, we can collect negative samples as well by cross matching pairs, e.g., A-D and C-B.

In our framework, good matching pairs are determined by the distance value calculated in (21). According to the characteristics of 2WGMMF feature, it has negative value when query and target images are very similar. In other words, there is rarely false positive pairs with negative value of 2WGMMF feature. Thus, good matching pairs are selected as follows:

$$d^{\mathrm{I}}_{\text{Final}} < 0, \tag{23}$$

for all the camera links. After obtaining two good matching pairs, we can start to build camera link models and update them by adding good matching pairs continuously.

B. Estimation of Time Window

People tend to walk in similar paths in most cases considering available pathways, obstructs, and shortest routes. Hence, the transition time t forms a certain distribution, and it can be exploited to infer and model the camera network topology [18], [28], [29]. We utilize the estimated distribution to determine the time window, which helps to reduce the number of candidates. Transition time distribution is modeled as a Gaussian distribution,

$$f(\forall t \in \mathbf{T} | \mu_T, \sigma_T) = \frac{1}{\sigma_T\sqrt{2\pi}} e^{-\frac{(t-\mu_T)^2}{2\sigma_T^2}}, \tag{24}$$

where $\mathbf{T} = [t_1, ..., t_N]$ represents a set of valid transition time values, μ_T and σ_T indicate mean and standard deviation of transition time distribution, respectively. Before having transition time distributions, we fix the length of time window $\tau = 120$ seconds, which can be set as different values for different known camera topologies. In all experiments, after estimating them, we set $\tau = \mu_T \pm 6\sigma_T$.

C. Estimation of Region Mapping Matrix

Both the regional color and texture features are exploited to estimate region mapping matrix as follows:

$$\mathbf{R}^A = \begin{bmatrix} \mathbf{h}^A_{r_1} \dots \mathbf{h}^A_{r_6} \\ \mathbf{h}^A_{r\mathrm{LBP}_1} \dots \mathbf{h}^A_{r\mathrm{LBP}_6} \end{bmatrix} = [\mathbf{r}_1 \dots \mathbf{r}_6], \tag{25}$$

where $\mathbf{r}_k \in \mathbb{R}^{n+l}$ for k = 1, 2,..., 6. We minimize the following objective function to get each vector $\bar{\mathbf{w}}_k$,

$$\begin{aligned} &\mathbf{w}_k = \arg\min_{\bar{\mathbf{w}}_k} g_{\mathbf{w}}(\bar{\mathbf{w}}_k) \\ &\text{s.t.} \quad g_{\mathbf{w}}(\bar{\mathbf{w}}_k) = \sum_{i=1}^{N_{\text{exit}}} \sum_{j=1}^{N_{\text{entry}}} \left\| \mathbf{R}^A_j \bar{\mathbf{w}}_k - r^A_{ik} \right\|_2^2, \\ &\bar{\mathbf{w}}_k \geq 0, \quad \left\| \bar{\mathbf{w}}_k \right\|_1 = 1, \end{aligned} \tag{26}$$

where N_{exit} and N_{entry} denote the number of exiting and entering observations respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

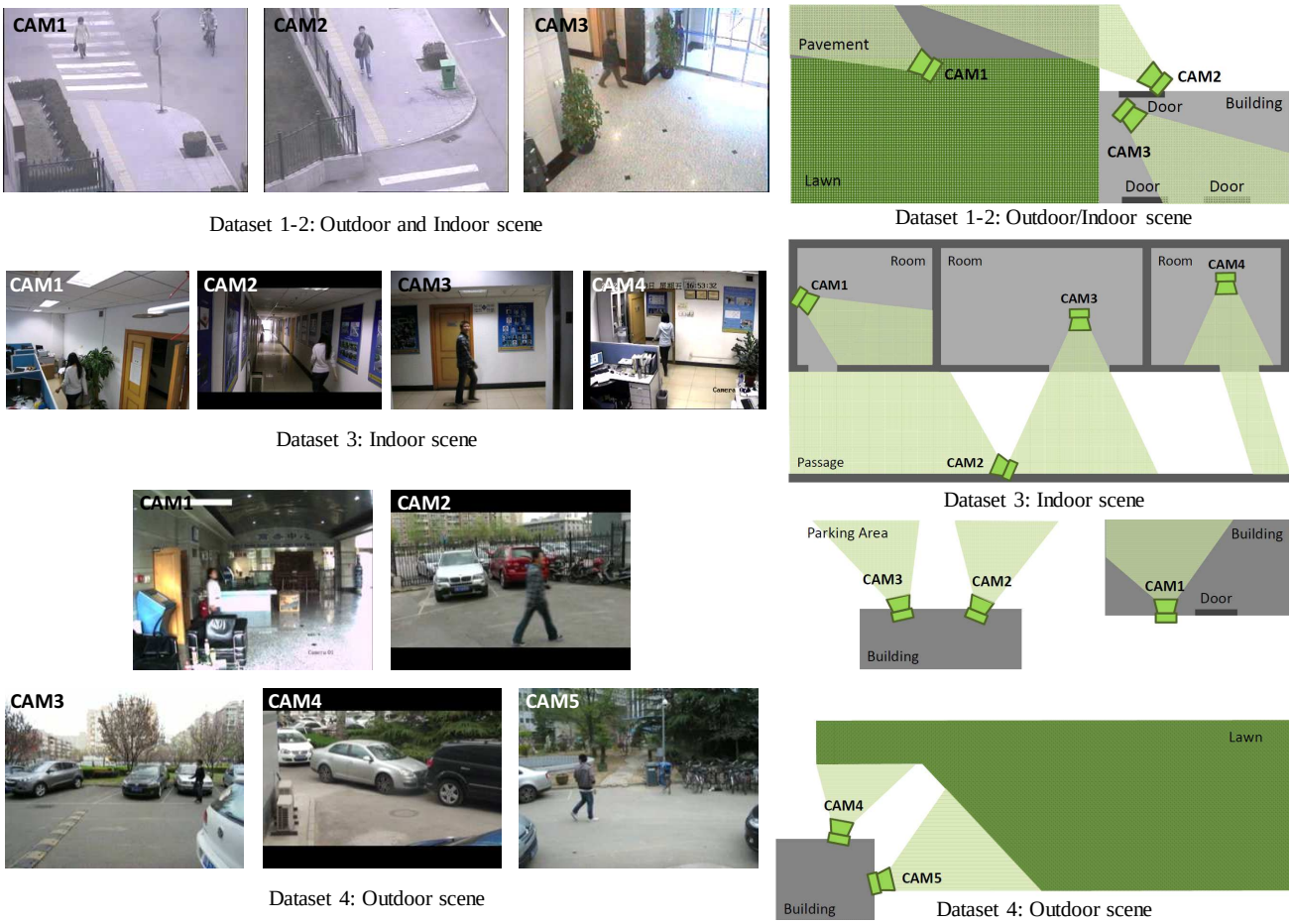


Fig. 10: Illustration of the topological relationship during tracking.

TABLE I: Details of NLPR_MCT Dataset [8].

Sub-dataset	Dataset1	Dataset2	Dataset3	Dataset4
# of cameras	3	3	4	5
Duration	20 min	20 min	3.5 min	24 min
Resolution	320×240	320×240	320×240	320×240
Frame rate	20 fps	20 fps	25 fps	25 fps
# of persons	235	255	14	49
GT^s	71853	88419	18187	42615
GT^c	334	408	152	256

D. Estimation of Region Matching Weights

The matching weights method in [46] is employed to determine the region matching weights, which are inversely proportional to the corresponding estimation error as follows:

$$q_k = \frac{1/\varepsilon_k^{error}}{\sum_{i=1}^{7} 1/\varepsilon_i^{error}} \quad k = 1 - 7, \tag{27}$$

where the estimation error of the mapping vector is defined as $\varepsilon_k^{error} = g_{\mathbf{w}}(\mathbf{w}_k)$ for k = 1, 2,..., 6 and $\varepsilon_7^{error} = \sum_{i=1}^{N_{\text{exit}}} \sum_{j=1}^{N_{\text{entry}}} \left\| \mathbf{r}_{j7}^A - \mathbf{r}_{i7}^B \right\|_2^2$. A large weight implies that the body region in one camera is well visible in the other camera as well.

E. Estimation of Feature Fusion Weights

Since there are four features, holistic color, 2WGMMF, region color, and region texture features, used in ICT, we need an efficient method to fuse them together. The feature fusion weights are systematically determined based on the degree of separation between the distributions of the values in the positive and negative sets. The separation is measured by the d-prime metric [46], [47],

$$d_j = \frac{\mu_j^{\text{N}} - \mu_j^{\text{P}}}{\sqrt{(\sigma_j^{\text{N}})^2 + (\sigma_j^{\text{P}})^2}}, \tag{28}$$

where μ_j and σ_j denote the mean and standard deviation of the distribution of the feature distance for each feature j; and superscripts P and N represent positive and negative sets, respectively. The feature fusion weights, α_j, j = 1, ..., 4, are calculated as

$$\alpha_j = \frac{d_j}{\sum_{i=1}^{4} d_i}. \tag{29}$$

VI. Experimental Results

This section presents the evaluation results of our approaches on the benchmark dataset, NLPR_MCT [8], which is specifically created for multi-camera pedestrian tracking over non-overlapping cameras. We introduce the details of the dataset and the evaluation criteria used in our experiments

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II: Performance comparison of multiple camera tracking without ground-truth of object detection. The best results are highlighted in colors (Underlined red font is rank-1 and *italicized green* font is rank-2).

Sub-dataset	Evaluation metric	Comb1	Comb2	Comb3	Comb4	USC-Vision [48] + [21]	NLPR [31]	Hfutdspmct	CRIPAC-MCT [30]
Dataset1	*Precision*	0.7724				0.6916	0.7967	0.7113	0.1488
	Recall	0.6088				0.6061	0.5929	0.3465	0.2154
	Detection	0.6809				0.6460	0.6799	0.4660	0.1760
	$Tracking^{\text{SCT}}$	0.9981				0.9981	0.9744	0.9229	0.9955
	SCTA	0.6796				0.6448	*0.6625*	0.4301	0.1752
	$Tracking^{\text{ICT}}$	0.8851	0.8851	0.8665	0.8789	0.9288	0.6220	0.6534	0.7111
	MCTA	0.6015	0.6015	0.5889	0.5973	*0.5989*	0.4120	0.2810	0.1246
Dataset2	*Precision*	0.8334				0.6948	0.7977	0.7461	0.1431
	Recall	0.7091				0.7843	0.6332	0.3669	0.1933
	Detection	0.7662				0.7368	0.7060	0.4919	0.1645
	$Tracking^{\text{SCT}}$	0.9991				0.9986	0.9779	0.9347	0.9945
	SCTA	0.7655				*0.7358*	0.6904	0.4598	0.1636
	$Tracking^{\text{ICT}}$	0.8842	0.8793	0.8818	0.8768	0.8691	0.6942	0.6122	0.7510
	MCTA	0.6769	0.6732	*0.6751*	0.6713	0.6260	0.4793	0.2815	0.1075
Dataset3	*Precision*	0.6597				0.4750	0.8207	0.3342	0.0853
	Recall	0.7260				0.6615	0.5345	0.0986	0.1206
	Detection	0.6913				0.5529	0.6474	0.1523	0.0999
	$Tracking^{\text{SCT}}$	0.9864				0.9904	0.9749	0.9682	0.9715
	SCTA	0.6819				0.5476	*0.6312*	0.1475	0.0971
	$Tracking^{\text{ICT}}$	0.5461	0.5329	0.5329	0.5000	0.1014	0.2953	0.2432	0.1143
	MCTA	0.3724	*0.3634*	*0.3634*	0.3410	0.0555	0.1864	0.0359	0.0111
Dataset4	*Precision*	0.8758				0.5216	0.8355	0.7720	0.0606
	Recall	0.8600				0.7938	0.6193	0.1210	0.0944
	Detection	0.8678				0.6295	0.7113	0.2092	0.0738
	$Tracking^{\text{SCT}}$	0.9977				0.9948	0.9275	0.9865	0.9762
	SCTA	0.8658				0.6262	*0.6597*	0.2064	0.0720
	$Tracking^{\text{ICT}}$	0.6270	0.6151	0.5992	0.6071	0.5437	0.4308	0.2944	0.2950
	MCTA	0.5429	*0.5326*	0.5188	0.5257	0.3404	0.2842	0.0608	0.0213
Average MCTA		0.5484	*0.5427*	0.5366	0.5338	0.4052	0.3405	0.1648	0.0661

TABLE III: Description of feature combination in evaluation.

Denotation	Feature combination
Comb1	2WGMMF, holistic color, regional color/texture, couple
Comb2	2WGMMF, regional color/texture, couple
Comb3	2WGMMF, holistic color, couple
Comb4	holistic color, regional color/texture, couple

in Section VI-A. The performance comparison with the state-of-the-art schemes is presented in Section VI-B. Finally, we discuss the experimental results in Section VI-C.

A. Dataset and Evaluation Criteria

The NLPR_MCT dataset consists of four sub-datasets. Every sub-dataset includes 3-5 cameras with non-overlapping FOVs and details of them are summarized in Table I, where GT^s is the number of ground truths in a single camera and GT^c is the number of ground truths across cameras. FOVs and the topological relationships of all the cameras are shown in Fig. 10. We assume that the connectivity between entry/exit zones in multiple cameras has already been specified [18], [21], [42].

The evaluation metric adopted is called Multi-Camera object Tracking Accuracy (MCTA) [31]:

$$\begin{aligned}&\text{MCTA}\\&= Detection \times Tracking^{\text{SCT}} \times Tracking^{\text{ICT}}\\&= \left(\frac{2 \times Precision \times Recall}{Precision + Recall}\right)\left(1-\frac{\sum_t mme_t^s}{\sum_t tp_t^s}\right)\left(1-\frac{\sum_t mme_t^c}{\sum_t tp_t^c}\right)\\&= \text{SCTA} \times \left(1-\frac{\sum_t mme_t^c}{\sum_t tp_t^c}\right),\end{aligned} \tag{30}$$

where mme_t and tp_t denote the number of mismatches and ground truths, respectively at time t. MCTA ranges from 0 to 1, and higher value indicates better performance. The metric can be divided into three parts, detection, SCT and ICT abilities, which are corresponding to the three brackets in (30). $Detection$ in (30) is also known as F_1-score, which reaches its best value at 1 and worst at 0. New object is counted as inter-camera ground truth, tp_t^c, by default in this criterion. The evaluation kit is available in [8]. Moreover, to evaluate the performance of SCT specifically, we define the Single-Camera object Tracking Accuracy (SCTA) by discarding the term of $Tracking^{\text{ICT}}$.

B. Multiple-Camera Object Tracking Results

To evaluate the performance, several combinations of the proposed features are compared with the state-of-the-art met-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

TABLE IV: Performance comparison of inter-camera tracking with ground-truth single camera tracking. The best results are highlighted in colors (Underlined red font is rank-1 and *italicized green* font is rank-2).

Sub-dataset	Evaluation metric	Comb1	Comb2	Comb3	Comb4	USC-Vision [21]	CRF [34]	NLPR [31]	CRIPAC-MCT [30]	Hfutdspmct
Dataset1	mme^c	13	14	19	17	27	54	55	113	86
	MCTA	0.9611	*0.9581*	0.9431	0.9491	0.9152	0.8383	0.8353	0.6617	0.7425
Dataset2	mme^c	30	46	31	36	34	81	121	167	141
	MCTA	0.9265	0.8873	*0.9240*	0.9118	0.9132	0.8015	0.7034	0.5907	0.6544
Dataset3	mme^c	32	35	41	36	70	51	39	44	40
	MCTA	0.7895	*0.7697*	0.7303	0.7632	0.5163	0.6645	0.7417	0.7105	0.7368
Dataset4	mme^c	62	69	72	80	72	70	157	110	155
	MCTA	0.7578	*0.7305*	0.7188	0.6875	0.7052	0.7266	0.3845	0.5703	0.3945
Average MCTA		0.8587	*0.8364*	0.8291	0.8279	0.7625	0.7577	0.6662	0.6333	0.6321

hods [21], [30], [31], [48] in Table II. The details of feature combinations are described in Table III. USC-Vision team [21], [48] is the winner of the MCT challenge in conjunction with ECCV 2014, and they utilize a two-step approach. They employ a four-body-part-based pedestrian detector [49] and a detection-based three-level hierarchical association approach [48], [50] for SCT. During the three-level association, detections are connected into the final trajectories by selecting detections discretely and complementing some missing detections. For ICT, they use two kinds of context information, spatio-temporal context and relative appearance context, to improve the ICT performance [21]. Spatio-temporal context is used to collect positive/negative training samples for each tracked object. Relative appearance context is used to model inter-object appearance similarities between the query person and the people in the neighboring set, which is defined as other people who enter/exit this zone in a time window with the query person. NLPR team [31] uses the DPM detector [2] to get the detection results, and applies an equalized global graph model that combines PMCSHR with an accurate similarity equalizer to compensate the weak invariance of appearance representation for ICT. CRIPAC-MCT team [30] exploits a head-shoulder detector [51] and an adaptive integrated feature (AIF) tracker [52] to get all the tracklets from each camera. Then, tracklets are merged into trajectories by a global tracklet association, which models ICT as a global MAP problem. Hfutdspmct team utilizes the Visual Background extractor (ViBe) algorithm [53] for detecting foreground. For SCT, their method is based on center location bi-directional matching. As for ICT, adjacency constrained patch matching as well as bi-directional weighted matching are applied.

In Table II, all the combinations, Comb1 to Comb4, of our proposed algorithm outperform the state-of-the art methods in terms of average MCTA. Especially, Comb1, which has combination of all the features, shows the best results. It proves that the proposed algorithm is robust in various environments.

We adopt the default setting of parameters in [6] for SCT. From the results of SCT in terms of F_1-score ($Detection$), mismatches ($Tracking^{\text{SCT}}$) within a single camera, and SCTA, it can be seen that our proposed method based on MAST also achieves the most robust overall performance in all the four scenarios. The main advantage of MAST over general tracking-by-detection is that we effectively combine the information from segmentation with local object detection, so that our method is less affected by the false positives generated by human detector in the entire frame. Moreover, the continuity of object tracking by segmentation is superior over those methods based on connecting trajectories, since there often exist many missing detections during tracking. This explains why the performance of MAST is more robust compared with the other state-of-the-art SCT used by each team. It can also be seen that the improved performance of intra-camera tracking and object segmentation also contribute to robust tracking across cameras. In addition, according to our parameters setting, the average runtime of our SCT together with object segmentation is 16.018 fps. The runtime is estimated on an Intel Core i7 PC with 2.67 GHz processor and 6G RAM in a Windows 7 environment.

C. Discussion

To focus on the ability of ICT, we have additional experiments, based on the ground-truth of SCT. More specifically, $Detection$ and $Tracking^{\text{SCT}}$ are both 1, thus, MCTA depends only on mme_t^c, which represents the number of mismatches in time t across different cameras in (30). In Table IV, the experimental results of the proposed method are compared with the state-of-the-art [21], [30], [31], [34]. Chen *et al.* [34] formulate ICT as an inference problem using the CRF framework. They first obtain the initial labels using Hungarian algorithm. Then, a global appearance model and an online learned target-specific appearance model using AdaBoost are combined with grouping information as high-level context feature to formulate the tracking task as an energy minimization problem. The problem is solved by their proposed iterative approximation algorithm. Our proposed method achieves the best result as shown in Table IV. With regard to average MCTA, Comb1 to Comb4 all perform better than the state-of-the-art methods. Moreover, Comb1 outperforms the other methods in every sub-dataset.

To validate the effectiveness of each single feature towards the final results, we also compare the performance of them in Table V. Compared to the performance of their combinations in Table IV, the performance based on each individual feature are worse. Note that 2WGMMF feature shows the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707399, IEEE Transactions on Circuits and Systems for Video Technology

TABLE V: Performance comparison of inter-camera tracking with single features. The best results are highlighted in colors (Underlined red font is rank-1 and *italicized green* font is rank-2).

		Holistic color		2WGMMF		Regional color		Regional texture	
Sub-dataset	Evaluation metric	SuBSENSE	Proposed	SuBSENSE	Proposed	SuBSENSE	Proposed	SuBSENSE	Proposed
Dataset1	mme^c	34	25	35	23	36	24	44	37
	MCTA	0.8982	0.9132	0.8952	0.9311	0.8922	*0.9281*	0.8683	0.8892
Dataset2	mme^c	52	49	60	55	82	64	88	78
	MCTA	*0.8725*	0.8800	0.8529	0.8652	0.7990	0.8431	0.7843	0.8088
Dataset3	mme^c	63	59	46	42	77	45	77	45
	MCTA	0.5855	0.6118	0.6974	0.7237	0.4934	*0.7039*	0.4934	*0.7039*
Dataset4	mme^c	95	87	75	72	87	73	94	90
	MCTA	0.6289	0.6602	0.7070	0.7188	0.6602	*0.7148*	0.6328	0.6484
Average MCTA		0.7457	0.7663	0.7881	0.8097	0.7112	*0.7975*	0.6947	0.7587

TABLE VI: Performance comparison of couple feature.

Sub-dataset	Evaluation metric	2WGMMF		Comb1	
		w/o couple	w/ couple	w/o couple	w/ couple
Dataset1	mme^c	23	19	20	13
	MCTA	0.9311	0.9431	0.9401	0.9611
Dataset2	mme^c	55	37	46	30
	MCTA	0.8652	0.9093	0.8873	0.9265

TABLE VII: Performance comparison of feature fusion weights and uniform weights on Comb1.

Sub-dataset	Evaluation metric	Uniform weights	Feature fusion weights
Dataset1	mme^c	19	13
	MCTA	0.9431	0.9611
Dataset2	mme^c	33	30
	MCTA	0.9191	0.9265
Dataset3	mme^c	37	32
	MCTA	0.7566	0.7895
Dataset4	mme^c	70	62
	MCTA	0.7266	0.7578
Average MCTA		0.8364	0.8587

best performance and regional color feature is the second-best in terms of average MCTA. However, the performance of holistic color feature is better than 2WGMMF feature in Dataset2 because many people are crossing the cameras, between CAM1 and CAM2, which have similar viewpoints. In addition, the performance of the proposed segmentation method and SuBSENSE are compared. From the performance of every single feature, it can be seen that the proposed segmentation gives better results. Because erroneously included background or cropped body part makes feature representation inaccurate, more precise masks lead to better ICT results. Specifically, the performance of regional color and texture features with SuBSENSE segmentation are much degraded. Because regional color and texture features are extracted from small areas comparably, they are more sensitive to the accuracy of segmentation.

In Table VI, the performance of couple feature in ICT is compared with 2WGMMF feature and Comb1. In all cases, the performance is improved when combining with the couple feature. Since couples only appear in Dataset1 and 2 only, there is no such comparisons for Dataset3 and Dataset4. USC-Vision [21] exploits relative appearance context learning, which is motivated by the fact that the same sets of people tend to re-appear in the neighboring camera. However, it is not applicable when FOV is limited with few people being captured. As a result, their performance is the worst in Dataset3 compared to the other methods in Table IV.

For showing the effect of feature fusion weights, we put uniform weights in (22) in case of Comb1 and the experimental results are shown in Table VII. With feature fusion weights, the overall accuracy is enhanced. It is because some features become more discriminative and other features are less effective according to relative difference of camera viewpoint. For example, the link between CAM1 and CAM2 is almost the same in Dataset1 and Dataset2 (see Fig. 10). Then, holistic color histogram is the most effective and its weight becomes larger. However, in case of the link between CAM2 and CAM3, or CAM3 and CAM4 in Dataset3, regional features, i.e., 2WGMMF and regional color/texture features, should have larger weights.

VII. Conclusion

We propose a robust approach for tracking the same identity across multiple cameras based on online learning in a fully unsupervised manner. In SCT, we introduce a method that depends on tracking by multi-kernel adaptive segmentation with assistance of local object detection, which also generates optimal foreground mask for the extraction of features in ICT. We make use of color transfer method to mitigate the change of illumination in ICT. The pose-invariant appearance features are exploited to overcome variation of poses and camera viewpoints between adjacent cameras. Moreover, the combination with context feature improves the performance of ICT. We demonstrate significant advantages compared to the state-of-the-art methods on the benchmark dataset representing real-world camera network scenarios. In the future, we will consider to employ self-calibration from tracking [54] to track

the person in 3D space so as to further improve the capability of tracking.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2. IEEE, 1999.

[4] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[5] C.-T. Chu, J.-N. Hwang, H.-I. Pai, and K.-M. Lan, "Tracking human under occlusion based on adaptive multiple kernels with projected gradients," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1602–1615, 2013.

[6] Z. Tang, J.-N. Hwang, Y.-S. Lin, and J.-H. Chuang, "Multiple-kernel adaptive segmentation and tracking (mast) for robust object tracking," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1115–1119.

[7] P. Shirley, "Color transfer between images," *IEEE Corn*, vol. 21, pp. 34–41, 2001.

[8] "Multi-Camera Object Tracking (MCT) challenge [online]," http://mct.idealtest.org/index.html.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[10] C.-T. Chu, J.-N. Hwang, S.-Z. Wang, and Y.-Y. Chen, "Human tracking by adaptive kalman filtering and multiple kernels tracking with projected gradients," in *Distributed Smart Cameras (ICDSC), 2011 Fifth ACM/IEEE International Conference on*. IEEE, 2011, pp. 1–6.

[11] H.-H. Lin, J.-H. Chuang, and T.-L. Liu, "Regularized background adaptation: a novel learning rate control scheme for gaussian mixture modeling," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 822–836, 2011.

[12] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 38–43.

[13] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2015.

[14] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 1–8.

[15] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3–19, 2013.

[16] G. Lian, J.-H. Lai, C. Y. Suen, and P. Chen, "Matching of tracked pedestrians across disjoint camera views using ci-dlbp," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1087–1099, 2012.

[17] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, 2008.

[18] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–205.

[19] C.-T. Chu and J.-N. Hwang, "Fully unsupervised learning of camera link models for tracking humans across nonoverlapping cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 979–994, 2014.

[20] Y.-G. Lee, J.-N. Hwang, and Z. Fang, "Combined estimation of camera link models for human tracking across nonoverlapping cameras," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2254–2258.

[21] Y. Cai and G. Medioni, "Exploring context information for inter-camera multiple target tracking," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 761–768.

[22] C.-H. Kuo, C. Huang, and R. Nevatia, "Inter-camera association of multi-target tracks by on-line learned appearance affinity models," in *European Conference on Computer Vision*. Springer, 2010, pp. 383–396.

[23] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions." in *BMVC*, vol. 8. Citeseer, 2008, pp. 164–173.

[24] T. D'Orazio, P. L. Mazzeo, and P. Spagnolo, "Color brightness transfer function evaluation for non overlapping multi camera tracking," in *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*. IEEE, 2009, pp. 1–6.

[25] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 26–33.

[26] H. Lim, O. I. Camps, M. Sznaier, and V. I. Morariu, "Dynamic appearance modeling for human tracking," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 751–757.

[27] B. C. Matei, H. S. Sawhney, and S. Samarasekera, "Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3465–3472.

[28] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *European conference on computer vision*. Springer, 2006, pp. 125–136.

[29] C.-C. Huang, W.-C. Chiu, S.-J. Wang, and J.-H. Chuang, "Probabilistic modeling of dynamic traffic flow across non-overlapping camera views," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3332–3335.

[30] W. Chen, L. Cao, X. Chen, and K. Huang, "A novel solution for multi-camera object tracking," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 2329–2333.

[31] L. Cao, W. Chen, X. Chen, S. Zheng, and K. Huang, "An equalised global graphical model-based approach for multi-camera object tracking," *arXiv preprint arXiv:1502.03532v2*, 2016.

[32] K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen, "An adaptive learning method for target tracking across multiple cameras," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[33] L. Wei and S. K. Shah, "Subject centric group feature for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 28–35.

[34] X. Chen and B. Bhanu, "Integrating social grouping for multi-target tracking across cameras in a crf model."

[35] X. Chen, Z. Qin, L. An, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1242–1249.

[36] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 509–515.

[37] J. Wu, C. Geyer, and J. M. Rehg, "Real-time human detection using contour cues," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 860–867.

[38] A. K. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, no. 35, pp. 99–109, 1943.

[39] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[40] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[41] H.-C. Shih and E.-R. Liu, "Automatic reference color selection for adaptive mathematical morphology and application in image segmentation," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4665–4676, 2016.

[42] X. Chen, K. Huang, and T. Tan, "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognition*, vol. 47, no. 3, pp. 1126–1137, 2014.

[43] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local

features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2360–2367.

[44] Y.-G. Lee, S.-C. Chen, J.-N. Hwang, and Y.-P. Hung, "An ensemble of invariant features for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 470–483, 2017.

[45] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[46] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 450–455, 2005.

[47] K.-W. Chen and Y.-P. Hung, "Multi-cue integration for multi-camera tracking," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 145–148.

[48] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *European Conference on Computer Vision*. Springer, 2008, pp. 788–801.

[49] C. Huang and R. Nevatia, "High performance object detection by collaborative learning of joint ranking of granules features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 41–48.

[50] C. Huang, Y. Li, and R. Nevatia, "Multiple target tracking by learning-based hierarchical association of detection responses," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 898–910, 2013.

[51] M. Li, Z. Zhang, K. Huang, and T. Tan, "Rapid and robust human detection and tracking based on omega-shape features," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 2545–2548.

[52] W. Chen, L. Cao, J. Zhang, and K. Huang, "An adaptive combination of multiple features for robust tracking in real scene," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 129–136.

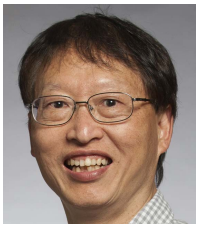
[53] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[54] Z. Tang, Y.-S. Lin, K.-H. Lee, J.-N. Hwang, J.-H. Chuang, and Z. Fang, "Camera self-calibration from tracking of moving persons," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 260–265.

Young-Gun Lee (S'07) received the B.S. degree in chemistry and physics from the Republic of Korea Air Force Academy in 2005, the M.S. degree in electrical engineering and computer science from the Seoul National University, Seoul, South Korea, in 2009. He is currently working toward the Ph.D. degree with University of Washington, Seattle, WA, USA.

His research interests include computer vision, image processing and video surveillance.

Zheng Tang (S'14) received the B.Sc. (Eng.) degree from the Joint Programme between Beijing University of Posts and Telecommunications and Queen Mary, University of London with First Class Honours in 2014. He received M.S. degree in electrical engineering from the University of Washington in 2016. He is currently a Ph.D. candidate in Information Processing Lab at the Department of Electrical Engineering of University of Washington. His current research interests include computer vision, machine learning, and video/image processing. Mr. Tang received the Finalist IBM Best Student Paper Award and the Finalist INTEL Best Student Paper Award on the 2016 International Conference on Pattern Recognition in Cancn, Mexico.



Jenq-Neng Hwang (F'01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from University of Southern California, Los Angeles, CA, USA.

In 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, where he was promoted to Full Professor in 1999. He has authored over 300 journal, conference papers, and book chapters in the areas of multimedia signal processing, and multimedia system integration and networking. He has authored the book $Multimedia Networking : From Theory to Practice$ (Cambridge University Press). His research interests include industry on multimedia signal processing and multimedia networking.

Dr. Hwang is a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society, where he is also a member. He is a member of the Multimedia Technical Committee of the IEEE Communication Society. He received the 1995 IEEE Signal Processing Societys Best Journal Paper Award. He served as the Program Co-Chair of the IEEE ICME 2016 and was the Program Co-Chair of the ICASSP 1998 and the ISCAS 2009. He served as the Associate Chair for Research from 2003 to 2005, and from 2011 to 2015. He is currently the Associate Chair for Global Affairs and International Development in the EE Department, University of Washington. He was the Societys representative to the IEEE Neural Network Council from 1996 to 2000. He served as an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE Signal Processing Magazine. He currently serves on the Editorial Board of ZTE Communications, Electronics and Telecommunications Research Institute, International Journal of Digital Multimedia Broadcasting, and Journal of Signal Processing Systems.