INTER-CAMERA TRACKING BASED ON FULLY UNSUPERVISED ONLINE LEARNING

Young-Gun Lee[⋆] *Zheng Tang*[⋆] *Jenq-Neng Hwang*[⋆] *Zhijun Fang*[†]

[⋆] University of Washington, Department of Electrical Engineering,
Seattle, WA 98195, USA
[†] Shanghai University of Engineering Science, School of Electronic and Electrical Engineering,
Shanghai, China

ABSTRACT

In this paper, we present a novel fully automatic approach to track the same human across multiple disjoint cameras. Our framework includes a two-phase feature extractor and an online-learning-based camera link model estimation. We introduce an effective and robust integration of appearance and context features. Couples are detected automatically, and the couple feature is also integrated with appearance features effectively. The proposed algorithm is scalable with the use of a fully unsupervised online learning framework. In the experiments, it outperforms all the state-of-the-art methods on the benchmark NLPR_MCT dataset.

Index Terms— Visual surveillance, human tracking, inter-camera tracking, NLPR_MCT dataset

1. INTRODUCTION

As the scale of camera networks is growing rapidly, human tracking across multiple disjoint cameras becomes more important. In particular, Inter-Camera Tracking (ICT), which establishes detected/tracked humans' correspondences across different cameras without overlapping Field Of Views (FOVs) to perform label handoff, is getting increasingly popular. Recently, many approaches have been proposed to address this problem. Most of them utilize brightness transfer functions [1, 2], spatio-temporal features [3, 4, 5] and appearance relationship [6, 7, 8] modeled from training data. However, the correspondences in training dataset need to be 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 graph modeling methods are exploited. More specifically, Chen *et al*. [9, 10] treat multi-camera object tracking as a global tracklet association, which is formulated as a global Maximum A Posteriori (MAP) problem with a Piecewise Major Color Spectrum Histogram Representation (PMCSHR) and minimum uncertainty gap measurement. However, disappearing points are manually selected and the use of MAP formulation is not a suitable solution in ICT [10]. To further improve the tracking performance, group information is also exploited as complementary features in recent approaches [11, 12]. Cai *et al*. [11] 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 indicates no clear social connection, therefore their assumption that the same set of people will reappear in the neighboring camera is not valid.

In this paper, we present a novel ICT method based on online learning, which systematically builds camera link model without any human intervention. Facilitated by the proposed 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 can effectively and robustly identify the same person across cameras. To be more specific, illumination variation is dealt with by a fully unsupervised color transfer method. Change of pose and camera viewpoint is overcome with pose-invariant features, 2WGMMF and regional color/texture features. The group feature is also incorporated to enhance the association performance. We evaluate our approach on the benchmark NLPR_MCT dataset [13] and achieve the best performance among all the state-of-the-art competing methods.

The rest of the paper is organized as follows. Our proposed framework is overviewed in Sec. 2. The proposed ICT method is detailed in Sec. 3. Comparative experimental results of our scheme with the state-of-the-art methods are shown in Sec. 4. Finally, we conclude this paper in Sec. 5.

2. SYSTEM OVERVIEW

An overview of the proposed approach is shown in Fig. 1. Robust segmentation results are first generated by a multi-kernel adaptive approach [14, 15]. To mitigate variation of illumination and color response among cameras, we transfer color characteristics of source image to target image before extracting features. Our feature extractor consists of two phases, and the phase change occurs after having at least two good matches to build camera link model, which includes region

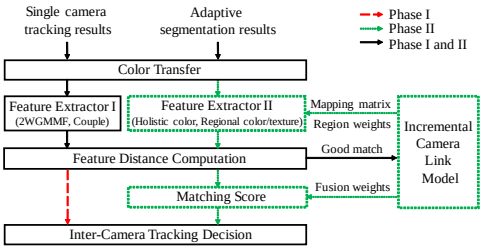


Fig. 1: An overview of the proposed ICT approach.

mapping matrix, region matching weights and feature fusion weights. Good matching pairs are determined by the score of feature combination. In phase I, ICT relies on 2WGMMF and couple features. Subsequently, holistic color and regional color/texture features are incorporated with feature fusion weights in phase II after the camera link model is systematically learned and continually updated.

3. ICT METHODOLOGY

In this section, main components of our proposed ICT methodology are addressed.

3.1. Color transfer

The same person may appear differently under two cameras because of changes in illumination and color responses. An effective approach is proposed in [16]. The RGB color space is transformed to the $l\alpha\beta$ color space and the data points composing the synthetic image are scaled by factors determined by respective standard deviation in each channel as follows:

$$I_s^{'} = \frac{\sigma_t}{\sigma_s}(I_s - \mu_s) + \mu_t, \tag{1}$$

where μ and σ denote mean and standard deviation of the I channel, and subscripts s and t denote source and target images, respectively. Transfer of color characteristics is applied within bounding boxes of objects. An example of color transfer is shown in Fig. 2(e).

3.2. Body partition

The most distinguishable body parts of pedestrians in FOVs are three: head, torso and legs [17]. Intuitively, we expect the color similarity between two different body parts to be low. A boundary line of body parts located at height T_i can be computed by solving the following problem:

$$\max_{T_i \in \{S,E\}} d(\mathbf{h}_{[T_i, T_i+\delta_h]}, \mathbf{h}_{[T_i-\delta_h, T_i]}), \tag{2}$$

where $d(\cdot)$ denotes Euclidean distance and $\mathbf{h}_{[a,b]} \in \mathbb{R}^n$ denotes the color histogram derived from the region of a to b. The boundary line is assumed to be located within $\{S, E\}$, e.g., $\{0.18H, 0.25H\}$ for head-torso and $\{0.48H, 0.70H\}$ for

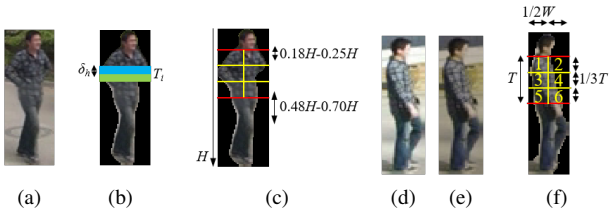


Fig. 2: Examples of body partition and color transfer in Dataset4 of NLPR_MCT. (a) Global ID 6 in CAM4. (b)-(c) Body partition of masked image of (a). (d) Global ID 6 in CAM5. (e) Color transferred result of (d). (f) Body partition of masked image of (e).

torso-legs. Examples are shown in Fig. 2(c) and (f). In our experiments, 8-bin histogram for each RGB channel is employed. The height δ_h value is empirically set to 5 pixels.

3.3. Holistic color feature

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 clothing of people from head-torso boundary to the foot point. The total cost function for the holistic color feature is

$$d_{\text{holistic color}}(A, B) = d(\mathbf{h}^A, \mathbf{h}^B), \tag{3}$$

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

3.4. 2WGMMF feature

The main idea of 2WGMMF feature [17] is that main color modes of the same identity in color histogram should be consistent across different viewpoints. The feature distance between color histogram of object A and the GMM of object B can be computed by Negative Loglikelihood (NL):

$$\begin{aligned} d_{NL}(\mathbf{h}_i^A, G(\mathbf{h}_i^B)) &= -\ln p(\mathbf{h}_i^A | \theta_1^B, ..., \theta_K^B) \\ &= -\ln \left(\sum_{k=1}^{K} \pi_k^B \mathcal{N}(\mathbf{h}_i^A | \boldsymbol{\mu}_k^B, \boldsymbol{\Sigma}_k^B) \right), \end{aligned} \tag{4}$$

where $\mathbf{h} \in \mathbb{R}^{m^c}$ denotes joint color histogram of m-bin and c-channel, that are obtained from either of the body part $i = \{\text{torso}, \text{legs}\}$, $G(\cdot)$ denotes GMM from inside color histogram and $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ denotes the set of parameters for component k. π_k denotes the mixing proportion, $\boldsymbol{\mu}_k \in \mathbb{R}^c$ the mean vector, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{c \times c}$ the covariance matrix and $\mathcal{N}(\cdot)$ the Gaussian distribution. K is the number of Gaussians, i.e., the number of dominant color modes.

The expression of (4) is regarded as one-way distance of i-part between two targets and a small value indicates that they are likely to be the same identity. The 2WGMMF feature

distance is

$$d_{\text{2WGMMF}}(A,B) = d_{NL}\left(\mathbf{h}^A_{\text{torso}}, G(\mathbf{h}^B_{\text{torso}})\right) + d_{NL}\left(\mathbf{h}^A_{\text{legs}}, G(\mathbf{h}^B_{\text{legs}})\right) + d_{NL}\left(\mathbf{h}^B_{\text{torso}}, G(\mathbf{h}^A_{\text{torso}})\right) + d_{NL}\left(\mathbf{h}^B_{\text{legs}}, G(\mathbf{h}^A_{\text{legs}})\right). \quad (5)$$

Here, we set 32-bin for each channel in RGB color space.

3.5. Regional color and texture features

To enhance the ability of ICT, we divide each human torso into multiple regions, since torso part usually carries the richest and most discriminant features. After body partition (Sec. 3.2), the torso part is divided into six regions based on the pre-defined ratios [18, 19] (see Fig. 2(f)). Since a specific region normally covers different areas of the torso due to different viewpoints (see Figs. 2(c) and 2(f)), the histogram extracted from one region of the torso can be modeled as a linear combination of the histograms extracted from multiple regions of the torso in the other camera, $\mathbf{h}^A_{map_k} = \left[\mathbf{h}^A_{r_1} \dots \mathbf{h}^A_{r_6}\right] \mathbf{w}_k$, where $\mathbf{h}^A_{r_k} \in \mathbb{R}^n$ denotes the regional color histograms and local binary pattern histograms [20] extracted from the region k of the observation A. $\mathbf{w}_k \in \mathbb{R}^6$ denote the mapping matrix of region k for linear combination. The total distance of regional features is the weighted sum of the distances from all seven regions derived from torso and legs as

$$d_{\text{regional feature}}(A,B) = \sum_{k=1}^{6} q_k \times d\left(\mathbf{h}^A_{map_k}, \mathbf{h}^B_{r_k}\right) + q_7 \times d\left(\mathbf{h}^A_{r_7}, \mathbf{h}^B_{r_7}\right), \quad (6)$$

where $\mathbf{q} = [q_1 \dots q_7]^T$ denote the weights for seven regional distances. 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 $\mathbf{W}_{map} = [\mathbf{w}_1 \dots \mathbf{w}_6]$.

3.6. Couple feature

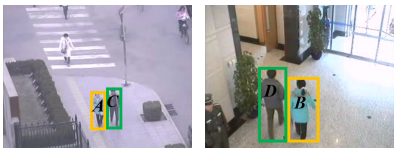
We present an effective group feature to improve the accuracy of ICT. A couple is defined as a pair of person traveling together through an FOV. Figure 3 shows an example of a couple on different cameras. After identifying the same couple across cameras with (7), persons are re-identified with (8) and (9). Couple identifier utilizes 2WGMMF feature as

$$\begin{aligned} &d_{\text{couple identifier}}(AC, BD) \\ &= \min\left(d_{\text{2WGMMF}}(A,B), d_{\text{2WGMMF}}(A,D)\right) \\ &+ \min\left(d_{\text{2WGMMF}}(C,B), d_{\text{2WGMMF}}(C,D)\right). \end{aligned} \quad (7)$$

To match person-to-person in a couple, in phase I, the negative 2WGMMF feature distance between a couple of target is utilized for measurement. In phase II, we adopt a combination of feature distances with feature fusion weights:

$$d^{\text{I}}_{\text{couple}}(A,B) = -d_{\text{2WGMMF}}(A, B_{\text{couple}}) = -d_{\text{2WGMMF}}(A,D), \quad (8)$$

$$d^{\text{II}}_{\text{couple}}(A,B) = -\sum_{j=1}^{N} \alpha_j d^{Norm}_{feature_j}(A,D), \quad (9)$$



(a) (b)

Fig. 3: An example of a couple in two cameras in Dataset1 of NLPR_MCT. (a) Couple in CAM1. (b) Couple in CAM3. (A-B and C-D correspond to the true matching pairs).

where α_j denote feature fusion weights and min-max normalization is used to normalize feature distances.

3.7. Final score

Since the value range of each feature distance is different, we exploit normalization and fusion methods to get the final score. By exploiting d-prime metric [21, 22], feature fusion weights are systematically determined based on the degree of separation between the distributions of the values in the positive and negative sets as $d_j = \mu^{\text{N}}_j - \mu^{\text{P}}_j \Big/ \sqrt{(\sigma^{\text{N}}_j)^2 + (\sigma^{\text{P}}_j)^2}$, where μ_j and σ_j denote the mean and standard deviation of the distribution of the feature distance for each feature j. Superscripts P and N represent positive and negative sets, respectively. The feature fusion weights are calculated as $\alpha_j = d_j \Big/ \sum_{i=1}^{4} d_i$. In phase I, the final score is a combination of 2WGMMF and couple features. The final score of phase II is a combination of normalized feature distances among of 2WGMMF, holistic color, regional color/texture and couple features with feature fusion weights as follow:

$$d^{\text{I}}_{\text{Final}}(A,B) = d_{\text{2WGMMF}}(A,B) + d^{\text{I}}_{\text{couple}}(A,B), \quad (10)$$

$$d^{\text{II}}_{\text{Final}}(A,B) = \sum_{j=1}^{N} \alpha_j d^{Norm}_{feature_j}(A,B) + d^{\text{II}}_{\text{couple}}(A,B). \quad (11)$$

3.8. Camera link model

In an FOV of a surveillance camera, a pedestrian is usually captured by dozens of frames. So one good matching pair is equivalent to dozens of positive sets. If we have two good matching pairs, i.e., A-B and C-D, we can collect negative sets by cross matching, i.e., A-D and B-C. Thus, camera link model estimation is available after having two good matching pairs. Phase transition occurs after establishing camera link models, which includes region mapping matrix ($\mathbf{W}_{map}$), region matching weights ($\mathbf{q}$), and feature fusion weights (α).

4. EXPERIMENTAL RESULTS

This section presents the evaluation results of our approaches on the benchmark, NLPR_MCT [13], which is collected for multi-camera tracking over non-overlapping cameras.

Table 2: Performance comparison of ICT with the state-of-the-art methods. The best results are highlighted in colors (Underlined red font is rank-1, *italicized green* font is rank-2 and **bold blue** font is rank-3).

Sub-dataset	Evaluation metric	Comb1	Comb2	Comb3	Comb4	USC-Vision [11]	CRF [23]	NLPR [10]	CRIPAC-MCT [9]	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

Table 1: Details of NLPR_MCT Dataset [13].

Sub-dataset	Dataset1	Dataset2	Dataset3	Dataset4
# of cameras	3	3	4	5
Duration	20 min	20 min	3.5 min	24 min
Frame rate	20 fps	20 fps	25 fps	25 fps
# of persons	235	255	14	49
GT^c	334	408	152	256

Table 3: 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

4.1. Dataset and evaluation criteria

The NLPR_MCT dataset consists of four sub-datasets and details are summarized in Table 1, where GT^c represents the number of ground truths across cameras. We assume that the connectivity between entry/exit zones in each camera is already specified [3, 11, 24].

The evaluation metric adopted is the widely used Multi-Camera object Tracking Accuracy (MCTA) [10]:

$$\begin{aligned}\text{MCTA} &= \text{Detection} \times \text{Tracking}^{\text{SCT}} \times \text{Tracking}^{\text{ICT}} \\ &= \left(\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)\left(1 - \frac{\sum_t mme_t^s}{\sum_t tp_t^s}\right)\left(1 - \frac{\sum_t mme_t^c}{\sum_t tp_t^c}\right),\end{aligned} \tag{12}$$

where mme_t and tp_t denote the number of mismatches and ground truths, respectively at time t. The MCTA ranges from 0 to 1. The ground truth of single camera tracking is used as inputs in our ICT experiments. Thus, MCTA depends only on tp_t^c and mme_t^c, which correspond to the accuracies across multiple cameras. The evaluation kit is available in [13].

4.2. Tracking results

The performance of the proposed method is compared with several state-of-the-art ICT methods [9, 10, 11, 13, 23]. USC-Vision [11] is the winner of the MCT challenge in conjunction with ECCV 2014. Cai *et al.* [11] propose to utilize spatio-temporal and relative appearance context to improve the performance. Chen *et al.* [23] formulate ICT as an inference problem using the CRF framework. Chen *et al.* [9] propose global tracklet association, which models ICT as a global MAP problem. Furthermore, Cao *et al.* [10] of NLPR propose an equalized global graph model by combining PMCSHR with a similarity equalizer to compensate the weak invariance of appearance representation in ICT. The results of Hfutdspmct are the rank-2 in MCT challenge [13].

In Table 2, several combinations of the proposed features are compared with the state-of-the-art methods. The details of feature combinations are described in Table 3. Comb1, which has combination of all the features, shows the best results in every sub-dataset. In terms of average MCTA, all the combinations of our proposed scheme outperform the competing methods. From the results, we can see that the most discriminative feature is 2WGMMF. In Dataset2, Comb3 and Comb4 perform better than Comb2, because many people cross between two cameras with a similar viewpoint, where holistic color feature is more effective. The performance of USC-Vision is the worst in Dataset3 compared to the other methods. It is because Dataset3 is recorded in an indoor environment, in which the FOVs are relatively narrow, so the assumption in [11] that the same sets of people tend to reappear in the neighboring camera is not applicable.

5. CONCLUSION

We propose a robust approach for tracking the same identity across multiple cameras. Our ICT algorithm based on fully unsupervised online-learning is scalable in practice. Two-phase feature extractor exploits the pose-invariant appearance features to overcome a variation of pose and camera viewpoint between two adjacent cameras. Moreover, context feature enhances the performance of ICT. We have shown significant advantages over the state-of-the-art methods on the real-world scenarios of camera network, NLPR_MCT dataset.

6. REFERENCES

[1] Bryan Prosser, Shaogang Gong, and Tao Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions.," in *Proc. BMVC*, 2008, vol. 8, pp. 74–84.

[2] Tiziana D'Orazio, Pier Luigi Mazzeo, and Paolo Spagnolo, "Color brightness transfer function evaluation for non overlapping multi camera tracking," in *Proc. ACM/IEEE ICDSC*, 2009, pp. 1–6.

[3] Dimitrios Makris, Tim Ellis, and James Black, "Bridging the gaps between cameras," in *Proc. IEEE CVPR*, 2004, vol. 2, pp. 205–210.

[4] Andrew Gilbert and Richard Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *Proc. ECCV*. Springer, 2006, pp. 125–136.

[5] Ching-Chun Huang, Wei-Chen Chiu, Sheng-Jyh Wang, and Jen-Hui Chuang, "Probabilistic modeling of dynamic traffic flow across non-overlapping camera views," in *Proc. IEEE ICPR*, 2010, pp. 3332–3335.

[6] Omar Javed, Khurram Shafique, and Mubarak Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Proc. IEEE CVPR*, 2005, vol. 2, pp. 26–33.

[7] Hwasup Lim, Octavia I Camps, Mario Sznaier, and Vlad I Morariu, "Dynamic appearance modeling for human tracking," in *Proc. IEEE CVPR*, 2006, vol. 1, pp. 751–757.

[8] Bogdan C Matei, Harpreet S Sawhney, and Supun Samarasekera, "Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features," in *Proc. IEEE CVPR*, 2011, pp. 3465–3472.

[9] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang, "A novel solution for multi-camera object tracking," in *Proc. IEEE ICIP*, 2014, pp. 2329–2333.

[10] Lijun Cao, Weihua Chen, Xiaotang Chen, Shuai Zheng, and Kaiqi Huang, "An equalised global graphical model-based approach for multi-camera object tracking," *arXiv:1502.03532v2*, 2016.

[11] Yinghao Cai and Gerard Medioni, "Exploring context information for inter-camera multiple target tracking," in *Proc. IEEE WACV*, 2014, pp. 761–768.

[12] Li Wei and Shishir K Shah, "Subject centric group feature for person re-identification," in *Proc. of the IEEE Conf. on CVPR Workshops*, 2015, pp. 28–35.

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

[14] Zheng Tang, Jenq-Neng Hwang, Yen-Shuo Lin, and Jen-Hui Chuang, "Multiple-kernel adaptive segmentation and tracking (mast) for robust object tracking," in *Proc. IEEE ICASSP*, 2016, pp. 1115–1119.

[15] Zheng Tang, Yen-Shuo Lin, Kuan-Hui Lee, Jenq-Neng Hwang, Jen-Hui Chuang, and Zhijun Fang, "Camera self-calibration from tracking of moving persons," in *Proc. IEEE ICPR*, 2016, pp. 260–265.

[16] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[17] Young-Gun Lee, Shen-Chi Chen, Jenq-Neng Hwang, and Yi-Ping Hung, "An ensemble of invariant features for person reidentification," *IEEE Trans. on CSVT*, vol. 27, no. 3, pp. 470–483, 2017.

[18] Chun-Te Chu and Jenq-Neng Hwang, "Fully unsupervised learning of camera link models for tracking humans across nonoverlapping cameras," *IEEE Trans. on CSVT*, vol. 24, no. 6, pp. 979–994, 2014.

[19] Young-Gun Lee, Jenq-Neng Hwang, and Zhijun Fang, "Combined estimation of camera link models for human tracking across nonoverlapping cameras," in *Proc. IEEE ICASSP*, 2015, pp. 2254–2258.

[20] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[21] Robert Snelick, Umut Uludag, Alan Mink, Mike Indovina, and Anil Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Trans. on PAMI*, vol. 27, no. 3, pp. 450–455, 2005.

[22] Kuan-Wen Chen and Yi-Ping Hung, "Multi-cue integration for multi-camera tracking," in *Proc. IEEE ICPR*, 2010, pp. 145–148.

[23] Xiaojing Chen and Bir Bhanu, "Integrating social grouping for multi-target tracking across cameras in a crf model," *IEEE Trans. on CSVT*, 2016.

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