Coupled Data Association and *l*1 Minimization for Multiple Object Tracking under Occlusion

Xue Wang, Qing Wang
School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, P. R. China

ABSTRACT

We propose a novel multiple object tracking algorithm in a particle filter framework, where the input is a set of candidate regions obtained from Robust Principle Component Analysis (RPCA) in each frame, and the goals is to recover trajectories of objects over time. Our method adapts to the changing appearance of objects, due to occlusion, illumination changes and large pose variations, by incorporating a *l*1 minimization-based appearance model into the Maximize A Posterior (MAP) inference. Though L1 trackers have showed impressive tracking accuracy, they are computationally demanding for multiple object tracking. Conventional data association methods using simple nonparametric appearance model, such as histogram-based descriptor, may suffer from drastic changing object appearance. The robust tracking performance of our approach has been validated with a comprehensive evaluation involving several challenging sequences and state-of-the-art multiple object trackers.

Keywords: Multiple object tracking, data association, L1 minimization, particle filter, MAP inference, RPCA

1. INTRODUCTION

Automatic detection and tracking of multiple objects is of prime importance for security systems and video surveillance applications. The fact that the camera remains stationary during the video sequence allow the use of background modeling techniques for the detection and tracking of moving objects such as [1-10].

For multiple object tracking, the data association between multiple objects and multiple observations is a first-line issue. A variety of data association algorithms have been proposed to tackle this problem, such as finding region correspondence using motion model [4], minimizing the cost function with network flows [11,12] and spatio-temporal Markov Chain Monte Carlo (MCMC) methods [13-15]. Most of these methods focus on motion consistency, and do not exploit appearance model. Usually a nonparametric histogram-based descriptor is used to represent the appearance of foreground area. However, data association only with such native appearance model apparently seems to have lost many discriminative visual features that could improve tracking. In [22], an appearance-adaptive model is incorporated in a particle filter to realize robust visual tracking and classification algorithms.

Sparse representation has been successfully applied to single object tracking by finding a sparse approximation in a template subspace [16-18]. The object is tracked through the video by extracting a template from the first frame and finding the object of interest in successive frames. This template update scheme could capture the appearance variations due to illumination or pose changes and occlusion by other scene objects. Nevertheless, for occlusion between multiple objects in a cluttered scene where moving objects interact frequently, updating the template set is not a good idea, because with a simple prior to model transition distribution, small errors are introduced each time the template is updated. The errors are accumulated and the tracker drifts from the target. Track ID switch errors often occur. In addition, the tracking methods with L1 minimization are computational ineffective.

In this paper we propose an improved data association algorithm for multi-object tracking (Fig.1). Given the candidate regions in each frame, obtained from RPCA and classic segmentation module, we aim to find the best association of observations, which maximizes the consistency of both motion and appearance of trajectories. Thanks to the explicit use of spatiotemporal smoothness in motion and appearance, we do not make a one-to-one assumption like [14]. Due to the high computational complexity of such an association scheme, a Data-Driven Markov Chain Monte Carlo (DD-MCMC) [23] method is adapted to sample the solution space. Both spatial and temporal association samples are incorporated into the Markov chain transitions.

Optoelectronic Imaging and Multimedia Technology III, edited by Qionghai Dai, Tsutomu Shimura,
Proc. of SPIE Vol. 9273, 927322 · © 2014 SPIE · CCC code: 0277-786X/14/$18
doi: 10.1117/12.2073887

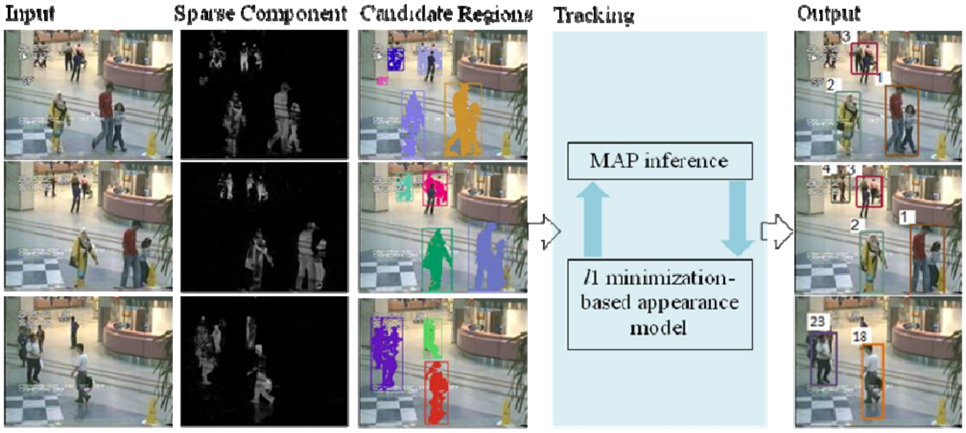

Figure 1. The framework of our proposed tracking algorithm. Given the candidate regions in each frame, obtained from RPCA and classic segmentation module, a $l1$ minimization-based appearance model is incorporated in the MAP inference, which leads to the best association of observations. A dynamic template update scheme [16] keeps track of the most representative templates throughout the tracking procedure.

The remainder of the paper is organized as follows. The related work is summarized in Section 2. The problem formulation and details of our multiple object tracking algorithm are represented in Section 3. Experimental results on several real surveillance videos and comparison with other methods are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. RELATED WORK

According to the depth of observations, the existing multiple object tracking methods can be categorized into either sequential inference or deferred logical inference. Compared with deferred logical inference, sequential inference methods are computationally more efficient [11-15], while considering more frames before association decision could generally help better overcome ambiguities caused by longer term occlusion and false or missed detection [19,20]. In [11,12], authors use an “out-of-the-box” pre-trained HOG pedestrian detector to generate the set of candidate locations prior to multi-object tracking. However, heavy occlusion will degrade their performance in a cluttered scene, because they only use motion constraint and native histogram-based appearance models, which can’t adapt to the drastic appearance changes between two successive trajectories.

The background maintains still in most surveillance videos, which enables many motion detection and tracking algorithms rely on the process of background subtraction, a technique which detects changes from a model of the background scene [1-10]. Background subtraction, binary morphology, and connected components analysis are generally the first processing steps in these methods. Since such object detection is a prerequisite for robust tracking, the background subtraction method should be capable of handling the problems caused by occlusion, long term or sudden illumination change and scene change. RPCA is proposed for recovering low-rank matrices from incomplete or corrupted observations and has been successfully used to solve back-ground subtraction [21]. By stacking the video frames as columns of a matrix D, the low-rank component A naturally corresponds to the stationary background and the sparse component E captures the moving objects in the foreground. However, considering the high dimensions of D, it would be impossible to decompose in such a way unless there is a truly scalable solution to this problem. So if each image frame has very high resolution, or the video fragment is long, generally we divide the whole video sequence into different segments.

Visual feature representation is also important for tracking problem. Sparse representation-based visual tracking algorithms are effective evidence. With the sparse assumption, the given signal can be approximately represented as the linear combination of a few basis vectors in the collected library. The target templates are stored and updated to form the dynamic dictionary in [16,17]. A new method is introduced to learn the dictionary as basis selection by gradient descent in [18]. Feature descriptions used include image patch containing the whole target [16], advanced feature vector as Haar-like features [17] and local image patch [18]. However, these tracking algorithm are all for one single target tracking and not robust for occlusion between multiple objects. The challenges of tracking multiple objects in sparse representation framework include dictionary construction and computational complexity.

3. MULTIPLE OBJECT TRACKING

3.1 Particle Filter

Suppose there are *K* unknown objects in the scene within the time interval [1,*T*]. The input for the tracking algorithm is a set of regions after RPCA and foreground segmentation. Let $Y = \bigcup_{t=1}^{T} y_t$ be the set of all available foreground regions within [1,*T*] and y_t be the set of foreground regions at time *t*. A track τ_k is a set of covering rectangles with the same label (track IDs). A cover *w* with *m* covering rectangles of *Y* can be written as follows:

$$w = \{x_i = (\mathbf{r}_i, t_i, l_i)\}, \mathbf{r}_i \in \prod_r, t_i \in [1,T], l_i \in [1,K], \quad (1)$$

where x_i is one covering rectangle, $\mathbf{r}_i$ and t_i represent the state (center position and size) and the timestamp for one rectangle, l_i indicates the label assigned to the rectangle $\mathbf{r}_i$, and *K* is the number of objects. $\prod_r$ is the set of all possible rectangles. Although the candidate space of possible rectangles is very large, it is still a finite number if we discretize the state of a rectangle in 2D image space. A cover with *K* tracks can also be written as $w = \{\tau_1, \cdots, \tau_K\}$. In a Bayesian formulation, the tracking problem is to find a cover to maximize a posterior of a cover of foreground regions, given the set of observations *Y*,

$$w^* = \arg\max P(w \mid Y). \quad (2)$$

We make inference about *w* from *Y* over a solution space $w \in \Omega$

$$w \square P(w \mid Y) \propto P(Y \mid w) P(w), w \in \Omega. \quad (3)$$

The likelihood $P(Y \mid w)$ is the observation likelihood representing how well the cover *w* explains the foreground regions *Y* in terms of the spatial-temporal smoothness in both motion and appearance. The prior model $P(w)$ regulates the cover to avoid overfitting the smoothness. To find a cover with reasonable properties, we define the prior model according to the following criterion: we prefer a small number of long tracks with little overlap with other tracks. As in [14], we adopt the prior probability of a cover *w* as the product of the following terms:

$$P(w) = P(N) P(L) P(O). \quad (4)$$

where $P(N)$, $P(L)$ and $P(O)$ are different exponential models for penalizing the number of tracks, length of each tracks and spatial overlap between different tracks respectively.

3.2 Observation Likelihood $P(Y \mid w)$

We assume the characteristics of motion and appearance of objects are independent; therefore, the likelihood can be written as follows:

$$P(Y \mid w) = f_F(w) \prod_{k=1}^{K} f(\tau_k), \quad (5)$$

where $f_F(w)$ represents the likelihood of the uncovered foreground area by *w* and $f(\tau_k)$ is the likelihood for each track. The area not covered by any rectangle indicates the false alarm in observations. The appearance of foreground regions covered by each track is supposed to be photo consistency and the motion of such a rectangle sequence should be smooth. Thus we can rewrite the likelihood for each track as the product of motion likelihood $f_M(\tau_k)$ and appearance likelihood $f_A(\tau_k)$. We represent the elements in track τ_k as $(\tau_k(t_1), \tau_k(t_2), \cdots, \tau_k(t_{|\tau_k|}))$, where $t_i \in [1,T]$ and $1 \le (t_{i+1} - t_i) \le \kappa$. Since missing detection may happen, we allow tracks to be composed of state vectors from non-consecutive frames e.g., we allow t_i and t_{i+1} to differ by up to κ frames. Motion likelihood is incorporated into a classic Markov chain transitions. Refer to [14] for more details. The appearance likelihood of one track is defined as follows:

$$f_A(\tau_k) = \prod_{i=1}^{|\tau_k|} L_A(\tau_k(t_i)) = \prod_{i=1}^{|\tau_k|} (1/\alpha) \exp(-\beta \, \mathrm{D}(\tau_k(t_i), \mathrm{T}_k \cdot \mathrm{a}_k^*(t_i))), \quad (6)$$

where $D(\cdot)$ represents the symmetric Kullback-Leibler Distance (KL) between the vectors of foreground covered by $\tau_k(t_i)$ and its nearest linear weighted combination $\mathrm{a}_k^*(t_i)\cdot \mathrm{T}_k$ of corresponding templates T_k , and $\mathrm{a}_k^*(t_i)$ is the coefficient vector obtained by solving a $l1$-regularized least squares problem, which is known to typically yield sparse solutions,

$$\mathbf{a}_k^*(t_i) = \arg\min \left\|\mathbf{T}_k \mathbf{a}_k(t_i) - \tau_k(t_i)\right\|_2^2 + \lambda \left\|\mathbf{a}_k(t_i)\right\|_1, \tag{7}$$

where $\mathbf{T}_k = [\mathbf{t}_{\mathrm{k},1}, \mathbf{t}_{\mathrm{k},2}, ..., \mathbf{t}_{\mathrm{k},n_k}] \in \mathrm{R}^{d\times n_k}$ $(d$ □ $n)$ contains n_k object templates. Any new rectangle $\tau_k(t_i) \in \mathrm{R}^d$ from the same object will approximately lie in the linear span of $\mathbf{T}_k$. Since in many visual tracking scenarios, target objects are often corrupted by noise or partially occluded. The occlusion leads to unpredictable errors. To incorporate the effect of occlusion and noise, we follow the scheme in [16] and use trivial templates to capture the occlusion. Since a fixed appearance template is not sufficient to handle recent changes in the video, we also adopt the dynamic template update scheme to deal with changing appearance. Since the appearance of pedestrian is textureless and in some real scenarios only low resolution surveillance videos are available, we use the most discriminative RGB color and stack template image to columns to form a 1D vector as one visual word.

Given the cover, the motion and appearance likelihood of an object is assumed to be independent of other objects. The joint observation likelihood of a cover can be factorized as follows:

$$P(Y \mid w) = f_F(w)\prod\nolimits_{k=1}^{K} f_M(\tau_k) f_A(\tau_k) = f_F(w)\prod\nolimits_{k=1}^{K}(\prod\nolimits_{i=3}^{|\tau_k|} L_M(\tau_k(t_i))\prod\nolimits_{i=1}^{|\tau_k|} L_A(\tau_k(t_i))), \tag{8}$$

where $L_M(\tau_k(t_i))$ is the likelihood between rectangle $\tau_k(t_i)$ and the predicted kinematic state.

3.3 Background Subtraction

Robust tracking of targets is impossible if the input regions are inaccurate. We use the RPCA method proposed by Candès et al. [21]. In this method, the video frames are stacked as columns of a matrix *D*, then the low-rank component *A* corresponds to the stationary background and the sparse component *E* captures the moving objects in the foreground. Since the rank *r* of is unknown, the problem is to find the matrix of lowest rank that could have generated when added to an unknown sparse matrix . Mathematically, the RPCA method is formulated as,

$$\min\nolimits_{A,E} \operatorname{rank}(A) + \lambda_1 \left\|E\right\|_0 \quad \text{subject to} \quad D = A + E, \tag{9}$$

where $\left\|\cdot\right\|_0$ denotes the counting norm (i.e., the number of non-zero entries in the matrix). Since this problem cannot be efficiently solved, Eq. (9) could be rewritten by its convex relaxation instead,

$$\min\nolimits_{A,E} \left\|A\right\|_* + \lambda_2 \left\|E\right\|_1 \quad \text{subject to} \quad D = A + E, \tag{10}$$

where $\left\|\cdot\right\|_1$ and $\left\|\cdot\right\|_*$ denote the matrix $l1$-norm and the nuclear norm respectively, and λ_2 is a positive constant to balance the two terms in $\left\|A\right\|_* + \lambda_2 \left\|E\right\|_1$ appropriately. The method reliably deals with changes in object appearance and illumination.

The candidate regions could be obtained from the sparse component *E*. In most general case, which is common in real scenarios, one foreground region may correspond to multiple objects (two examples are shown in Fig. 2), and one object may correspond to multiple foreground regions. In this case, without using any model information, it is difficult to segment the foreground regions in a single frame. However, if we consider this task in space-time, the smoothness in motion and appearance of objects can be used to solve this problem. Fig. 2 shows some results from RPCA and candidate regions for data association.

3.4 Optimization

Directly optimizing a posterior by enumerating all possible solutions in the solution space is simply not feasible. We use the data-driven MCMC proposed in [16] to estimate the best spatiotemporal cover of foreground regions. A track can be extended in both the positive and negative time directions; such bidirectional sampling technique has more flexibility and reduces the total number of samples.

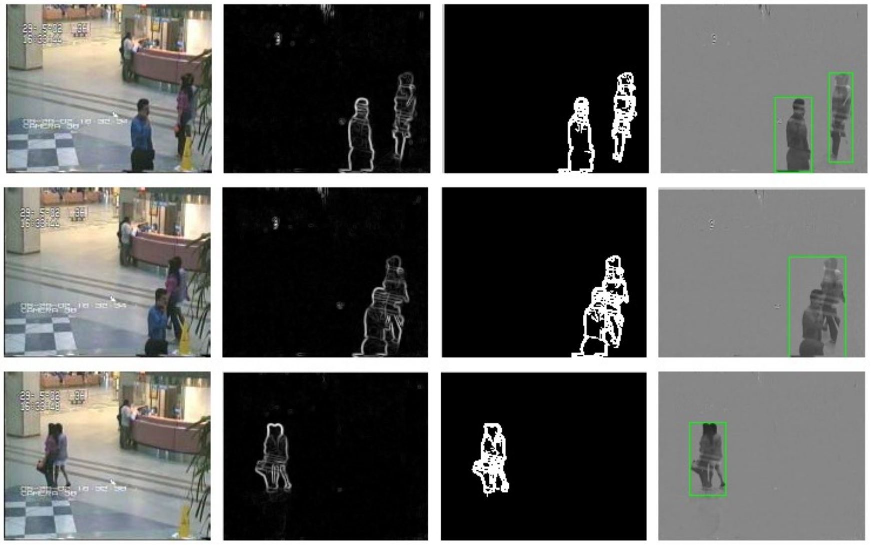

Figure 2. Some candidate regions containing foreground objects on the hall sequence from [10]. A minimal bounding rectangle is used here to represent foreground region. The frame indexes are 2435, 2438 and 2479.

4. EXPERIMENTS

The proposed coupled tracking algorithm is tested using three challenging surveillance videos (*hall, EnterExitCrossingPaths1cor, view01*) with 4761 frames in total, involving severe occlusion and pose changes. The method is compared with two latest state-of-art tracking methods, L1 tracker [7] and global data association with network flow [3], abbreviated by GlobalDA for convenience. The tracking results of the compared algorithms are obtained by running the source code provided by their authors using default parameters setting. Since the L1 tracker is proposed for tracking one single object, in order to deal with multiple objects scenario, we keep one L1 tracker for each object and get the final tracking results. The targets are selected by hand in the beginning frame.

The first sequence is obtained from [20]. As shown is Fig.3(a)(b), GlobalDA obtains many short tracklets and misses persons who have large pose change or do not match pedestrian HOG detector very well. We show the robustness of our algorithm to large pose changes.

In the sequence *EnterExitCrossingPaths1cor* from CAVIAR dataset, severe occlusion happens among three persons and the appearance model of two persons are similar. The results by our approach and L1 tracker are shown in Fig.3 (c) and (d) respectively. L1 tracker mixes up these three targets which occlude each other. The results show our approach is able to track objects throughout the occlusion.

The third sequence is *view01* from PETS09 dataset. As shown in Fig.3 (e), our approach can get reliable longer tracklets than GlobalDA and perform better than L1 tracker as the pose change and occlusion.

Given the ground truth in every frame of the second test sequence, we adopt position and overlap error quantitative comparisons [21]. We plot the overlap as well as the position error curves for 150 frames for three persons in Fig. 4. Apparently, our approach performs better than L1 tracker measuring with either position error or overlap. Occlusions degrade the performance of L1 tracker. The results of tracking No.2 person with L1 tracker seem good from the quantitative criterion, however, actually this track drift to another person close to No.2, which can be seen from Fig. 3(d). Because GlobalDA obtains short tracklets and there are several tracking IDs correspond to the same object, which make statistics over long period sequence difficult, so we do not compare our approach with GlobalDA quantitatively.

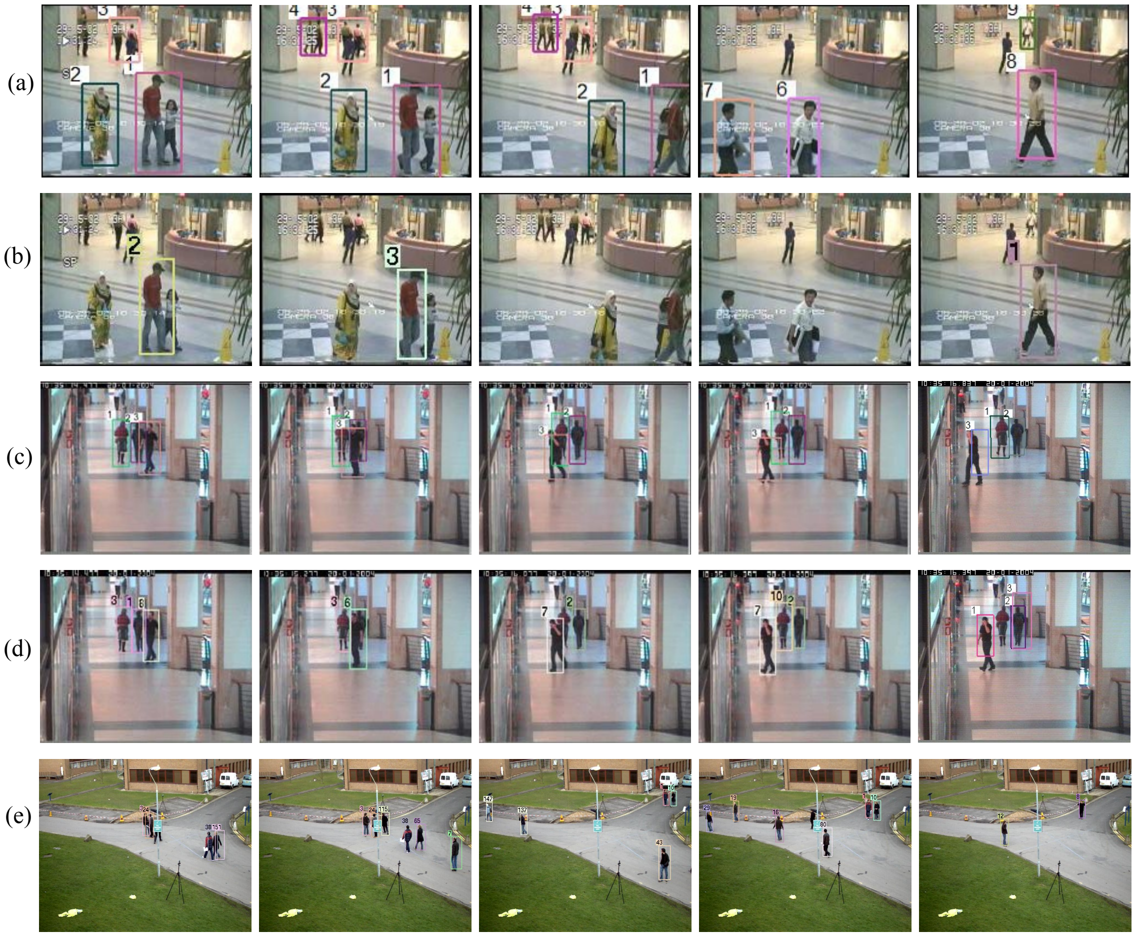

Figure 3. Tracking results of different algorithms on 3 challenging sequence: (a) our approach on hall, (b) GlobalDA on hall, (c) our approach on EnterExitCrossingPaths1cor, (d) L1 tracker on EnterExitCrossingPaths1cor, (e) our approach on view01.

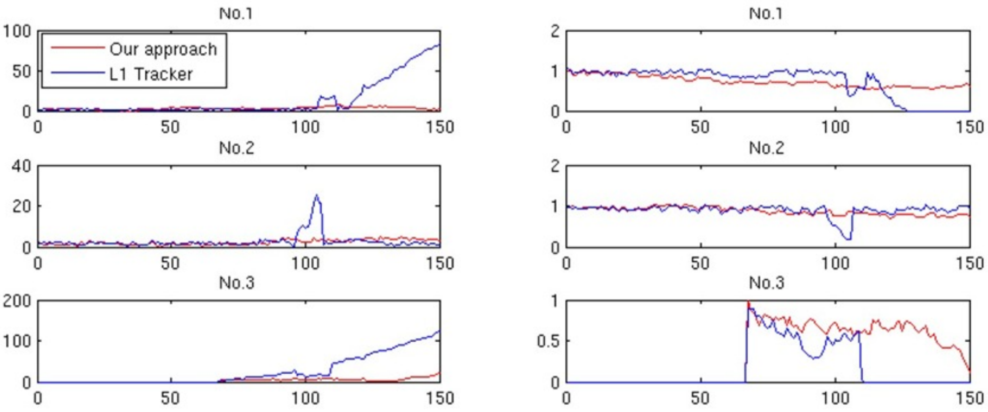

Figure 4. Quantitative results on EnterExitCrossingPaths1cor using our approach and L1 tracker. (Left) Position error (pixel) between tracking results and ground truth. x-coordinate: frames, y-coordinate: position error. (Right) Overlap of tracking results and ground truth. x-coordinate: frames, y-coordinate: overlap.

5. CONCLUSION

We have proposed a framework to find a global optimal spatiotemporal association with maximizes the consistency of motion and appearance of objects over time. To overcome the drastic appearance changes of objects, our method incorporates a *l*1 minimization-based appearance model into the Maximize A Posterior (MAP) inference within a particle filter framework. Compared to other multiple object tracking algorithms, the experimental results show that our proposed method outperforms both temporally (i.e., consistency of labels) and spatially (i.e. accuracy of outlined regions).

REFERENCES

[1] Perera, A.G.A., Srinivas, C., Hoogs, A., Brooksby, G., Hu, W., “Multi-object tracking through simultaneous long occlusions and split-merge conditions,” In CVPR (2006).

[2] Kim, Z., “Real time object tracking based on dynamic feature grouping with background subtraction,” In CVPR (2008).

[3] McKenna, S.J., Jabri, S., Duric, Z., Wechsler, H., “Tracking groups of people,” Computer Vision and Image Understanding, 80: 42-56 (2000).

[4] Javed, O., Shah, M., “Tracking and object classification for automated surveillance,” In ECCV (2002).

[5] Zhou, J.,Hoang, J., “Real time robust human detection and tracking system,” In CVPR (2005).

[6] Oliver, N.M., Rosario, B., Pentland, A.P., “A Bayesian computer vision system for modeling human interactions,” IEEE Transactions on Pattern Analysis, 28: 831-843 (2000).

[7] Orwell, J., Remagnino P.M., Jones G.A., “From connected components to object sequences,” IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2000).

[8] Robert, G.A., Williams, L.R., “Multiple target tracking with lazy background subtraction and connected components analysis,” Machine Vision and Application, 20: 93-101 (2009).

[9] Shao, J., Jia, Z., Li, Z., Liu, F., Zhao J., Peng, P., “A closed-loop background subtraction approach for multiple models based multiple objects tracking,” Journal of Multimedia, 6 (2011).

[10] Huang, L., Gu, W., Tian, Q., “Statistical modeling of complex backgrounds for foreground detection,” IEEE Transaction on Image Processing, 13**:** 1459-1472 (2004).

[11] Zhang, L., Li, Y., Nevatia, R., “Global data association for multi-object tracking using network flows,” In CVPR (2008).

[12] Pirsiavash, H., Ramanan, D., Fowlkes C.C., “Globally-optimal greedy algorithms for tracking a variable number of objects,” In CVPR (2011).

[13] Khan, Z., Balch, T., Dellaert, F., “MCMC-based particle filtering for tracking a variable number of interacting targets,” PAMI, 27 (2005).

[14] Yu, Q., Medioni, G., Cohen, I., “Multiple-target tracking by spatiotemporal monte carlo markov chain data association,” In CVPR (2007).

[15] Benfold, B., Reid, I., “Stable multi-target tracking in real-time surveillance video,” In CVPR (2011).

[16] Mei, X., Ling, H., “Robust visual tracking using *l*1 minimization,” In ICCV (2009).

[17] Liu, B., Yang, L., Huang, J., Meer P., Gong, L., Kulikowski, C., “Robust and fast collaborative tracking with two stage sparse optimization,” In ECCV (2010).

[18] Liu, B., Huang, J., Yang, L., Kulikowsk, C., “Robust tracking using local sparse appearance model and k-selection,” In CVPR (2011).

[19] Berclaz, J., Fleuret, F., Fua, P., “Robust people tracking with global trajectory optimization,” In CVPR (2006).

[20] Oh, S, Russell, S., Sastry, S,, “Markov chain monte carlo data association for general multiple-target tracking problems,” IEEE Conference on Decision and Control (2004).

[21] Candès, E.J., Li, X., Ma, Y., Wright, J., “Robust Principal Component Analysis?” Journal of the ACM (2011).

[22] Zhou, S.K., Chellappa, R., Moghaddam, B., “Visual tracking and recognition using appearance-adaptive models in particle filters,” IEEE Transactions on Image Processing, 11: 1491-1506 (2004).

[23] Tu, Z and Zhu, S., “Image segmentation by data driven markov chain monte carlo,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5): 657-673 (2002).