International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.5 (2014), pp.125-136
http://dx.doi.org/10.14257/ijsip.2014.7.5.11

Target Handoff with Appearance Model Inheriting and Learning

Wenhui Dong and Peishu Qu

College of Physics and Electronic engineering, Dezhou University, Dezhou 253023, China
dongwh_81@163.com, qupsh@163.com

Abstract

We address the issue of continuous tracking of the target in an environment covered by multiple cameras. In such a scenario, target handoff is a key problem. In this paper, we propose a novel target handoff method based on appearance model inheriting and learning. The appearance model is initially learned by sparse representation using the tracking results in the first camera. The next camera inherits the appearance model for target handoff and updates it after getting the whole tracking results. Then, the appearance model is transferred to the following camera. By the appearance model inheriting and learning, the appearance model can describe the target more and more precisely, which will make the target handoff more accurately and effectively. We also demonstrate the performance of our method on several video surveillance sequences.

Keywords: *Multi-camera tracking; Target handoff; Appearance model*

1. Introduction

As the scale and complexity increase in the surveillance, it is more difficult to track and monitor the target in a single camera with the required resolution and continuity. Thus, camera networks emerge and find extensive applications in visual surveillance. The goal is to locate targets, track their trajectories, and maintain their identities when they travel within or across cameras. The employment of multiple cameras not only improves coverage but also brings in more flexibility. Such a system usually consists of two main parts: 1) intra-camera tracking: tracking the target within a camera; 2) inter-camera tracking: tracking the target from one camera to another. Although there have been significant improvements in intra-camera tracking, inter-camera tracking is a less explored topic and many problems in it still need to be solved. Among the problems, target handoff is a crucial one. Target handoff is a process of transferring the target from one camera to next, which solves the target correspondence and identity problem among multiple observing cameras. To establish correspondence between objects, a robust appearance model is essential. However, obtaining a robust appearance is very challenging. The same object observed in different cameras undergoes significant variations of resolutions, lightings, poses and viewpoints. Besides, objects captured by surveillance cameras are often small in size and a lot of visual details such as facial components are indistinguishable in images, some of them look similar in appearance. Therefore, many authors want to find features that can be highly discriminative and robust to those inter-camera variations. In this paper, we propose a different idea. Instead of finding robust features for all cameras, a multi-mode appearance model of the target can be obtained by different cameras. The appearance model is initially a complete description of the tracked target in the first camera, which is learned based on sparse representation. The next camera inherits the appearance model for target handoff and updates it with the variations

ISSN: 2005-4254 IJSIP

after getting the whole tracking results. So do the following cameras. By the appearance model inheriting and learning, the appearance model will contain all the variations of the target in the different cameras and can describe the target more and more precisely, which will make the target handoff more accurately and effectively.

The reminder of this paper is organized as follows. We review the related work in Section 2. Section 3 describes the proposed target handoff with appearance model inheriting and learning. Section 4 contains the experimental results, which demonstrate the merits of the proposed target handoff method using several video surveillance sequences. Finally, Section 5 summarizes this work.

2. Related Work

Compared with single camera tracking, multi-camera tracking brings many new challenges. It is much less reliable to predict the spatio-temporal information of objects across different camera views than in the single camera view. There may be dramatic changes in the appearance of the target because of the variations of camera settings, viewpoints and lighting conditions in different camera views. So target handoff which solves the target correspondence and identity problem among multiple observing cameras becomes a key problem.

In general, target handoff methods could be divided into three main categories: geometry-based, appearance-based, and hybrid-based approaches. The geometry-based approach contains three sub-categories: location-based, alignment-based, and homograph-based approaches. In location-based approach [1, 2], the trace of the tracked object is projected back to the world coordinate system, and then the objects projected onto the same location are considered as the same target. It is usually assumed in these methods that the camera calibration or the topology of the cameras has already been solved before tracking. In alignment-based approach [3, 4], using the geometric transformation between cameras, the tracks of the same object are being aligned and recovered across different cameras. Take reference [3] as example, the tracking result of each camera is complete alignment both in time and space, so that the same object in different cameras will map to the same location. The homography-based approach [5-8] obtains position correspondences between overlapped views in the 2D image plane. Calderara, *et al.*, [7] warp the vertical axis of the object on the FOV of another camera to compute the amount of the match therein. This improves the capability in handling both the cases of single individuals and groups. In appearance-based approach [9-11], distinguishing features of the tracked objects are extracted and matched, generating correspondence among cameras. Appearance model is established by using the object feature captured in the different cameras which includes color, shape, texture, and so on. Researchers often choose one or combine some of these features to build the appearance model for matching. Color based appearance models often choose color histograms of the whole image regions as global features to match objects across camera views because they are robust to the variations of poses and viewpoints [12-14]. However, they also have the disadvantages that they are sensitive to the variations of lighting conditions and photometric settings of cameras. In reference [17], the brightness transfer function is computed between two cameras to handle this disadvantage. But different camera couples have different brightness transfer functions, which leads to the overhead computational cost. Shape based appearance model describes the target in the aspect of edges or gradient structures. Histogram of oriented gradients belongs to this kind of features. It computes the histograms of gradient orientations within cells which are placed on a dense grid and undergo local photometric normalization. It is robust to small translations and rotations of object parts. There are also other models [18, 19] proposed to characterize the geometric configuration of different local parts of objects. In texture based appearance model, Many local descriptors such as Scale-invariant feature transform(SIFT) [20], color SIFT [21], Local Binary Patterns (LBP) [22], Speeded Up Robust Feature (SURF) [23], Maximally Stable Extremal Regions (MSER) [24], have been proposed to characterize local texture and they can be applied to match the target between cameras. The hybrid-approach [25] is a combination of geometry and appearance-based methods.

Our target handoff method also belongs to the appearance-based method. Instead of finding robust features for all cameras, tracking results of all the cameras are used to establish the appearance model. Thus the model contains different modes of the target in different cameras. The appearance model is initially established by the tracking results in the first camera. The next camera inherits the appearance model for target handoff and updates it with the variations after getting the whole tracking results. Then, the appearance is inherited by the following camera. By the complementary of the following cameras, the appearance model will describe the target more and more precisely, which makes the target handoff more accurately and effectively.

3. Target Handoff with Appearance Model Inheriting and Learning

In this paper we achieve target handoff by appearance model inheriting and learning. Every camera uses the appearance model transferred from the former camera to initialize the tracker. The tracker searches the target in the first few frames and continues to track the target after the target being detected. Before the target leaves the FOV (Field of view) of the camera, all the tracking results in this camera are used to complement the appearance model for the variations. Finally, the updated appearance model is transferred again to the next camera. For the purposes of this paper, we assume that reasonably correct single-camera tracking results are available through whatever method is preferred by the user.

3.1. Appearance Model based on Sparse Representation

In this paper, we establish the model by sparse representation using gray level feature. The kernel of sparse representation is to represent the target on a over-complete dictionary and the target can be described by sparse linear combinations of the atoms in it. Suppose we have gotten N target templates in the first camera. Each target template is represented by gray level feature and the columns of the template are stacked to form a 1D vector. We seek the dictionary D that leads to the best possible representations for each member in this set with strict sparsity constraints. K-SVD method is used here to train the dictionary. The K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary, and an update process for the dictionary atoms to better fit the data [26].

Given the target template set $Y = \{y_i\}_{i=1}^{N}$, Our objective function is

$$\min_{D,X}\{\|Y - DX\|_2^2\} \quad subject \quad to\,\forall i, \|x_i\|_0 \le T_0 \tag{1}$$

Where, X is the sparse coefficient set and x_i is the element in it.

Sparse coding stage: This is a process that searches for sparse representations with coefficients summarized in X .The dictionary D is fixed in this stage and we random choose K target templates as the atoms of the dictionary. Then, the problem in (1) can be decoupled to

$$i = 1, 2, \cdots N, \min_{x_i}\{\left\| y_i - Dx_i \right\|_2^2\} \quad subject \quad to \forall i, \left\| x_i \right\|_0 \le T_0 \tag{2}$$

The system in (2) is underdetermined and does not have a unique solution for every x_i. We solves the problem as an l1-regularized least squares problem, which has been known to typically yield sparse solutions [27].

$$\min \left\| y_i - Dx_i \right\|_2^2 + \lambda \left\| x_i \right\|_1 \tag{3}$$

Where, $\| \ \|_2$ and $\| \ \|_1$ are the l_2 norm and l_1 norm respectively.

The minimization task in (3) can be solved using the preconditioned conjugate gradients (PCG) algorithm in reference [28].

Update process for the dictionary atoms: This stage is performed to search for a better dictionary. This process updates one column at a time, except the column d_k ,all other columns in D are fixed. The purpose is to find a new column $\tilde{d}_k$ and new values for its coefficients that best reduce the MSE.

In order to update the columns of the dictionary, we should rewrite the object function in (1) as following:

$$\left\| Y - DX \right\|_2^2 = \left\| (Y - \sum_{j \ne k}^{K} d_j x_T^j) - d_k x_T^k \right\|_2^2 = \left\| E_k - d_k x_T^k \right\|_2^2 \tag{4}$$

Where, x_T^k denotes the k-th row of X and E_k is the error for all the N templates when the k-th column is removed.

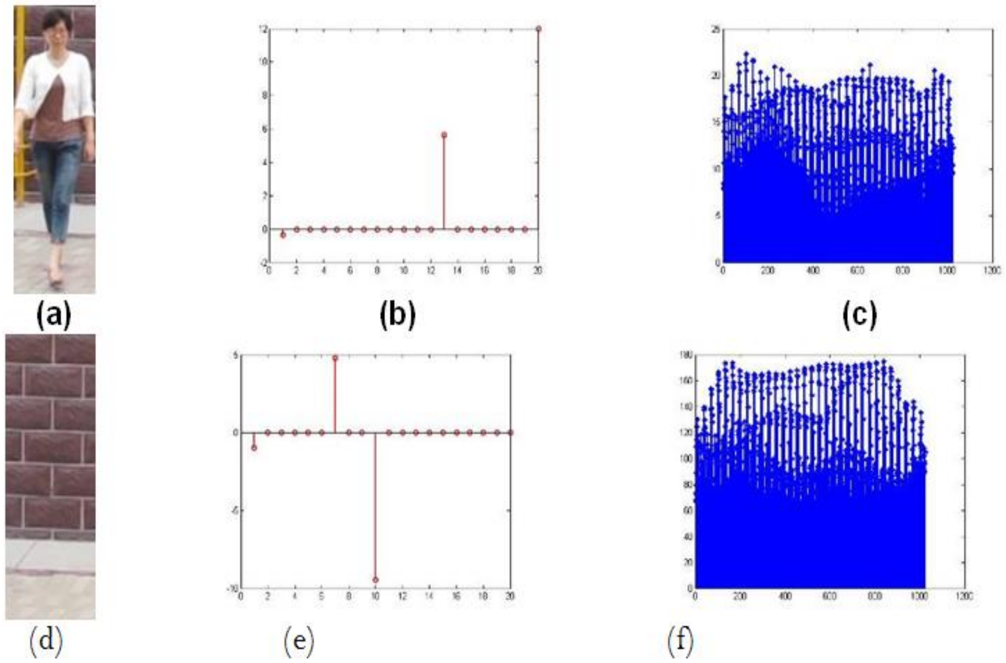
In order to find alternative $\tilde{d}_k$ and x_T^k under the sparsity constraint, define ω_i as the indices pointing to examples $\{y_i\}$ that use the column d_k ,namely, $\omega_k = \{i \mid 1 \le i \le K, x_T^k(i) \ne 0\}$ .Then a matrix Ω_k can be constructed with ones on the $(\omega_k(i), i)$ elements and zeros elsewhere. By multiplying the matrix, (4) can be rewritten as

$$\left\| E_k \Omega_k - d_k x_T^k \Omega_k \right\|_2^2 = \left\| E_k^R - d_k x_R^k \right\|_2^2 \tag{5}$$

Then, we can use SVD to find $\tilde{d}_k$ as the first column of U and x_T^k as the first column of V multiplied by $\Delta(1,1)$.

$$E_k^R = U \Delta V^T \tag{6}$$

Before getting the solution, it is necessary that the columns of D remain normalized and the support of all representations either stays the same or gets smaller by possible nulling of the terms.



(a) (b) (c)

(d) (e) (f)

Figure 1. Discriminative Character of the Dictionary, (a) Target Image, (b) Sparse Coefficients of the Target, (c) Reconstruction error of the Target Pixels, (d) Background Image,(e) Sparse Coefficients of the Background, (f) Reconstruction error of the Background

The two processes will be continued until convergence, and then a discriminative dictionary D is gotten. If the target is represented on D, the reconstruction error of each pixel will be very small. If other thing is represented on it, the reconstruction error will be much larger. Figure 1 shows the discriminative character of the dictionary. Both the target and the non target are represented on the dictionary trained by the target templates. As can be seen in the figure, the reconstruction error of each pixel in the target is below 25 while that of the background is much larger.

3.2 Appearance Model Inheriting for Target Handoff

Suppose we have known the handoff camera, the dictionary D is transferred to the next camera before the target is being out of the view. Then, the target handoff is equivalent to the target detection in the first few frames in the next camera. We search the target exhaustively in the frame and extract target candidates. Each candidate is sparsely represented on the dictionary D. Then, it is reconstructed again by the dictionary D and the sparse coefficients. A confidence map can be obtained based on the reconstruction error using (7) for all the candidates and if the maximum value is above the threshold τ, then the corresponding candidate is considered as the target. The whole scheme is shown in Figure 2.

$$S(c) = e^{-\|c-Dx\|_2^2} \tag{7}$$

Where, c is the candidate.

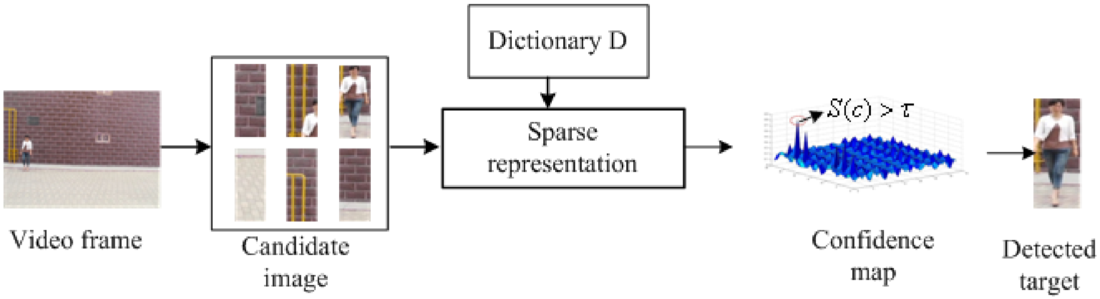
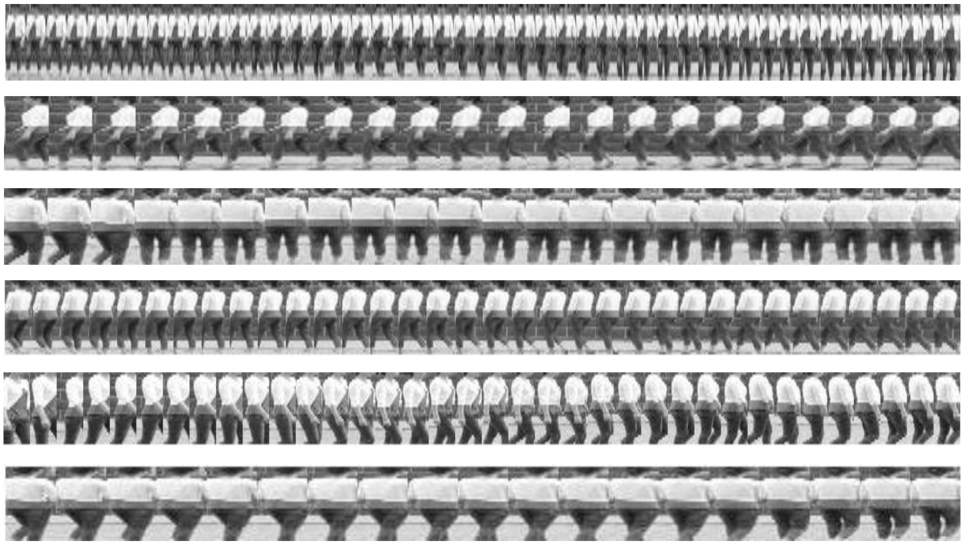
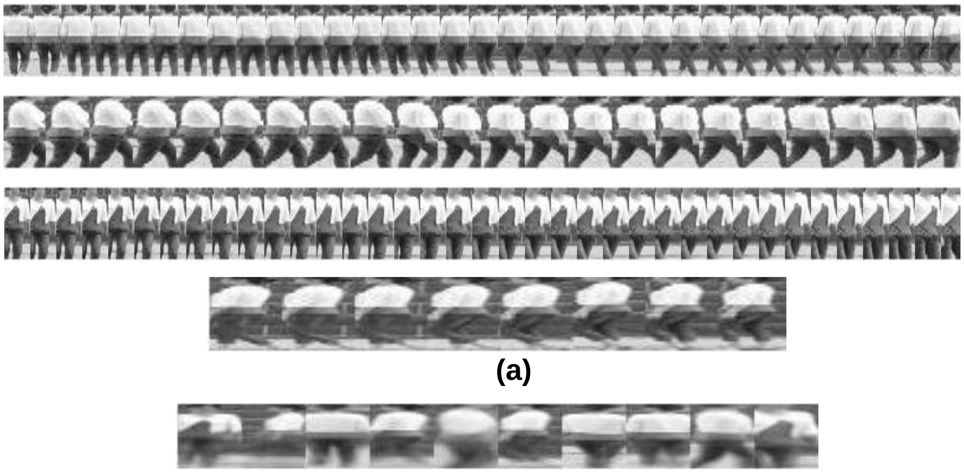


Figure 2. Target Handoff by Appearance Model Inheriting, The Next Camera Inheriting the Dictionary and the Target Handoff is Equivalent to a Target Detection based on the Sparse Representation on the Dictionary

3.3 Appearance Model Learning

The appearance of the target varies from one camera to another. So we need update it in every camera after the target handoff process. When the tracking has been done, we save all the tracking results in this camera. If we only use the tracking results in the current camera to train the dictionary D using K-SVD method, the dictionary will change to adapt to the variations of the target in the current camera but it will forget the case in the former camera. So the tracking results in both former and current camera should all be used to train the dictionary D. However, it will be too time cost to transfer all the tracking results in the former camera to current one. Intuitively, only the representative results need to be transferred. So in this paper, we cluster the tracking results in the former camera and only the centers of the clustering are transferred to the current camera. Any clustering method can be used here and we use the ISODATA algorithm to achieve this clustering process. As an example, we show the clustering result for 290 tracking results and the centers in Figure 3.In these 290 frames the target walks along a circle. Then all the centers of the clustering of the tracking results in former camera and the tracking results in current camera are used as templates to train the dictionary using K-SVD method. We summarize the appearance modeling process in Figure 4. The updated dictionary will be transferred again to the next handoff camera for target handoff.





(a)

(b)

Figure 3. Tracking Results Clustering, (a) Ten Clustering of 290 Tracking Results, (b) Centers of the Clustering

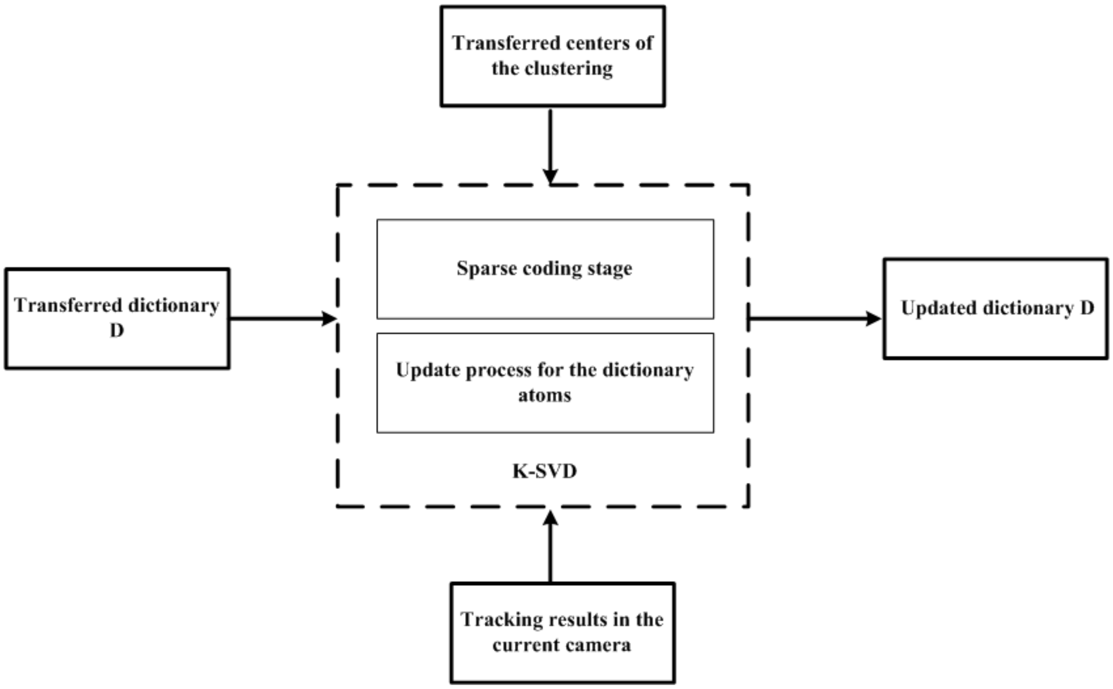

Transferred centers of the clustering
Transferred dictionary D
Sparse coding stage
Update process for the dictionary atoms
K-SVD
Updated dictionary D
Tracking results in the current camera

Figure 4. Appearance Model Learning for Variations: after Learning, the Updated Dictionary D can Represent all the Target Variations in all Former Cameras and the Current Camera

4. Experimental Results

4.1 Target Handoff between Two Cameras

In this experiment, we evaluate our method in a complex environment with two cameras. The two cameras are set in our building. Camera 0 can monitor the view around the door of the building and Camera 1 monitors the hall inside of the building. We show some surveillance scenes and the relative locations of the two cameras in Figure 5. The green circles denote the two cameras. As can be seen, there is an overlapped FOV between the two cameras (Scene 4). In this experiment, a man walks along the red line and enters the building. Camera 0 tracks the man singly about 400 frames. Then the man enters scene 4 and target handoff is trigged. Using the appearance model transferred from Camera 0, the Camera 1 detected the target in frame 407. Target handoff is achieved. Figure 6 demonstrates some examples in the whole process.

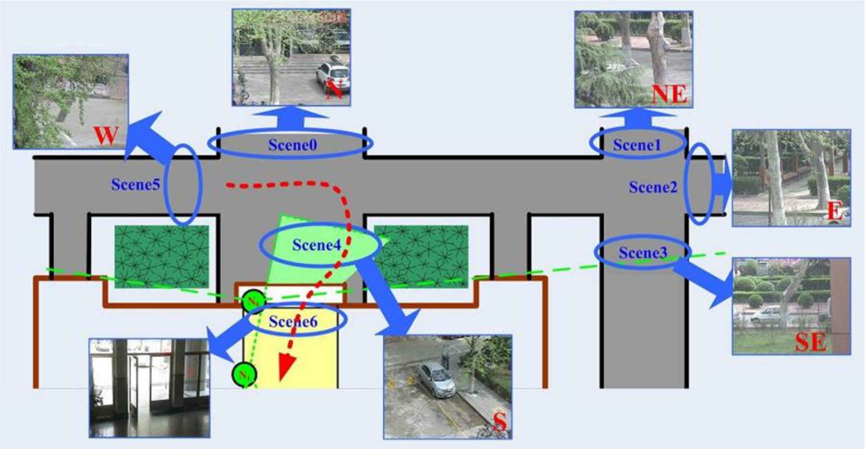
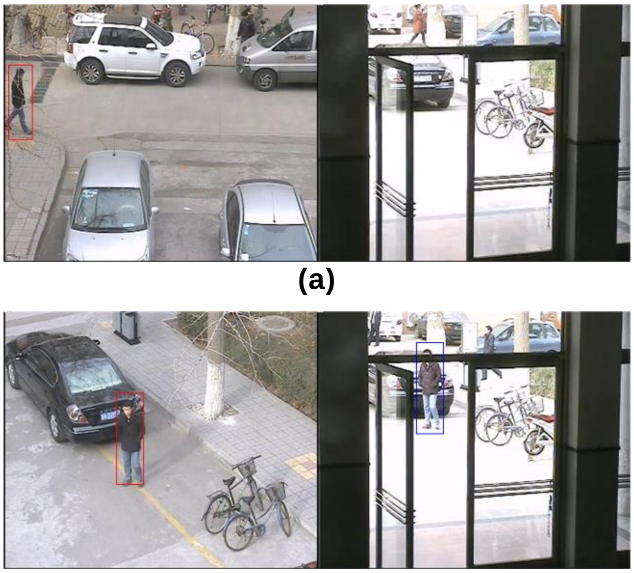
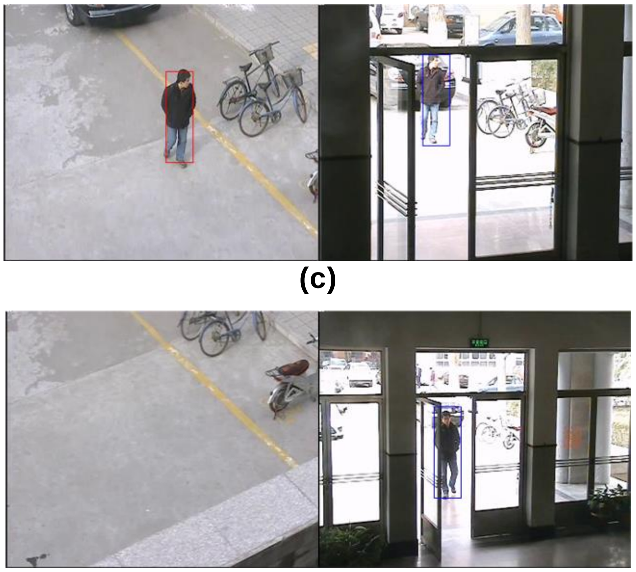


Figure 5. Scenes and the Position Relationship between the Two Cameras around the Building



(a)

(b)



(c)

(d)

Figure 6. Target Handoff between the Two Cameras, (a) Frame 3, (b) Frame 407, (c) Frame 428, (d) Frame 523

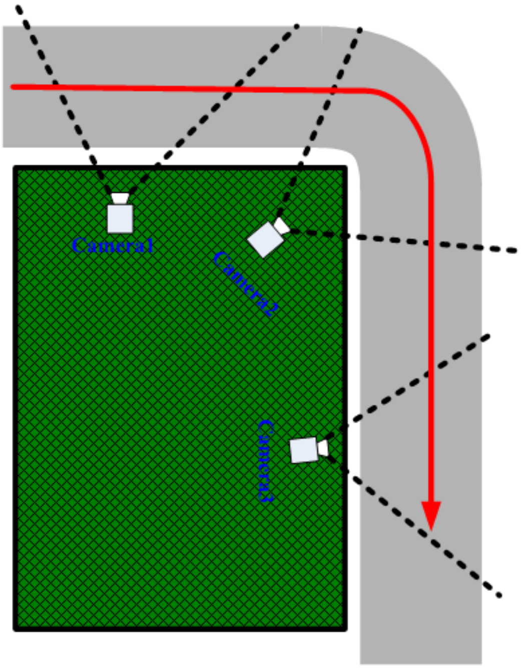
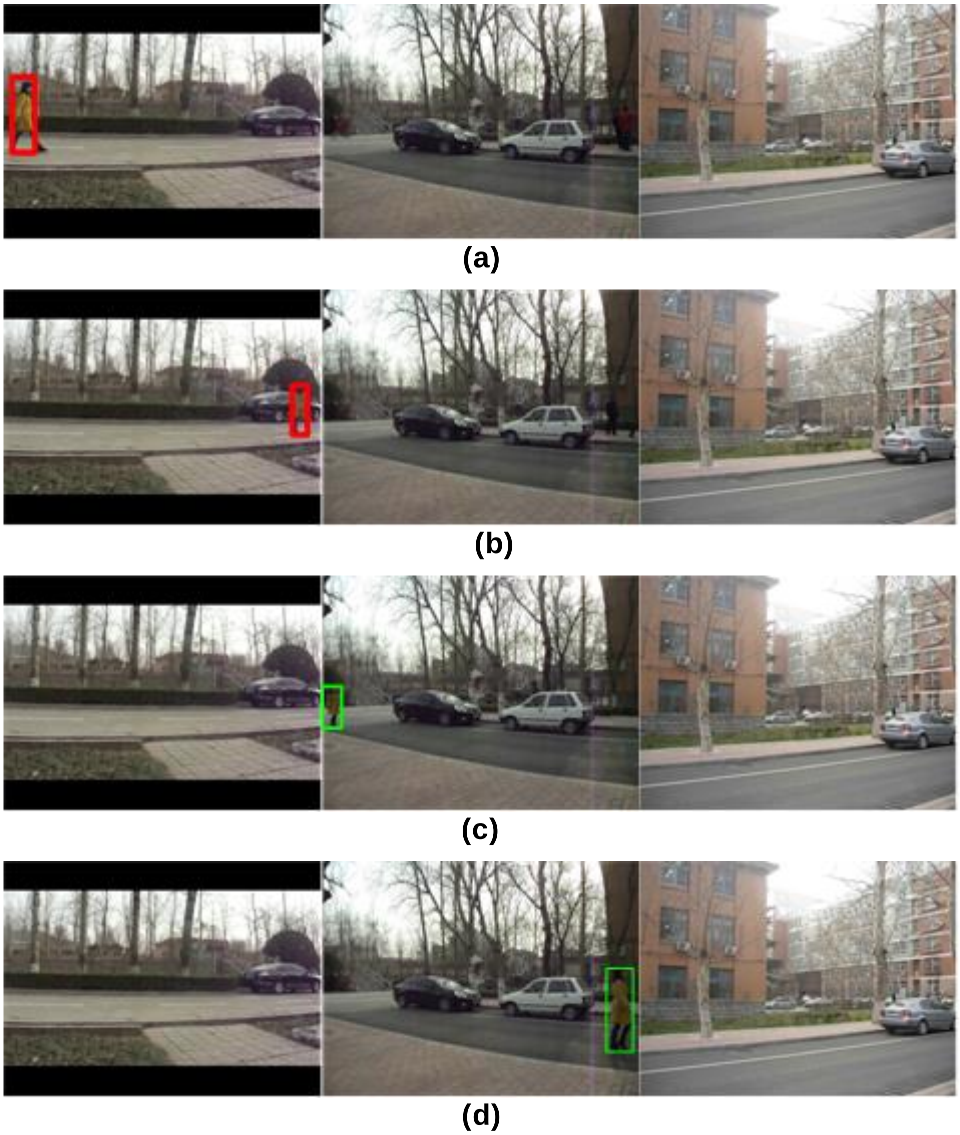


Figure 7. Scenes and the Position Relationships for Three Cameras

4.2 Target Handoff in Three Cameras

In this experiment, we test our method in a network that has three cameras. Figure 7 shows the scenes and the position relationships for the three cameras. The cameras are set at the same side of the road and the FOV of them are non-overlapped. Camera 1 is under a tree and the light is dark. While light in the Camera 2 and Camera 3 are much brighter. In our experiment, a woman walks along the road and crosses the three cameras from Camera 1 to Camera 3. When the woman is going to be out of the view of Camera 1, the dictionary trained by the tracking results and the centers of the clustering in Camera 1 are transferred to Camera 2 for target handoff. After target handoff, Camera 2 begins to track the woman and all the tracking results are saved. Then, all the centers of the clustering of the tracking results from Camera 1 and the tracking results in Camera 2 are used as templates to update the dictionary. Finally, the updated dictionary and the centers of the clustering in Camera 2 are transferred to Camera 3 for target handoff. As can be seen in Figure 8, the target handoff is achieved accurately in Camera 2 and Camera 3.

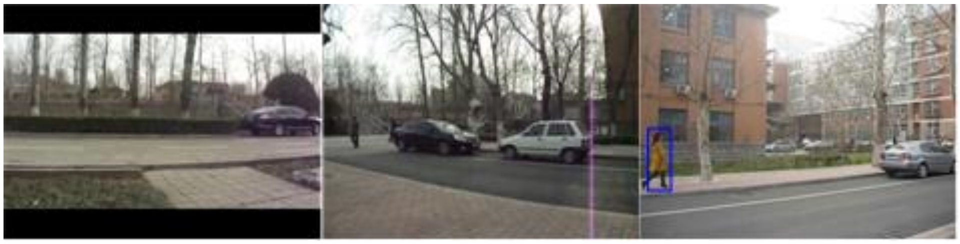


(a)

(b)

(c)

(d)



(e)

Figure 8. Target Handoff among the Three Cameras: (a) Frame 79,(b)Frame 180,(c)Frame 195,(d)Frame 300 (e) Frame 321

5. Conclusion

In this paper, we propose a novel idea for target handoff which is based on appearance model inheriting and learning. The appearance model is initialized by tracked target in the first camera using sparse representation. The next camera inherits the appearance model for target handoff and updates it with the variations after getting the whole tracking results. So do the following cameras. By the appearance model inheriting and learning, the appearance model will contains all the variations of the target in the different cameras and can describe the target more and more precisely, which will make the target handoff more accurately and effectively.

References

[1] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, S. Chatterjee and R. Jain, “An architecture for multiple perspective interactive video”, Proceedings of the third ACM international conference on Multimedia, **(1995)** May 10-12, New York, USA.

[2] J. Black and T. Ellis, “Multiple camera image tracking”, Image and Vision Computing, vol. 24, no.11, **(2006)**, pp.1256-1267.

[3] S. Khan and M. Shah, “Consistent labeling of tracked objects in multiple cameras with overlapping fields of view”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, **(2003)**, pp. 1355-1361.

[4] F. Fluuret, J. Berclaz, R. Lengagne and P. Fua, “Multi camera people tracking with a probabilistic occupancy map”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, **(2008)**, pp. 267-273.

[5] L. Lee, R. Romano and G. Stein, “Monitoring activities from multiple video streams: establishing a common coordinate frame”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, **(2000)**, pp. 758-767.

[6] S. Calderara, A. Prati, R. Vezzani and R. Cucchiara, “Consistent labeling for multi camera object tracking”, 13th International Conference, **(2005)** September 6-8, Cagliari, Italy.

[7] S. Calderara, R. Cucchiara and A. Prati, “Bayesian-competitive consistent labeling for people surveillance”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, **(2008)**, pp. 354-360.

[8] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou and S. Maybank, “ Principal axis-based correspondence between multiple cameras for people tracking”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, **(2006)**, pp. 663-671.

[9] M. Balcells, D. DeMenthon and D. Doermann, “An appearance-based approach for consistent labeling of humans and objects in video”, Pattern and Application, vol. 7, no. 4, **(2005)**, pp. 373-385.

[10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, vol. 60, no. 2, **(2004)**, pp. 91-110.

[11] J.-Y. Choi, J.-W. Choi and Y.-K. Yang, “Improved tracking of multiple vehicles using invariant feature-based matching”, Pattern Recognition and Machine Intelligence, vol. 4815, **(2007)**, pp. 649-656.

[12] J. Orwell, P. Remagnino and G. A. Jones, “Multiple camera color tracking”, IEEE Workshop on Visual Surveillance, **(1999)** June 26-28.

[13] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale and S. Shafer, “Multi-camera multi-person tracking for easyliving”, IEEE Workshop on Visual Surveillance, **(2000)**, July 1-2, Dublin, Ireland.

[14] A. Mittal and L. S. Davis, “M2tracker: A multi-view approach to segmenting and tracking people in a

cluttered scene", International Journal of Computer Vision, vol. 51, **(2003)**, pp. 189–203.

[15] B. Prosser, S. Gong and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer function", British Machine Vision Conference, **(2008)** September 1-4, Leeds, UK.

[16] A. Agarwal and B. Triggs, "Hyper features-multilevel local coding for visual recognition", 9th European Conference on Computer Vision, **(2006)** May 7-13, Graz, Austria.

[17] G. Carneiro and D. Lowe, "Sparse flexible models of local features", European Conference on Computer Vision, **(2006)** May 7-13, Graz, Austria.

[18] D. Lowe, "Distinctive image features from scale-invariant key points", International Journal of Computer Vision, vol. 60, no. 2, **(2004)**, pp. 91–110.

[19] A. E. Abdel-Hakim and A. A. Farag, "Csift: A sift descriptor with color invariant characteristics", European Conference on Computer Vision, **(2006)** May 7-13, Graz, Austria.

[20] T. Ojala, M. Pietik¨ainen and T. M¨aenp¨a¨a, "Multi resolution gray-scale and rotation invariant texture classification with local binary patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, **(2002)**, pp. 971–987.

[21] H. Bay, T. Tuytelaars and L. V. Gool, "Surf: Speed up robust features", European Conference on Computer Vision, **(2006)** May 7-13, Graz, Austria.

[22] P. E. Forssen, "Maximally stable color regions for recognition and matching", IEEE conference on Computer Vision and Pattern Recognition, **(2007)** June 17-22, Minnesota, USA.

[23] O. Tuzel, F. Porikli and P. Meer, "Region covariance: A fast descriptor for detection and classification", European Conference on Computer Vision, **(2006)** May 7-13, Graz, Austria.

[24] Z. Qiang and L. Baoxin, "Discriminative K-SVD for dictionary learning in face recognition", 2010 IEEE Conference on Computer Vision and Pattern Recognition, **(2010)** June 13-18, San Francisco, USA.

[25] J. Kang, I. Cohen and G. Medioni, "Continuous tracking within and across camera Streams", IEEE International Conference on Computer Vision and Pattern Recognition, **(2003)** June 16-22, Wiscosin, USA.

[26] S.-J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky. "A method for large-scale l1-regularized least squares", IEEE Journal on Selected Topics in Signal Processing, vol. 1, no. 4, **(2007)**, pp. 606-617.

Authors

Wenhui Dong, she received her B.S. degree in electronic engineering from Qufu Normal University in 2003 and M.S. degree in communication and information systems from Shandong University in 2006.Now she is a lecturer of College of Physics and Electronic engineering, Dezhou University. She focuses on computer vision, pattern recognition and image processing.

Peishu Qu, he received his B.S. degree in electronic engineering from Qufu Normal University in 2003 and M.S. degree in communication and information systems from Civil Aviation University of China in 2009. He focuses on computer vision, pattern recognition and image processing.