Multi Camera-Based Person Tracking Using Region Covariance and Homography Constraint

Bogdan Kwolek
Rzeszów University of Technology
35-959 Rzeszów, Poland
bkwolek@prz.edu.pl

Abstract

In this paper, an algorithm for multiple camera based person tracking is presented. Region covariance matrixes are used to model the target appearance. The correspondence between multiple camera views is established via homography. It is utilized to improve the tracking of people under assumption that they are at the common ground plane. If there is occlusion in one view, the homography to this view from another view is utilized to locate the object template. The information about the true location of the template helps the tracker to resume, even in case of substantial temporal occlusions or large object movements. The object template is represented by multiple non-overlapping patches. Owing to such an object representation the tracker is capable both detecting the occlusion and handling considerable partial occlusions. The object tracking is achieved using particle swarm optimization. The objective function is based on the Log-Euclidean Riemannian metric. Experimental results that were obtained on surveillance videos show the feasibility of the presented approach.

1. Introduction

People tracking is an important problem for surveillance applications. The goal of the tracking is to automatically find the same target in an adjacent frames from an image sequence once it is initialized. The challenge is to track the object irrespective of scale, rotation, perspective projection, occlusions, changes of appearance and illumination. Several methods were proposed to achieve object tracking [1][7], including surveillance applications. However, most current surveillance systems still treat multiple cameras as a set of single cameras. They are typically utilized to extend the viewing area. Therefore, reliable tracking of people in multiple cameras is very desirable capability, particularly for surveillance systems. This is because by using multiple cameras we can handle occlusions better. However, the use of multiple cameras is connected with difficulties such as camera calibration and synchronization, as well as correspondence. Correspondence between multiple cameras involves establishing instant coherence between objects in different views. It is one of the most important and difficult problems in visual surveillance based on multiple cameras.

Some methods were proposed to achieve people tracking using multiple views. In [9] the homography is used to align the foreground of the ground plane in images that are acquired by cameras with overlapping fields of view. However, the results obtained by this method might be inaccurate since the feet are typically segmented erroneously due to small sizes as well as possible occlusions. Moreover, this method tends to segment a person into several parts and in consequence it often leads to large number of false positive locations. The discussed method has been extended in [10] [2] to planes at multiple heights. The major drawback of such an approach is large number of so called false positive candidates. Therefore, in [3] both 2D and 3D information is used to reduce the amount of false positive candidates. The discussed above methods operate on foreground extracted via a background subtraction. Thus, the results strongly depend on the quality of object detection. Hence, they might be unsatisfactory in varying illumination conditions, in case of shadows around the feet, etc. In [5] a homography based method for tracking people in dense crowd via a multiple camera system has been proposed. It uses multiple height homographies for extraction of the head top and assumes that the scene is observed by a set of overlooking cameras. The intensity correlation on the projected patches is used to detect the candidate blobs. However, due to well known reduced discriminative capability of the intensity correlation, which does not consider both spatial and statistical properties of the object, the results might be not good enough.

In this paper we present an algorithm for tracking people in multiple views. It is based on the region covariance [16], which describes both spatial and statistical properties of the objects. The correspondence between multiple views is established via homography that is estimated using pairs of corresponding landmarks on the ground plane. It is then

utilized to improve the tracking of people under assumption that they are at the common ground plane. If there is occlusion in one view, the homography from another view to this view is utilized to locate the template of the target. The information about the true location of the template helps the tracker to resume, even in case of substantial temporal occlusions or large object movements. The object template is represented by multiple non-overlapping patches. Owing to such an object representation the tracker is able both detecting the occlusion and handling considerable partial occlusions. The object tracking is achieved using particle swarm optimization [8]. The objective function is based on the recently proposed Log-Euclidean Riemannian metric [4].

The paper is organized as follows. The next section is devoted to region covariance based people tracking. At the beginning, we overview the particle swarm optimization. Afterwards, we discuss how the covariance can be used to represent image regions. Then we present the Log-Euclidean Riemannian metric to compute the similarity between region covariance matrixes. The last part of the section is dedicated to presentation of tracking results, which were obtained using single view. Section 3 is devoted to multi camera based people tracking. In section 4 we present the experimental results that were obtained using multiple views. Finally, we conclude the paper in section 5.

2. Region covariance based people tracking

One way to achieve object tracking is searching for the best match of the predefined object model in the image. In the simplest solution the object tracking can be accomplished via the deterministic searching of window location whose content best matches the content of a reference window. Particle Swarm Optimization (PSO) [8], which is a population based stochastic optimization technique, allows us to avoid such time consuming exhaustive searching for the best match. Region covariance [16] is a robust descriptor for object detection and classification. Region covariance based statistics of an image was utilized in [13] to achieve reliable object tracking. In this section we overview the PSO and show how this algorithm can be utilized to realize object tracking. The section explains also how region covariance matrix can be employed to accomplish reliable tracking of objects which undergo considerable occlusions. Afterwards, we demonstrate how the PSO built on the region covariance matrix copes with temporal occlusions of an object being tracked.

2.1. Particle Swarm Optimization

PSO is a population based algorithm that exploits a set of particles representing a potential solutions of the optimization task [8]. This technique differs from other evolutionary techniques by inclusion of particle velocity. The particles fly through the n-dimensional problem space with a velocity subject to both stochastic and deterministic update rules. They undergo evaluation according to some fitness function after each time step. In the course of the optimization the particles iteratively evaluate their candidate solutions and remember the coordinates of their best location with the smallest objective value so far, making this information available to their neighbors. Particles communicate good positions to each other and adjust their own velocity using such good positions. Additionally each particle employs a best value, which can be:

- a global best, which is immediately updated when a new best position is found by any particle in the swarm
- neighborhood best, where only a specific number of particles is affected if a new best position is found by any particle in the sub-population

The topology with the global best converges faster as all particles are attracted simultaneously to the best part of the search space. The neighborhood best allows parallel exploration of the search space and decreases the susceptibility of falling into local minima. However, it slows down the convergence speed.

In the ordinary PSO algorithm the update of the particle velocity is realized in accordance with the following equation:

$$v_j^{(i)} \leftarrow w v_j^{(i)} + c_1 r_{1,j}(p_j^{(i)} - x_j^{(i)}) + c_2 r_{2,j}(p_{\mathrm{g},j} - x_j^{(i)}) \quad (1)$$

where $v_j^{(i)}$ is the velocity in the $j-$th dimension of the $i-$th particle, w is the positive inertia weight, c_1, c_2 denote the acceleration coefficients, $r_{1,j}$ and $r_{2,j}$ are uniquely generated random numbers with the uniform distribution in the interval $[0.0,\ 1.0]$, $p^{(i)}$ is the best position that the particle i has found, p_{g} denotes best position that is found by any particle in the swarm. The new position of a particle is calculated in the following manner:

$$x_j^{(i)} \leftarrow x_j^{(i)} + v_j^{(i)} \quad (2)$$

The local best position of each particle is updated according to the following formula:

$$p^{(i)} \leftarrow \begin{cases} x^{(i)}, & \text{if } f(x^{(i)}) < f(p^{(i)}) \\ p^{(i)}, & \text{otherwise} \end{cases} \quad (3)$$

and the global best position p_{g} is defined as follows:

$$p_{\mathrm{g}} \leftarrow \arg\min_{p^{(i)}} \{f(p^{(i)})\} \quad (4)$$

The value of velocity $v^{(i)}$ should be restricted to the range $[-v_{max}, v_{max}]$ to prevent particles from moving out of the search space. In the evaluation phase the fitness value of each particle is determined by a predefined observation model according to the following formula:

$$f(x_t^{(i)}) = p(z_t^{(i)} | x_t^{(i)}) \quad (5)$$

where $z_t^{(i)}$ is the observation corresponding to $x_t^{(i)}$.

2.2. Covariance as a region descriptor

In our approach we utilize the region covariance matrix (RC) to represent the object template. For every pixel i of the $M \times N$ template we calculate a feature vector b_i

$$b_i = (x \;\; y \;\; R \;\; G \;\; B \;\; I_x \;\; I_y)^T \tag{6}$$

where x, y represent the Cartesian coordinates of pixel i, R, G, B stands for color components, and I_x, I_y are image derivatives. The region covariance descriptor is given by:

$$C = \frac{1}{MN-1} \sum_{i=1}^{MN} (b_i - \bar{b})(b_i - \bar{b})^T \tag{7}$$

where $\bar{b}$ denotes the vector of means of corresponding features for the pixels in the template. Such a region descriptor can be computed fast using integral images [16]. The region covariance descriptor has many advantages. In particular, RC indicates both spatial and statistical properties of the objects, it allows to combine multiple modalities and features, and last but not least, it is capable of relating regions of different sizes. This descriptor is also robust to the variations in illumination conditions, pose and view. Although the covariance matrixes are positive semi-definite in general, in practice they should be regularized by adding a small constant multiple of the identity matrix, making them strictly positive.

2.3. Similarity measure

Recently, a novel Log-Euclidean Riemannian metric [4] has been proposed to obtain statistics on symmetric positive definite matrixes. Under such a metric the distances and Riemannian means assume an easier form in contrast to widely used affine-invariant Riemannian metric [12].

The Singular Value Decomposition (SVD) of symmetric matrix A of size $n \times n$ is $U\Sigma U^T$, where U is an orthonormal matrix, and $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal matrix with nonnegative eigenvalues. The matrix exponential $\exp(A)$ of symmetric matrix is given by: $\exp(A) = U \cdot \text{diag}(\exp(\lambda_1, \ldots, \exp(\lambda_n)) \cdot U^T$, conversely, the matrix logarithm of symmetric positive definite matrix is given by: $\log(A) = U \cdot \text{diag}(\log(\lambda_1, \ldots, \log(\lambda_n)) \cdot U^T$. Each symmetric matrix is associated to a tensor by the exponential, conversely, a tensor has a unique symmetric matrix logarithm. The distance between two symmetric positive definite matrixes X and Y under the Log-Euclidean Riemannian metric is calculated according to:

$$dist(X, Y) = \| \log(X) - \log(Y) \|_2 \tag{8}$$

The Riemannian mean of several elements is an arithmetic mean of matrix elements. Using the Log-Euclidean metric we can directly employ the algorithm [14] for the incremental subspace update.

2.4. Object tracking with occlusion handling

In the algorithm utilized in this work, the particle score (5) is evaluated using the distance (8). Given the estimate of the object state in the previous time, when a new image is available, the particles are drawn from a Gaussian distribution in order to cover the promising object locations. The PSO is employed afterwards in order to concentrate the particles near the true state of the object. The optimization aims at shifting the particles towards more promising regions in the search area.

In order to evaluate the tracking performance of the PSO built on the region covariance matrix we conducted experiments on PETS2009 data sets. Here we follow a similar idea for multi-region covariance, where the object template is represented by multiple non-overlapping patches [11]. Owing to robust combining of such patch votes the object tracker is able to handle considerable partial occlusions. As we can observe in Fig. 1, thanks to multi-patch representation of the object, the template does not undergo shifting during the occlusion.

Figure 1. Tracking a person under occlusion. Frames #204, 209, 210, 211, 212, 213, 214, 217

Similar behavior can be observed in Fig. 2, where the template also keeps well the location despite considerable occlusion. As we can see, the object being tracked undergoes several occlusions, see frames #150 and #152. Moreover, the colors of the objects, i.e. the color of the jacket and the signboard, are quite alike.

Figure 2. Tracking a person under considerable occlusion. Frames #145, 146, 148, 149, 150, 151, 152, 153, 154, 155, 156, 160

Figure 3 illustrates the behavior of the algorithm in the same image sequence, but with slightly different initial location of the template. As we can notice in the discussed images, despite small drift of the template, the tracker temporally fails and then recovers quickly.

Figure 3. Tracking a person in sequence of images from Fig. 2 using different initial location of the template. Frames #149, 150, 151, 152, 153, 154, 155, 156

Figure 4 depicts some experimental results of person tracking in another camera view. The quality of these images is poorer than the quality of images from Fig. 1. Besides, the upper body of the pedestrian is severely occluded. Despite this the tracker follows the target as well as keeps precisely the location of the template.

Figure 4. Person tracking in another camera's view. Frames #222, 231, 240, 241, 242, 243, 244, 245, 246, 247, 251, 261

The results that are presented above were obtained using identical settings. The number of particles was equal to 50, whereas the maximal number of iterations was set to 5. The views of the scene are shown in Fig. 5. The above results indicated the great tracking performance of the PSO built on multi-patch covariances. However, as expected, in some situations, single view based tracking might lose the target due to considerable occlusion, even if a tracker is built on robust image statistics like region covariance.

3. Multi camera based people tracking

The results obtained in the previous section acknowledged that in some circumstances the single camera methods

Figure 5. Input images in view #1 and #3

might be insufficient for handling the tracking of people under occlusion or for handling dense crowds. Occlusion and lack of visibility in crowded scenes are the central difficulties in tracking individual people correctly and consistently. Most current surveillance systems still treat multiple cameras as a set of single cameras. To take advantage of synergy of multiple cameras and to achieve advantageous cooperation, it is necessary to establish correspondence between the different views. Finding such correspondence is one of the most important problems in the visual surveillance.

Several recent methods take advantage of targets, which move on a common plane and which are observed by cameras with overlapping fields of view. In [9] it was developed a planar homography constraint to determine occlusions and robustly establish locations on the ground plane, corresponding to the feet of the people. In order to find tracks the algorithm extracts feet regions over a window of frames and stacks them creating a space time volume. In [10][2] the same idea is extended for multiple parallel planes to obtain 3D volume of the target. In such a framework we can distinguish the following stages: (i) foreground detection to get moving targets, (ii) the use of the homography constraint to project the targets to a common plane, (iii) processing the projected data to find the correspondence. Figure 6 illustrates the mentioned above stages. The depicted results were obtained on PETS2009 datasets, see Fig. 5. The homographic transformation of the pedestrian into the ground plane was done through the use of accompanying calibration data. The cameras were calibrated using a method that was proposed in [15].

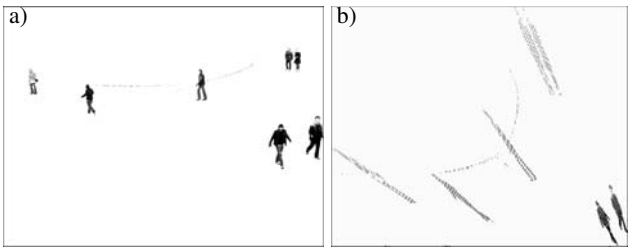

Figure 6. The foreground image of Fig. 5 (a), homographic transformation of the foreground into ground plane (b)

In order to determine the homography between cameras,

the different views should share a common ground plane. Let x_i, y_i and x_i', y_i' be a pair of corresponding points on the ground plane in the two views. The homography H is given by the following equation [6]:

$$\begin{bmatrix} x_i' \\ y_i' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \quad (9)$$

Although H can be determined from at least 4 pairs of corresponding points on the plane, the more points is used, the more reliable H will be.

Figure 7 illustrates the correspondence established via the ground plane homography. The homography has been estimated on the basis of ten pairs of points lying on the ground plane, which have been selected manually. As can be seen the homography induced by be the planar surfaces permits precise warping of images taken by different cameras.

Figure 7. Homography between different camera views

The shadows that are cast by persons, as depicted at Fig. 6b, can by utilized to establish a correspondence between different camera views. The fusion of such shadows, which are projected from different views amounts to carrying out the visual hull intersection on the ground plane. But due to perspective distortions the process of establishing the correspondence in 3D might be not an easy task. Moreover, the method often leads to considerable number of false positives. What is more, in real surveillance scenarios, the points of the feet may not be correctly detected or even they may be undetectable due to occlusions. In addition, in real conditions the feature points of a person in different views do not always correspond to the same physical point.

Taking into account the above shortcomings, our method does not employ the background subtraction as the primary modality for object extraction, but instead it relies more strongly on tracking correspondences. The correspondence between multiple views is established via homography. Such correspondence is used to improve the tracking of people under assumption that they are at the common ground plane. If there is occlusion in one view, the homography to this view from another view is used to locate the target template. The information about the template location helps the tracker to resume, even in case of substantial temporal occlusions or large object movements. Even if the target is partially or fully occluded by another object, the tracker still follows the target as long as it is visible in another view. Assuming that occlusions produce large image differences, the multiple patch based object representation allows the tracker to detect occlusion as well as allows to resume the tracking. On the other side, through such reliable object representation the tracker can cope with considerable occlusions.

4. Experiments

Experiments were conducted on the PETS2009 test sequences. Figure 8 shows some tracking results that were obtained using view 1 and 5. Upper row depicts results obtained on the images from camera 1, whereas bottom row contains results obtained on the images from view 5. As we can see, despite the occlusion in frame #130 from view 1 the template is located correctly. Similar effect can be observed in frame #151, where this time a woman is under next occlusion in the another view. Owing to the proper tracking in frame #145 in view from camera 1, see middle image in the upper row, the template in view 5 reflects the true location of the target. As we observed, even when the target is fully occluded, see frame #145 in bottom row, or partially occluded, see frames #130 and #151 in upper row, the algorithm still follows the target as long as it is visible from another view, see corresponding images in Fig. 8.

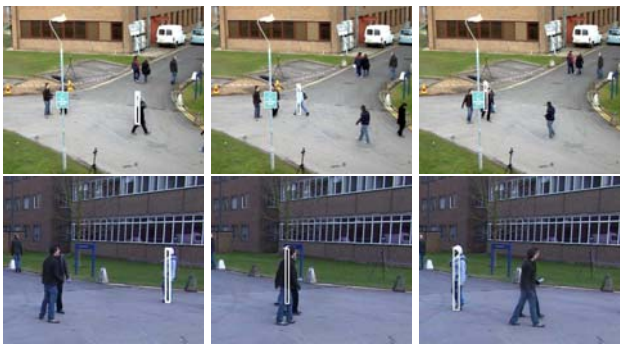

Figure 8. Multi view based person tracking. Frames #130 (left), #145 (middle), #151 (right). Upper row contains images from camera 1, whereas bottom row contains images from view 5

The detection of outliers takes place in each tracker by sorting the similarity scores. If the number of outliers exceeds some value, say 50%, the algorithm calculates the distances between the projected locations and the current locations of the bottom edges of the templates. If the distance is larger than a certain threshold as well as the number of outliers exceeds the mentioned above percentage, the algorithm selects the template with the larger number of outliers.

Afterwards, it validates if the outliers on such a template compose a strip at a certain part of the template, particularly in the upper/left/right part of the template. If yes, then the algorithm verifies if the occlusion was detected in previous frame and then finally marks the template as occluded. The location of such a template is determined on the basis of the another template via the homographic projection.

We realized also tracking using a template modeling only the region of torso. On the basis of location of the feet and the heads or alternatively on the basis of the principal axes we determine the corresponding lines at the ground plane and then their intersection point. Then given the calibration data and location of the head, and assuming that upright person is located on such a point we determine the height of the person. On the basis of the person height and image coordinates in one view we determine the coordinates in the corresponding view, see example results in Fig. 9.

Figure 9. Multi view based person tracking. Frame #130 in view 1 (left), #130 in view 5 (middle), #145 in view 1 (right).

The algorithm has been implemented in C/C++. A typical laptop computer equipped with 2.4 GHz Pentium IV is utilized to run the tracker. The experimental results described in this section were obtained on color images of size $768(720) \times 576$ using 100 particles. The maximal number of iterations in the PSO algorithm has been set to 5.

5. Conclusions

We have presented an algorithm for multiple camera based person tracking. The homography between camera views is utilized to establish the correspondence. The object appearance is modeled via the region covariance that is calculated within the template. The object template is represented by multiple object patches. Owing to such an object representation the tracker is capable of handling considerable partial occlusions. The multiple patch based object representation allows the tracker to detect occlusion as well as it allows resuming the tracking. The information about the template location in one view helps the tracker to recover in an another view, even in case of substantial temporal occlusions or large object movements. Experimental results that were obtained on surveillance videos demonstrated that the proposed multi view tracking algorithm can follow the person even when he/she is fully occluded by an unknown object.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *J. on Comp. Vision and Image Understanding*, 73(3):428–440, 1999.

[2] D. Arsić, N. Lehment, E. Hristov, B. Hörnler, B. Schuller, and G. Rigoll. Applying multi layer homography for multi camera tracking. In *Proc. ACM/IEEE Int. Conf. on Distr. Smart Cameras, Stanford, CA*, pages 1–9, 2008.

[3] D. Arsić, B. Schuller, and G. Rigoll. Multiple camera person tracking in multiple layers combining 2d and 3d information. In *Proc. Workshop on Multi-camera and Multi-modal Sensor Fusion Alg. and Appl., Marseille, France*, Oct. 2008.

[4] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56:411–421, 2006.

[5] R. Eshel and Y. Moses. Homography based multiple camera detection and tracking of people in a dense crowd. In *Proc. IEEE Conf. on Comp. Vision and Pattern Rec., Anchorage, Alaska, USA*, pages 1–8, June 2008.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2000.

[7] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, and Cyb., Part C*, 34(3):334–352, 2004.

[8] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. of IEEE Int. Conf. on Neural Networks*, pages 1942–1948. IEEE Press, Piscataway, NJ, 1995.

[9] S. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *Proc. of the 10th European Conf. on Computer Vision, Graz, Austria*, pages 133–146, 2006.

[10] S. M. Khan, P. Yan, and M. Shah. A homographic framework for the fusion of multi-view silhouettes. In *Proc. IEEE Int. Conf. on Comp. Vision*, pages 1–8, Oct 2007.

[11] B. Kwolek. Object tracking via multi-region covariance and particle swarm optimization. In *Proc. of the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 418–423, Washington, DC, 2009. IEEE Comp. Society.

[12] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *Int. J. Comput. Vision*, 66(1):41–66, 2006.

[13] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on Lie algebra. In *Proc. Int. Conf. on Comp. Vision and Pattern Recognition*, pages 728–735. vol. 1, 2006.

[14] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *Int. J. Comput. Vision*, 77(1-3):125–141, 2008.

[15] R. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE J. of Robotics and Automation*, 3(4):323–344, 1987.

[16] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conf. on Computer Vision*, pages 589–600, Graz, Austria, 2006. LNAI, vol. 3952, Springer-Verlag Berlin Heidelberg.