MULTIVIEW PEDESTRIAN LOCALISATION VIA A PRIME CANDIDATE CHART BASED ON OCCUPANCY LIKELIHOODS

Yuyao Yan[†⋆] *Ming Xu*[†] *Jeremy S. Smith*[⋆]

[†] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China. Email: {yuyao.yan,m.xu}@xjtlu.edu.cn
[⋆] Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, Liverpool, UK. Email: j.s.smith@liv.ac.uk

ABSTRACT

A sound way to localize occluded people is to project the foregrounds from multiple camera views to a reference view by homographies and find the foreground intersections. However, this may give rise to phantoms due to foreground intersections from different people. In this paper, each intersection region is warped back to the original camera view and is associated with a candidate box of the average size of pedestrians at that location. Then a joint occupancy likelihood is calculated for each intersection region. In the second step, essential candidate boxes are identified first, each of which covers at least a part of the foreground that is not covered by another candidate box. The non-essential candidate boxes are selected to cover the remaining foregrounds in the order of their joint occupancy likelihoods. Experiments on benchmark video datasets have demonstrated the good performance of our algorithm in comparison with other state-of-the-art methods.

Index Terms— image motion analysis, object detection, image fusion, visual surveillance

1. INTRODUCTION

An important task in video surveillance is to detect multiple pedestrians. These pedestrians may be partially occluded by each other in a camera view. To overcome this problem, multiple cameras can be deployed to provide complementary information about the moving targets, because the overlapped pedestrians in one camera view may be separated in another camera view. The information provided by the multiple cameras is able to make detection more robust and accurate [1]. When working with multiple cameras, homography has been widely used for the association and fusion of multi-camera observations. Khan and Shah [2] projected the foreground likelihoods from multiple camera views to a reference view by using ground-plane homographies and identified the intersection regions as the locations of pedestrians. This approach adds robustness to the detection of pedestrians. However, the foreground projection of one pedestrian from a camera view may intersect with that of another pedestrian from another camera view, which leads to phantom detections.

Deploying more cameras or placing the cameras at a high elevation may reduce phantoms. The geometric approach to phantom removal compares the heights and sizes of foreground intersection regions with those of pedestrians [3] [4] [5] [6]. The temporal approach to phantom removal employs tracking processes to check the temporal coherence of each foreground intersection region [4] [7] [8]. The colour approach compares the intensities or colours of the original foregrounds for each foreground intersection region [5].

Multiview pedestrian detection is sometimes thought of as an optimization problem. Fleuret et al. [7] calculated a probabilistic occupancy map (POM) in the ground plane which is divided into grids. A pedestrian is modeled as a rectangle of the average size of pedestrians standing in each grid. Then an iterative algorithm is utilized to find the optimal rectangles which cover more foreground pixels and less background pixels in both camera views. Ge et al. [9] proposed a generative sampling-based approach that models a pedestrian as an upright cylinder. Gibbs sampling is used to estimate the number and the locations of pedestrians in a crowd. Utasi and Benedek [10] extended the classical Bayesian Marked Point Process (MPP) model [11] to a 3DMPP model which utilizes the pixel-level features from pedestrians' heads and feet, instead of the whole silhouettes, to reduce the number of phantoms. Peng et al. [12] modeled each pedestrian as a rectangle as in [7] and analyzed the occlusion relationship among such rectangles to identify phantoms by using a Bayesian network.

In this paper an algorithm is proposed for multiview pedestrian localisation. The foregrounds from two camera views are warped to a top view using homographies. Then each intersection region is warped back to both camera views. Each warped back region is associated with a candidate box standing on that region and of the average size of pedestrians. The joint occupancy likelihood of each candidate is calculated by taking into account the foreground likelihood and the

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 60975082 and an XJTLU PhD scholarship under Grant PGRS-12-02-07.

observability of the candidate boxes in both camera views. At the second stage, a prime candidate chart is developed to select the essential candidates, each of which covers at least a foreground region that is not covered by another candidate. Afterwards the non-essential candidates are selected to cover the remaining foreground regions in terms of the joint occupancy likelihoods.

The contributions of this paper are twofold: the use of the prime candidate chart greatly reduces the search space of the optimized solution; the joint occupancy likelihood considers the foreground likelihood and the observability of each candidate.

2. FOREGROUND SEGMENTATION

Background subtraction is used for the foreground detection in each camera view, in which the colour of each pixel is modelled as a Gaussian mixture model [13][14]. In the real world, people may be, or appear to be, walking side by side. This complicates the foreground projections from multiple camera views. In this paper the convex hull of each foreground region is used to separate such pedestrians. The spaces between the contour and the convex hull are defined as convexity defects. Each convexity defect has three main points: the start point p_s, the end point p_e and the farthest defect point p_d.

The convex hull of a group of side-by-side pedestrians usually have one or more large convexity defects facing upwards and between their heads. In order to locate the convexity defects, the direction of each convexity defect is calculated as the bisector of the angle $\angle p_s p_d p_e$:

$$\beta = \arctan\left(\frac{\overrightarrow{p_d p_s}}{|\overrightarrow{p_d p_s}|} + \frac{\overrightarrow{p_d p_e}}{|\overrightarrow{p_d p_e}|}\right) . \quad (1)$$

By thresholding the area of the convexity defect triangles and limiting the angle of β from $-\frac{\pi}{6}$ to $\frac{\pi}{6}$, which ensures the convexity defect is facing upwards, the farthest defect points can be identified and the side-by-side pedestrians are split at that location. The same process is recursively applied to the split foreground regions so that more than two side-by-side pedestrians in a group can be separated.

3. HOMOGRAPHY ESTIMATION

Planar homography is defined by a 3×3 transformation matrix between a pair of captured images of the same plane with two cameras. Let $\mathbf{x}^c$ and $\mathbf{x}^t$ be the homogeneous coordinates of a point in camera view c and its corresponding point in a virtual top view. They are associated by the homography matrix $\mathbf{H}^{t,c}$ as $\mathbf{x}^c \cong \mathbf{H}^{t,c}\mathbf{x}^t$. After each camera is calibrated, a 3×4 projection matrix can be calculated using the intrinsic and extrinsic parameters of the camera: $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]$. Then the homography matrix for the ground plane is:

$$\mathbf{H}_0^{t,c} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_4] . \quad (2)$$

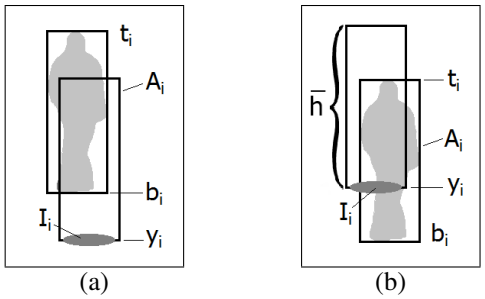


Fig. 1. A schematic diagram of the variables related to the estimation of d_i and h_i.

The homography matrix for a plane parallel to and at a height of h above the ground plane is:

$$\mathbf{H}_h^{t,c} = [\mathbf{m}_1, \mathbf{m}_2, h\mathbf{m}_3 + \mathbf{m}_4] = \mathbf{H}_0^{t,c} + [\mathbf{0}|h\mathbf{m}_3] , \quad (3)$$

where $[\mathbf{0}]$ is a 3×2 zero matrix [15].

In this paper the foregrounds in the individual camera views are projected to the top view using the homographies for the waist plane [16]. Then the foreground intersections are warped back to the individual camera views by using the ground-plane homographies. The warped region for a pedestrian is ideally located at the bottom of the foreground region. If it is well below the bottom of the foreground region, it is a phantom; If it is above the bottom, it may be either a phantom or a pedestrian standing behind another.

4. JOINT OCCUPANCY LIKELIHOODS

Suppose there are N cameras. F_i represents the foreground observation in camera view i. For a specific intersection region I in the top view, there are N warped back intersection regions $\{I_1, I_2, \ldots, I_N\}$, each of which is cast in an individual camera view and is associated with a rectangular box A_i of the average size of pedestrians who are standing there. Let X be the event that there is a pedestrian at intersection region I in the top view. Given foreground observations $F_1, F_2, \ldots, F_N$, we are interested in finding the posterior probability of event X happening. Three independent measurements derived from each foreground region are the foreground pixel set f, the foot location d and the height observation h.

By using Bayes law and considering the conditional independence between the three measurements, we have:

$$\begin{aligned} &P(X|F_1, F_2, \ldots, F_N) \propto P(F_1, F_2, \ldots, F_N|X)P(X) \\ &\propto \prod_{i=1}^{N} P(F_i|X) = \prod_{i=1}^{N} P(f_i, d_i, h_i|X) \\ &= \prod_{i=1}^{N} [P(f_i|X)P(d_i|X)P(h_i|X)] . \end{aligned} \quad (4)$$

f_i is the foreground pixel set enclosed by candidate box A_i, i.e. $f_i = F_i \cap A_i$. $P(f_i|X)$ can be approximated by the

foreground pixel ratio:

$$P(f_i|X) = \frac{number\ of\ foreground\ pixels\ in\ A_i}{number\ of all\ pixels\ in\ A_i} \ . \quad (5)$$

d_i is used to measure the distance between the bottom of the candidate box and that of the corresponding foreground box in camera view i. Suppose the vertical coordinate of the warped region centroid is y_i with a variance $\sigma^2_{y,i}$ which is determined by the height of the warped region, the vertical coordinate of the foreground region bottom is b_i with a variance $\sigma^2_{b,i}$, and the vertical coordinate of the foreground region top is t_i (see Fig. 1(a)). Then the Mahalanobis distance is:

$$d_i = \begin{cases} 0 & if\ b_i \le y_i \le t_i \\ \frac{(y_i - b_i)^2}{(\sigma^2_{y,i} + \sigma^2_{b,i})} & otherwise \end{cases} \ . \quad (6)$$

d_i is chi-square distributed with $n = 1$ degree of freedom, i.e. $d_i \sim \chi^2_1$. Suppose the tail probability on the chi-square distribution is denoted by $Q_{\chi^2}(x, 1) = \int_x^\infty p_{\chi^2}(t, 1)dt$. Given the value of d_i, $P(d_i|X)$ is determined as:

$$P(d_i|X) = Q_{\chi^2}(d_i, 1) \ . \quad (7)$$

h_i is the maximum height of the pedestrian candidate. It is the distance between the bottom of the candidate box and the top of the corresponding foreground box in camera view i (see Fig. 1(b)). Suppose the heights of adults are Gaussian distributed as $h \sim G(\overline{h}, \sigma^2_h)$ and the tail probability on the Gaussian distribution is denoted by $Q_G(X) = \int_x^\infty p_G(t)dt$. Then the maximum height h_i and $P(h_i|X)$ are defined as:

$$h_i = t_i - y_i \quad (8)$$

$$P(h_i|X) = 1 - Q_G(h_i) \ . \quad (9)$$

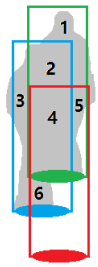
Both d_i and h_i are normalized by the average height $\overline{h}$ of the pedestrians standing at the warped back region.

5. PRIME CANDIDATE CHARTS

The joint occupancy likelihood is derived separately for each pedestrian candidate. To encode the interactivity such as occlusion and grouping between pedestrians, global optimization is carried out for the multiview pedestrian localization. We borrowed the idea from the Quine-McCluskey method [17] [18] for the minimization of Boolean functions.

Each foreground region is decomposed into sub-regions according to the overlapping relationship of all the candidate boxes associated with that foreground region. Each sub-region must be made as large as possible while ensuring that there is no transition on the overlapping candidate boxes inside the sub-region (see Fig. 2(a)). Only the sub-regions which are big enough and contain a significant portion of foregrounds are used.

A prime candidate chart is introduced to select a minimum set of pedestrian candidates to cover all the foreground



(a)

Sub-region:	1	2	3	4	5	6
Red	+	+	+	X	X	X
Green	X	X	+	X	X	+
Blue	+	X	X	X	+	X

(b)

Fig. 2. (a) Decomposition of a foreground region into sub-regions. (b) The corresponding prime candidate chart.

sub-regions of interest. In the prime candidate chart (see Fig. 2(b)), the foreground sub-regions in all the camera views are listed across the top of the chart, and the pedestrian candidates are listed down the left side. If a candidate covers a given sub-region, an X is placed at the intersection of the corresponding row and column; otherwise a plus sign is placed.

The prime candidate chart is updated as follows:

(1) All the candidates are scanned. If the joint occupancy likelihood of any candidate is too low, it is removed from this chart by replacing the X's in the corresponding row by plus signs.

(2) The sub-regions are scanned. If a sub-region is covered by only one candidate, the candidate is labeled as an essential candidate and a pedestrian. The Xs in the corresponding row and in the columns of the sub-regions covered by this candidate is replaced by plus signs.

(3) If there are sub-regions not covered, find any candidate with its sub-regions fully contained in another candidate. Remove the X's of the contained candidate, which may leave a single X in some columns. Proceed as if the corresponding candidate is an essential one.

(4) Scan the sub-regions. If there are sub-region not covered, each must be covered by two or more candidates. Select a column with two X's. Assume the candidate corresponding to an X is essential and repeat steps 2-3. Then try the other X. In the two groups of resultant 'pedestrians', select the one with a greater joint occupancy likelihood.

6. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, experiments were performed on the PETS2009 City Center (CC) dataset [19] which is a benchmark dataset containing a crowd of pedestrians in 8 calibrated camera views. Only two camera views (views 1 and 2) were used in our experiments. Each view has 795 frames, in which the first 200 frames were used to train the background model and the remaining frames were used to evaluate the performance.

Fig. 3 shows the detection results at frame 726 on the

(a) (b) (c)

Fig. 3. The detection results at frame 726 on the PETS2009 CC dataset. (a)(b) Camera views 1 and 2, and (c) the top view.

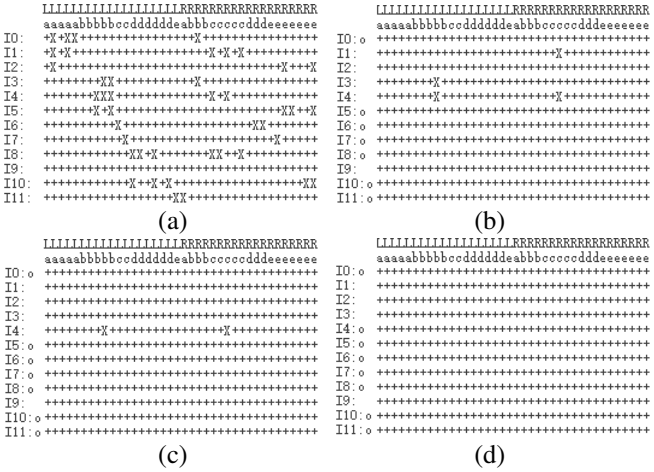
Table 1. Evaluation results on PETS 2009 CC dataset.

Method	Evaluation	RECALL	PRECISION	TER
3DMPP	[10]	N/A	N/A	0.31
POM	[10]	N/A	N/A	0.27
POM	us	0.91	0.82	0.29
MvBN	[12]	0.90	0.97	0.13
Proposed	us	0.96	0.99	0.05

PETS2009 CC dataset. The borderlines of the overlapping field of views are shown as black dashed lines. The contour and bounding box of each foreground region are in green and black, respectively. Each foreground intersection region in the top view, which is represented in a distinguished colour, corresponds to a pair of candidate boxes represented by the same colour in both camera views. The intersection region IDs are shown at the top of Fig. 3(c) and also in the same distinguished colour. An identified pedestrian is labeled with a circle at the bottom of its candidate box, while each phantom is labeled with a cross.

Fig. 4 is the prime candidate chart at the same frame as Fig. 3. Down the left side of the chart is the list of pedestrian candidates. If a candidate is identified as a pedestrian, then it is labeled with a circle. At the top of each chart, L and R represent the left and right camera views. In the second row, a-e are foreground region IDs. If a foreground region ID appears successively several times, they refer to the sub-regions decomposed from the same foreground region. Fig. 4(a) shows the chart after step 1 by removing invalid candidates, Fig. 4(b) is the chart after step 2 by removing essential candidates, Fig. 4(c) and 4(d) are those in and after step 3 by merging non-essential candidates.

For a performance comparison with some state-of-the-art algorithms, three metrics were evaluated: PRECISION, RECALL and TER (total error rate) [10] [12], which are defined as the ratios $TP/(TP+FP)$, $TP/(TP+FN)$ and $(FN+FP)/(TP+FN)$, respectively (TP, FP and FN are



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

Fig. 4. The prime candidate chart at frame 726: (a) after step 1, (b) after step 2, (c) in step 3 and (d) after step 3.

the numbers of true positives, false positives and false negatives). A larger PRECISION/RECALL value or a lower TER value indicates a better performance. The comparison results based on camera views 1 and 2 in PETS2009 CC dataset are shown in Table 1. The proposed algorithm outperforms these algorithms in terms of PRECISION, RECALL and TER.

The number of steps used in the prime candidate chart in each frame was investigated. In the 595 frames for the test, the algorithm came to an end in 561 frames (94.3%) after essential candidates were identified and it had to resort to the selection of non-essential candidates in only 34 frames (5.7%).

7. CONCLUSIONS

We have proposed an algorithm for multiview pedestrian detection, which is based on foreground intersections in a top view. The joint occupancy likelihoods and the prime candidate chart used in this paper add the robustness to the pedestrian localization. Experiment results have shown its much better performance than some state-of-the-art algorithms. The structure of the algorithm facilitates the use of more cameras.

8. REFERENCES

[1] M. Xu, J. Orwell, L. Lowey, and D. Thirde, "Architecture and algorithms for tracking football players with multiple cameras," *IEE Proc. Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 232–241, 2005.

[2] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *European Conference on Computer Vision*, 2006, pp. 133–146.

[3] D. B. Yang, H. H. Gonzlez-Baos, and L. J. Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *International Conference on Computer Vision*, 2003, pp. 122–129.

[4] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, 2009.

[5] R. Eshel and Y. Moses, "Tracking in a dense crowd using multiple cameras," *International Journal of Computer Vision*, vol. 88, no. 1, pp. 129–143, 2010.

[6] D. Arsic, E. Hristov, N. Lehment, B. Hornler, B. Schuller, and G. Rigoll, "Applying multi layer homography for multi camera person tracking," in *ACM/IEEE International Conference on Distributed Smart Cameras*, 2008, pp. 1–9.

[7] F. Fleuret, J. Berclaz, R., and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, 2008.

[8] M. Liem and D. M. Gavrila, "Multi-person tracking with overlapping cameras in complex, dynamic environments," in *British Machine Vision Conference*, 2009, pp. 199–218.

[9] W. Ge and R. T. Collins, "Crowd detection with a multiview sampler," in *European Conference on Computer Vision*, 2010, pp. 324–337.

[10] A. Utasi and C. Benedek, "A bayesian approach on people localization in multicamera systems," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 23, no. 1, pp. 105–115, 2013.

[11] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2913–2920.

[12] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang, "Robust multiple cameras pedestrian detection with multi-view bayesian network," *Pattern Recognition*, vol. 48, no. 5, pp. 1760–1772, 2015.

[13] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, vol. 2, pp. 246–252.

[14] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *International Conference on Pattern Recognition*. IEEE, 2004, vol. 2, pp. 28–31.

[15] J. Ren, M. Xu, J. S. Smith, and S. Cheng, "Multi-view and multi-plane data fusion for effective pedestrian detection in intelligent visual surveillance," *Multidimensional Systems and Signal Processing*, vol. 27, no. 4, pp. 1007–1029, 2016.

[16] M. Xu, J. Ren, D. Chen, J. S. Smith, and G. Wang, "Real-time detection via homography mapping of foreground polygons from multiple cameras," in *Proc. IEEE International Conference on Image Processing*, 2011, pp. 3593–3596.

[17] W. V. Quine, "The problem of simplifying truth functions," *The American Mathematical Monthly*, vol. 59, no. 8, pp. 521–531, 1952.

[18] E. J. McCluskey, "Minimization of Boolean functions," *Bell System Technical Journal*, vol. 35, no. 6, pp. 1417–1444, 1956.

[19] A. Ellis, A. Shahrokni, and J. Ferryman, "PETS2009 and winter-PETS2009 results: A combined evaluation," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*. IEEE, 2009, pp. 1–8.