Person Detection using Temporal and Geometric Context with a Pan Tilt Zoom Camera

Alberto Del Bimbo, Giuseppe Lisanti, Iacopo Masi, Federico Pernici
University of Florence
Media Integration and Communication Center (MICC)
Florence, Italy
{*delbimbo, lisanti, masi, pernici*}*@dsi.unifi.it*

***Abstract*—In this paper we present a system that integrates automatic camera geometry estimation and object detection from a Pan Tilt Zoom camera.**

We estimate camera pose with respect to a world scene plane in real-time and perform human detection exploiting the relative space-time context. Using camera self-localization, 2D object detections are clustered in a 3D world coordinate frame. Target scale inference is further exploited to reduce the number of false alarms and to increase also the detection rate in the final non-maximum suppression stage.

Our integrated system applied on real-world data shows superior performance with respect to the standard detector used.

***Keywords*-person detection; PTZ camera; context; structure from motion; SVM;**

I. Introduction and paper contributions

The task we address in this paper is 3D multiple pedestrian detection from a PTZ camera. At any point in time, we want to detect pedestrian, or other objects, localize them on the 3D scene plane, estimate their trajectories without losing target association and eventually predict their future motion for the purpose of sensor management. We are indeed ultimately interested in performing sensor management in order to determine the best way to task the visual sensor, trough pan tilt zoom, to detect and track or recognize targets.

However 3D tracking from a moving and zooming camera is difficult, since there are many sources of uncertainty in object localization. Indeed beyond measurement noise, clutter, changing background and occlusions, camera pose estimation while zooming causes several difficulties.

This challenging problem has been largely neglected until recently [1], in which the authors propose a real time system capable of internal and external camera parameters estimation in video sequence taken with large focal length changes. They address the problem as a real time retrieval system in which a database of pre-build features with associated bundle adjusted homographies are retrieved and subsequently refined online. Pose estimate is then used to perform multiple target tracking in 3D assuming a planar scene. A major limit of this work is that target templates used to acquire image measurements are not automatically detected and have to be manually specified in the first frame of the sequence. This narrows the general applicability of the framework in real surveillance scenario.

In this paper we build upon this work to propose a detection module that exploits the camera pose geometric context to obtain better performances in pedestrian detection. Scene context has been recently shown to facilitate object detection methods in outdoor image dataset, when coarse estimated camera pose are available [2]. We follow a similar approach but introducing both geometric and temporal contextual knowledge. Context allows objects to be physically placed within the 3D world and this further allows reasoning between these objects and the 3D environment. Estimated camera geometry adds a context to individual object detections and temporal coherence provides them with a history, supporting their presence in the video frames. To this end 3D objects localization are accumulated over a time-based sliding windows and are clustered in 3D. Redundant detections in 3D, originating from the same target or clutter are used to improve target detection performance.

We provide the following improvements to the results proposed in [1]: (1) we improve scanning windows based human detector, considering a PTZ camera tracking system with 3D space-time context at frame-rate; (2) we demonstrate how space-time context improves precision and recall performance of recent state of the art for object detectors.

II. Related work

Besides the difficulties of Data Association and Filtering [3], 3D Multiple Target Tracking (MTT) with a moving camera is notoriously difficult because of target detection and sensor registration (i.e. manage visual data to account for camera zooming and pose changes). While these three components are far from being solved individually, their integration and interplay considerably enhance overall tracking performance. Among these three components, object detectors have made tremendous progress over the last few years and are getting applicable in complex surveillance scenes [4], [5], [6]. The work [6] in particular improves upon others using multi-scale deformable part models, learned with discriminative training and demonstraes good performance on difficult benchmarks. Recently [2] shows that

geometric scene context, estimated from a single image, improves object detection and recognition. Authors propose a probabilistic inference model to merge pre-trained detector responses with scene knowledge as coarse camera viewpoint and surface orientation estimation. The work [7] (closely related to our work) combines detectors and geometric context extracted using real time Structure from Motion (SfM) with a moving camera. In [8], [9] miss-detection are reduced by image rectification providing the detector with a better viewpoint.

Further methods jointly addressing detection and tracking are [10] and [11] among others. These works assume an existence variable that follows a discrete Markov chain parameterized by object birth and death probabilities. Yet principled, these methods are very computationally intensive and not always suitable for real time automated surveillance purpose.

Specific multiple target detection and tracking with a PTZ camera is addressed in [12]. Authors proposed a system built upon [13] for tracking targets in 3D. The world-to-image homography is computed from the hockey rink model and adaptation to target scale changes is performed by examining windows slightly larger/smaller than the current target size. Available geometric context is not taken into account in object detection.

III. Real Time PTZ Camera Pose and Focal Length Estimation

Our real time SfM module is based on the approach described in [1]. SURF interest points [14] are matched onto a set of bundle adjusted images (covering the whole field of regards of the PTZ camera) each of which has an associated world to image homography $\mathtt{G}_t$. The homography associated to the closest retrieved image is subsequently refined trough RANSAC and the focal length for the current frame is extracted from the combination of the nearest homography with the world to image homography. Pedestrian have been assumed to be closely vertical in the 3D scene plane position of the two extremities and the imaged feet and heads location for humans are related by a time variant planar homology:

$$\mathtt{W}_t = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_{t,\infty} \cdot \mathbf{l}_{t,\infty}^{\mathrm{T}}}{\mathbf{v}_{t,\infty}^{\mathrm{T}} \cdot \mathbf{l}_{t,\infty}} \quad . \tag{1}$$

For the individual images of the sequence, the vanishing line $\mathbf{l}_{t,\infty}$ and the vanishing point $\mathbf{v}_{t,\infty}$ change according to the variation of the camera parameters due to the pan-tilt-zoom operation, for each image at time t the $\mathtt{W}_t$ is completely defined by: $\mathbf{l}_{t,\infty} = \mathtt{G}_t \cdot [0,\ 0,\ 1]^{\mathrm{T}}$ and $\mathbf{v}_{t,\infty} = \mathtt{K}_t\mathtt{K}_t^{\mathrm{T}} \cdot \mathbf{l}_{t,\infty}$. The cross-ratio μ, being projective invariant, remains constant throughout the sequence while $\mathtt{K}_t$ is the internal camera matrix parameterized by the extracted focal length.

IV. Human Detection with Context

State of the art object detectors as [5], [6] use one or more SVM trained linear filters to detect an object model in the image. Objects from a particular category are localized in the image by thresholding the filter response evaluated at each position and scale using HOG features pyramid. Besides the excellent performance, their application depends specifically on the task to be accomplished. In the case of real time tracking, detector recall performance is more appropriate since contextual knowledge (if present) could be exploited to refine the precision performance as well.

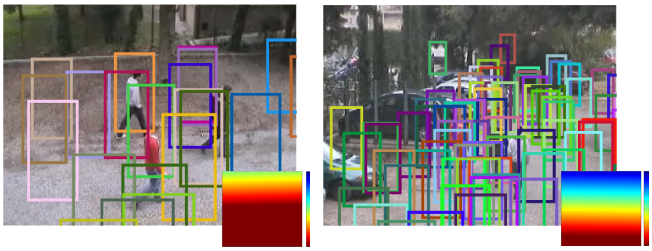

Figure 1. Randomly generated feasible pedestrian detections with context. The geometric context is estimated according to [1]. In particular two views with different scene depth are shown. Each *bottom-right* frame shows the imaged height likelihood.

This refinement stage is generally handled by a suppression of non-maximum responses using some kind of prior knowledge over detections. Detector [5] implements this stage by performing mean-shift [15] in the detection space (i.e. position and scale). While [6] sorts the detections by their filter score, and greedily select the highest ones by excluding detections with bounding boxes that are at least 50% overlapped with a bounding box of a previously selected detection. The underline assumption here is that detector response performance gracefully degrades in the neighborhood of the target image location (locality context).

A. Imaged Height Context

In the case of object detection throughout a video sequence taken with a PTZ camera, space-time context over frames is used to further refine over the locality of detection responses. More formally we exploit the facts that: (1) the detector gives in general a strong positive response in the neighborhood of the object; (2) the detector does not report with same frequency and uncertainty non-object image region; (3) the higher the peak detection score, the higher the probability for the image region to be a true positive; (4) imaged scale and 3D world coordinate at which a target can be detected in a given image are known (i.e. they are estimated with the method in [1] as shown in Fig. 1). Under these assumptions we proceed by filtering out all the detections which does not correspond to the geometric context. Given a set of n detection $D = \{B_i, s_i\}_{i=1}^{n}$ defined by their bounding box $B = (x_1, y_1, x_2, y_2)$ (upper left corner coordinates and lower right corner respectively) and score s, for each detection in the set D, eq. (1) is evaluated at the foot location $\mathbf{z}_l = (\frac{x_1+x_2}{2}, y_1, 1)$ to obtain the corresponding

head location $\mathbf{z}_u = \mathtt{W}_t \mathbf{z}_l$.

For each detection in D, the corresponding detected height and width are computed as $h_d = y_2 - y_1$, $w_d = x_2 - x_1$; while the contextual height and width are respectively estimated as: $\hat{h} = ||\mathbf{z}_l - \mathbf{z}_u||$ and $\hat{w} = \frac{\hat{h}}{\alpha}$, where α is the imaged human aspect ratio. The detections in D are finally filtered out as:

$$D' = \{(B_i, s_i) \in D : |h_d - \hat{h}| < \epsilon_h, \left|\frac{h_d}{w_d} - \frac{\hat{h}}{\hat{w}}\right| < \epsilon_r\} \quad (2)$$

where ϵ_h and ϵ_r are system thresholds which values depend on the accuracy of the SfM module.

B. Temporal Context in 3D Localizations

Object detections are converted to 3D observations using the image to world homography $\mathtt{G}_t^{-1}$ and accumulated in a 3D world coordinate plane to form clusters. Our idea is to group a collection of 3D detected observations in a sliding window buffer $\mathbf{X}_t = \{\mathbf{x}_{t'}\}_{t'=t-\tau}^{t}$ of length τ, where $\mathbf{x}_{t'} = \mathtt{G}_t^{-1}\mathbf{z}_t$, with $\mathbf{z}_t \in D'$ are the filtered detection responses provided by the height context in eq. 2. Cluster analysis is performed in $\mathbf{X}_t$ using parametric Gaussian Mixture Model Expectation Maximization [16]. Clusters coordinate that are close each other are automatically joined together by the EM process. The mode of the most confident clusters provides the final detection results.

Observation in $\mathbf{X}_t$ are in the spatial vicinity of each other at different time steps, therefore, the number of detected localization grows with time for persistent detections. On the other hand, if a detection event originates from clutter, it is less likely to form any pattern or cluster of detections in the vicinity of each other within τ. According to this the mixture model selection-clustering algorithm will ignore it automatically. Mutual exclusion principle is implicitly enforced being the parameters space of clustering defined in the Euclidean 3D scene plane. We've noted that the life span of a detection event represented by this false alarm region is very short when compared with that of a true target.

However targets motion may still compromise clustering performance. According to this, further 3D knowledge is enforced with 3D kinematic priors of moving targets by removing clusters with large eigenvalues or with an eigenvalues ratio that exceeds a reasonable kinematic motion.

Finally in order to keep real time performance and to select clusters that represent with high probability true positive detections, we limit number of samples used to feed EM algorithm. According to this, elements in D' are resampled based on their SVM score. Posterior class probability used to sample $\mathbf{X}_t$ are computed from the $s_i \in D'$, as described in [17]. This resampling step gives also some kind of randomization to the entire process, which may help to avoid local minima during EM iterations.

V. Experimental Results

Experimental results are performed on ten video sequences at a resolution of 320x240 at 20 fps, for a total of 7920 frames. The video dataset is taken with a PTZ camera framing moving targets that enter and leave the scene. We used the human detector in [6]. Detection performance is quantitatively evaluated using recall-precision curves (RPCs) on manually labeled bounding boxes. A correct detection is scored when: $A_o = \frac{area(B_d \cap B_{gt})}{area(B_d \cup B_{gt})} > 0.5$, where A_o is the area of overlap between the predicted bounding box B_d and the manually annotated ground truth B_{gt}.

Figure 2. Qualitative Comparison: *(left)* part-based detector *(center)* enforced with geometric context *(right)* proposed method with both geometric and temporal context. *(a):* targets enter the scene *(b):* large focal length.

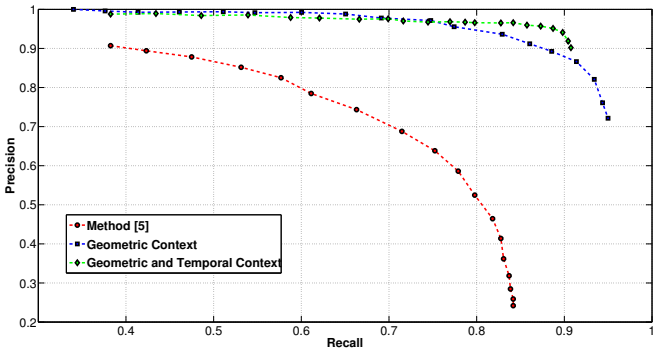

Figure 3. RPCs Comparison.

Fig. 2 shows qualitatively the improvements achieved by introducing our method in the part-based detector described in [6]. In particular the figure shows two frames out of the ten sequences with different levels of zoom and perspective effects. It can be noted that less false positive detections are generally present and that bounding boxes surrounding pedestrians are more accurately localized. Increased accuracy in localization is achieved by using the modes of the estimated clusters.

Quantitative comparisons are summarized in Fig. 3 through precision recall curves. The figure shows that a significant improvement is obtained by our proposed method in exploiting contextual knowledge.

VI. Conclusion

We have presented a system for human detection exploiting contextual knowledge about camera/scene geometry and temporal coherence in sequence taken with a PTZ camera. Target scale inference, camera self-localization, object detection and clustering in 3D world coordinate frame are combined to reduce the number of false alarms and to increase also the detection rate in the final non-maximum suppression stage. We have shown that object detector performance in [6] improves considerably with the proposed method. Further research may be investigated by integrating the proposed system to automatically initialize targets template in order to improve tracking data association.

Acknowledgment

This work is partially supported by Thales Italia, Florence, Italy.

References

[1] A. Del Bimbo, G. Lisanti, and F. Pernici, "Scale invariant 3D multi-person tracking using a base set of bundle adjusted visual landmarks" in *Proc. of ICCV Int'l Workshop on Visual Surveillance*, 2009.

[2] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective" in *Internation Conference on Computer Vision and Pattern Recognition*, 2006.

[3] J. Vermaak, S. Godsill, and P. Perez, "Monte carlo filtering for multi target tracking and data association" *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 309–332, Jan. 2005.

[4] P. Viola and M. J. Jones, "Robust real-time face detection" *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection" in *Internation Conference on Computer Vision and Pattern Recognition*, 2005.

[6] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model" in *Internation Conference on Computer Vision and Pattern Recognition*, 2008.

[7] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool, "Dynamic 3D scene analysis from a moving vehicle" in *Internation Conference on Computer Vision and Pattern Recognition*, 2007.

[8] L. Yuan, W. Bo, and N. Ram, "Human detection by searching in 3D space using camera and scene knowledge" in *International Conference on Pattern Recognition*, 2008.

[9] J. Yao and J. Odobez, "Multi-Camera Multi-Person 3D Space Tracking with MCMC in Surveillance Scenarios" in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2*, 2008.

[10] J. Vermaak, S. Maskell, M. Briers, and P. Prez, "Bayesian visual tracking with existence process" in *International Conference on Image Processing*, 2005.

[11] C. Kreucher, K. Kastella, and A. Hero, "Multitarget tracking using a particle filter representation of the joint multitarget probability density" *IEEE Transaction on Aerospace Electronics Systems*, vol. AES-39, no. 4, pp. 1396–1414, 2005.

[12] N. d. F. Yizheng Cai and J. Little., "Robust visual tracking for multiple targets." in *European Conference on Computer Vision*, 2006.

[13] K. Okuma, J. J. Little, and D. G. Lowe, "Automatic rectification of long image sequences" in *Asian Conference on Computer Vision*, 2004.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)" *Computer Vision Image Understanding*, vol. 110, pp. 346–359, 2008.

[15] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–575, 2003.

[16] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 381–396, 2000.

[17] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines" *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, 2007.