Person detection, tracking and masking for automated annotation of large CCTV datasets

Marcos Nieto, Peter Leškovský and Juan Diego Ortega

Vicomtech-IK4, Paseo Mikeletegi 57, San Sebastian 20009, Spain,
{mnieto, pleskovsky, jdortega}@vicomtech.org

Abstract. In this paper we describe a real-time approach for person detection in video footage, joint with a privacy masking tool, in the framework of forensic applications in CCTV systems. Particularly, this paper summarizes our results in these domains within the European FP7 SAVASA and P-REACT projects. Our main contributions have been focused on real-time performance of detection algorithms using a novel perspective-based approach, and the creation of a methodology for privacy masking content such as the faces of the persons in the images.

Keywords: computer vision, real-time, detection and tracking, privacy masking

1 Introduction

Analyzing large volumes of video footage in CCTV systems is troublesome and expensive for CCTV operators and law enforcement agencies. Companies, public institutions and the research community are pushing forward to provide technology solutions, especially in the field of video analytics, providing semantic-aware, remotely accessible, reduced-size annotations. These annotations can be stored and made accessible through VSaaS (Video Surveillance as a Service) with applications like the SAVASA system [1], or the P-REACT platform [2].

In this work we present our developments related to person detection, which can provide rich information of the scene and that can be used further by high-level semantic analysis to track persons or to recognise actions between persons. Our approach is based on a novel perspective-based for enhanced efficiency compared with traditional detection approaches.

Besides, the effective exploitation of this information must be compliant with the privacy and ethical rules of local jurisprudences, which may be very restrictive in some cases such as in Europe. In that sense, we have also worked on creating tools for masking private or protected content (e.g. detected faces in images) using standard tools and facilitating the protected reconstruction of the material with secure keys and a dedicated player.

2 Automatic content annotation

Figure 1 illustrates the conceptual architecture of the analysis modules of the automatic annotation scheme which includes the mentioned modules of person detection, tracking and privacy masking. Basically, this architecture shows the integration layers that can be used to connect to any CCTV system, using the proprietary SDKs of the NVR providers or standard interfaces like ONVIF. After a transcoding stage, the platform launches the person detection and tracking module, and additional feature extraction methods for high-level video analytics (more details can be found in [3]).

The detected persons are described in XML files that can then be used to build protected video files, using cryptographic methods and adding watermarking for providing digital evidence services.

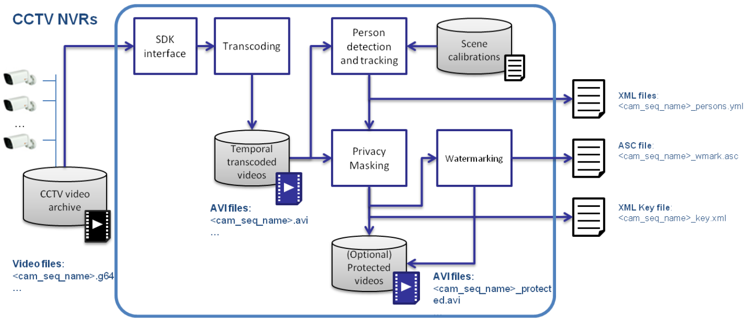

Fig. 1. Block diagram of the video analytic modules of the automatic annotation scheme.

3 Person detection

Using a default person model, we generate a grid of positions of parallelepipeds in the 3D world lying on a dominant plane, separated by defined steps (in metric units) [4]. The projection of the volumes defined this way provides the set of rectangles candidates to contain persons in the scene. The great advantage of this approach is that the hypotheses are all potentially correct compared to multiscale approaches, where there are a number of hypotheses whose sizes do not fit with their positions in the dominant plane. Hence the number of candidates is greatly reduced, and there is no more need to filter out absurd hypotheses (with respect to perspective) as it happens when using traditional scanning window approaches [5].

This grid-based approach can be used with any detector scheme (e.g. classifiers, foreground detection, etc) that provides a weighted output between 0 and 1. In our work, we have used two types of detectors: background subtraction and appearance-based detectors. The former are helpful to detect moving objects; the latter enhances the reliability of the system searching patterns of head-and-shoulder and full-body shapes. Analogously, when the perspective of the scene give a close view of the persons, it is possible to use face detection methods.

When densely sampling an image, several candidates partially represent each person. In Fig. 2 (right), the top view of the projected grid with the associated detection values, represented by the radius of each red circle at each cell position,

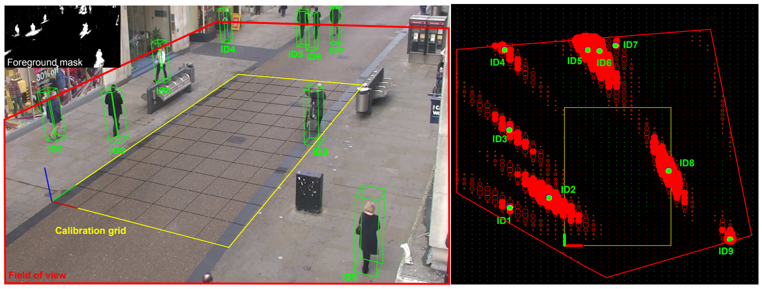

Fig. 2. Detection of 3D persons using the perspective grid (left) and an example top view of the grid (right).

is shown. Given this detection map, we obtain a refined map by removing noise and detections corresponding to figures larger or smaller than the person model. We do this by deconvolution of the map with a kernel corresponding to pixel occupancy of our person model at each grid cell.

The 3D detections are then mapped back to the image and a tracking stage associates them to existing tracks, which are then updated using the new position and size of the detections using a Kalman filter to smooth the trajectories.

The main achievement using the proposed perspective-grid detection approach is the reduction of the number of candidate regions in the images to be evaluated by the detector, along with the possibility to combine this methodology with any existing detector. To measure this feature, we have compared the use of the perspective-grid (PG) approach with a brute-force multiscale scanning window (BF-MSW) and a fine-tuned multiscale scanning window (FT-MSW) [4] in three different scenes with low, medium and far perspectives, respectively.

Sequence	BF-MSW	FT-MSW	PG	PG improvement (%)
(a) TRECVID Cam1 [6]	27998	3051	2344	14,76
(b) UT-Interaction Dataset [7]	21105	2603	390	85,01
(c) IKUSI Cam3	27895	21588	8744	59,49

Table 1. Comparison of the number of candidates generated by the multiscale methods and the proposed perspective-grid (PG) method.

The fourth column of Table 1 shows the improvement of the proposed method over the results of the FT-MSW that also uses the perspective information of the scene to select the best parameters of the scanning window method. The datasets used correspond to a close view (a), an intermediate view (b) and a distant view of a parking area (c) scenes.

4 Privacy masking

After personal or other privacy information (e.g. faces, license plates) is detected, we apply reversible occlusion of the corresponding regions in the video. We first extract the confidential regions from and occlude them in the main stream. The extracted regions are formed into privacy streams which are entirely encrypted once their encoding into desired video format has been obtained. Finally, all streams are encapsulated in common video container file. Upon video reply, we first decrypt the privacy streams and position them correctly within the main stream in order to reconstruct the original video.

Our approach to video encryption is non-standard due to the fact that there are no open or public standards which would define the application of encryption schemes to videos. We therefore followed general guidelines on encryption and key distribution applied to binary information. Especially, we followed the general cryptographic approaches applied in proprietary solutions like are DVD or Blue-ray media. The video is encrypted in its binary form with a symmetric AES cipher and the secret key used is then distributed in encrypted form, applying an asymmetric RSA encryption scheme. Private and public keys in the form defined by the OpenPGP standard protocol were used for the RSA encryption and distribution of the AES secret key among users.

5 Conclusions

We have presented an efficient methodology to exploit the perspective information of the scene to dramatically reduce the computational complexity of person detection algorithms for video surveillance applications turning it available for real-time processing. Moreover, our proposed method can conveniently be used with any existing input detector, such as background detectors, or detection-by-classification methods.

The privacy masking applied allows us to mask private information otherwise recognisable in the captured recordings and thus to comply to the strict privacy protection regulations set by the EU for distribution of surveillance videos. Nevertheless, it depends on the automatic detection of regions holding private information, which, for example for face detection, is still not 100% accurate.

References

1. FP7-SEC-2011-1 SAVASA project `http://www.savasa.eu`
2. FP7-SEC-2013.7.2-1 P-REACT project `http://www.p-react.eu`
3. Jargalsaikhan, I., Direkoglu, C., Little, S., O'Connor, N. E.: An evaluation of local action descriptors for human action classification in the presence of occlusion. In MultiMedia Modeling. Dublin, Ireland. January 2014.
4. Nieto, M., Ortega, J. D., Cortes, A., Gaines, S.: Perspective Multiscale Detection and Tracking of persons. MMM 2014, PartII, LNCS 8326, pp. 92-103, 2014.
5. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision, pp.3–7, 2008.
6. Smeaton, A. F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. MIR'06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp.312–330, 2006.
7. Ryoo, M. S., Aggarwal, J. K.,: UT-Interaction Dataset ICPR contest on Semantic Description of Human Activities (SDHA). 2010.