A system to automatically track humans and vehicles with a PTZ camera

M. Lalonde[a], S. Foucher[a], L. Gagnon*[a], E. Pronovost[b], M. Derenne[b], A. Janelle[b]
[a]R&D Dept., CRIM, 550 Sherbrooke Street West, Suite 100, Montreal, QC, Canada, H3A 1B9
[b]VideoStream Technologies Inc., 7493 Trans-Canada Highway, Suite 103, St-Laurent, QC, Canada, H4T 1T3

ABSTRACT

The paper reports about the development of a software module that allows autonomous object detection, recognition and tracking in outdoor urban environment. The purpose of the project was to endow a commercial PTZ camera with object tracking and recognition capability to automate some surveillance tasks. The module can discriminate between various moving objects and identify the presence of pedestrians or vehicles, track them, and zoom on them, in near real-time. The paper gives an overview of the module characteristics and its operational uses within the commercial system.

Keywords: Video surveillance, object tracking, object recognition, real-time implementation

1. INTRODUCTION

The aim of this paper is to give an overview of a practical object tracking modules that was developed for a video surveillance application with specific software and hardware constraints. The goal of the project was to develop and integrate a software module, for a commercial surveillance system equipped with Pan-Tilt-Zoom (PTZ) cameras that allows autonomous object detection, tracking and recognition in an outdoor urban environment. The module had to be able to discriminate between various moving objects, discard non-relevant ones, identify the presence of pedestrians or vehicles, track them, and optionally zoom on them, all in near real-time.

Video surveillance is now an active research topic. An important R&D initiative at the origin of this field was the U.S. government-funded program "Video Surveillance and Monitoring" (VSAM) [1]. The Defense Advanced Research Projects Agency (DARPA) Information Systems Office launched the three-year VSAM program in 1997 to develop automated video understanding technology for use in future urban and battlefield surveillance applications. The VSAM program looked at several fundamental issues in detection, tracking, auto-calibration, and multi-camera systems and motivated many other academic researches (for instance, [5-7]). Collins et al. [5] have developed a system that allows a human operator to monitor activities over a large area using multiple calibrated cameras with a geospatial site model. Tracking approach is based on image correlation mapping, followed by computation of the 3D location on the site model. Inter-sensor communication consists in a "handing off" mechanism between sensors situated along the object's trajectory.

A decentralized architecture has also been proposed using multiple calibrated cameras to learn patterns of activities from motion observation [13]. The basic assumption for learning is the preservation of the object identity throughout the tracking process. Another wide area surveillance system using client-server architecture has proposed by Javed et al. [9-10]. It uses non-calibrated cameras with overlapping and/or non-overlapping fields of view (FOVs). The system is trained to learn the topology of the FOVs. The inter-camera correspondence is established based on linear velocity prediction and on a spatio-temporal constraint based on the FOVs topology. Another real-time visual surveillance system for detecting and tracking multiple people and monitoring their activities in an outdoor environment was proposed by Haritaoglu et al. [8]. It operates on monocular grayscale video or infrared imagery and employs a combination of shape analysis and tracking to locate people and create appearance models. It can determine whether a

* langis.gagnon@crim.ca; phone 514-840-1235; fax 514-840-1244; crim.ca

foreground region contains multiple people and can segment the region into its constituents. More recently, a system with a decentralized architecture has been developed [2,7] with no dependence on a central server that could fail during an operational mode. The intelligent nodes send and receive information between them and a pair of cameras are attached to each node (one of them is an infrared camera) to improve performance in low-light conditions).

The work presented here differs mainly from the above by its non-academic and industrially-oriented nature. The module has been developed for VideoStream Technologies Inc. and integrated within their VST OneTrack system, an intelligent security/CCTV Windows application that controls PTZ cameras. The target application of the VST OneTrack system is the monitoring of an outdoor scene (for example a parking lot) without an operator, for the detection and tracking of pertinent moving events in order to capture quality images (close-up shots) that can be used, for instance, to confirm the identification of a person or a vehicle (with its plate number). Whereas most existing security video intelligence systems use fixed cameras, in this application the camera can be in motion horizontally, vertically and has zoom capability. The project necessitated the development of practical original solutions to the following specifications:

- Robust and fast foreground/background segmentation, ideally unaffected by any undesirable effect caused by e.g. lighting changes;
- Object recognition and classification to allow intelligent tracking (zoom-in) of pertinent objects (person, car, truck) versus useless ones (e.g. animals, trees, birds, puddles of water, windows reflections, rain, etc.);
- Low cost.

This paper gives an overview of the system characteristics (Section 2) and its operational uses (Section 3). Some algorithmic and implementation details are deliberately not detailed due to their sensitive industrial nature.

2. SYSTEM CHARACTERISTICS

The intelligent video surveillance module is composed of four main components: (1) background modeling, (2) track management, (3) object recognition and (4) light tracking. These components enable two use cases: In the first mode (using the two first components), the system observes a scene in a wide FOV, tracks humans and vehicles and moves the PTZ camera to zoom on a specific object if needed. More than one scene can be visited thanks to the pan-tilt programming feature, but each scene is analyzed independently. In the second mode (using the fourth component) initiated by the first mode or by the user, the system concentrates on a specific potentially moving object and locks the camera onto it.

2.1 Background modeling

The vast majority of intelligent video surveillance systems include a background analysis component (e.g. background suppression, removal or subtraction). Most of them assume a simple static background. Although a few number of dynamic background suppression models have been published recently (e.g. [11]), they were not considered in this study due to uncertainties regarding performance detection and, more importantly, real-time constraints. The hypothesis is that moving objects that should belong to the background will generate incoherent motion and can be eliminated with post-processing filtering. Our background model is based on an optimized variant of a probabilistic neural network (PNN) and it allows producing a background probability map. Each pixel has its own PNN with the current color as input data and the output of the PNN gives the probability for this pixel to be part of the background. The weights of each PNN are updated according to the recent color history of the corresponding pixel. In order to avoid contaminating the background model with color pixels from foreground objects, coarse areas that include significant motion are excluded from the model update. These areas are found using a pixel change history technique [15]. Groups of "connected" pixels in the background probability map form blobs of potential interest (moving objects in the foreground) that are further analyzed. Figure 1 shows an example of a frame with its background/foreground segmentation.

Figure 1: Input frame (top), its motion map (bottom-left) and the background probability map (bottom-right).

2.2 Track management

For each frame, the corresponding background probability map is segmented into blobs. Blobs appearing in successive frames are fused together if there are similar in shape and color, in which case they form a track. Tracks represent objects of interest that are being tracked by the system and they are characterized by their shape, color, motion and an object class label (pedestrian, vehicle or other). Object class labeling is discussed in the next section.

At any given time, many active tracks are managed by the track management module. Blobs from the current frame are assigned to the most similar tracks. However, there may be numerous cases where such blob assignment fails: objects leaving the camera field of view or simply disappearing, confusion arising when two or more objects get together, bad blob segmentation due to image noise or environmental conditions (e.g. pedestrian walking behind a telephone pole). In such cases, extension of an active track is attempted using a bootstrap filter that is applied locally in order to find the foreground region that is the most coherent with the blob motion and color. If the bootstrap filter fails, the track simply terminates. Figure 2 shows an example of a situation where the bootstrap filter kicks in to disambiguate tracking.

During track evolution, object position and label are smoothed over time to increase identification robustness. Tracks with incoherent motion are rejected as they most probably refer to noise or stationary objects (e.g. tree leaves moving in the wind). Information about the currently active tracks, such as estimated speed and object label, helps guide the surveillance system to zoom in on relevant parts of the tracked objects such as faces (for pedestrians) and license plates (for vehicles).

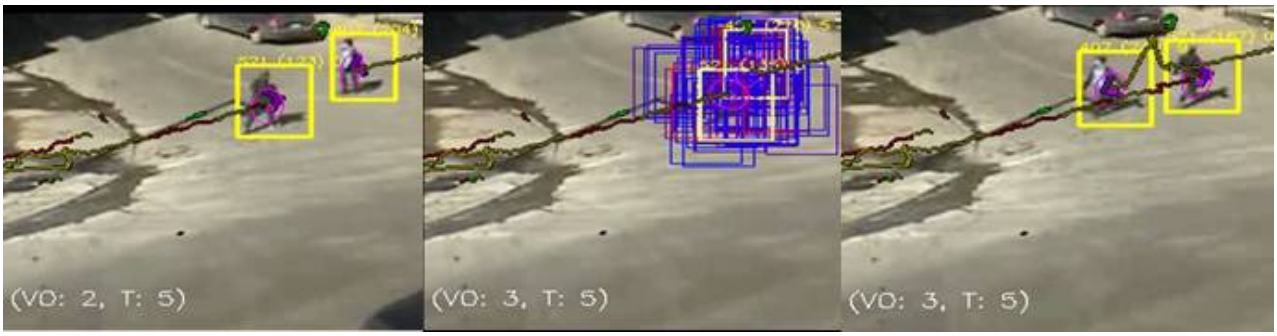

Figure 2 : Two pedestrians walking in opposite directions are being tracked; as they get near one another, the bootstrap filter kicks in to properly extend each track.

2.3 Object recognition

For each processed frame, extracted blobs are classified as being 'pedestrians', 'vehicles' or 'unknown' according to their geometrical properties. Many papers addressing this topic have been found in the literature, however it became apparent that most of the techniques proposed assume that pedestrians are imaged with a relatively high resolution (many tens of pixels high), which is not the case in this study. We turned our attention to scene-independent techniques (e.g. [1], [3]) which try to recognize pedestrians and vehicles in low resolution (or far-field) images.

A collection of features proposed by various authors were retained for analysis (Table 1). Evaluation was performed by first extracting these features from 4215 shapes of pedestrians and 2233 shapes of vehicles, manually segmented from typical low-resolution image sequences, and then training an Adaboost classifier with stump classifiers as weak learners. In this context, Adaboost was used as a feature selector capable of ranking the features by relevance. After training, three features were found to be the most discriminant: the orientation of the ellipse around the blob, the near horizontal/near vertical axis ratio, and the blob occupancy inside its oriented bounding box. Classification error on a validation set was around 3%. A fuzzy classifier was then designed. Simple membership functions based on the best features found and tuned for the 'pedestrian' and 'vehicle' classes were established so that whenever a blob's joint membership value for one class is higher that its joint membership value for the other class, it is assigned to this class provided that the value is higher that a minimum.

Table 1. Some features tested for pedestrian/vehicle recognition

Perimeter
Dispersedness = $(\text{Perimeter})^2$/area
Area
Y-normalized area = Area / blob_centroid.y
Area of convex hull
Ratio of axes of fitted ellipse
Occupancy = Area / (area of oriented bounding box)
Variation of area
Var(dispersedness) over n frames
Var(occupancy) over n frames
Min(occupancy) over n frames
Average speed (top part, bottom part of bounding box)
Y-Normalized average speed = speed / blob_centroid.y
Orientation of fitted ellipse

2.4 Light tracking

This component allows tracking specific objects despites changes in camera point of view. Unlike the wide field-of-view use case described above where the camera is fixed and uses background modeling to locate foreground objects, this operating mode allows single object tracking and it is meant to be used in true PTZ mode as the system tracks a single object of interest, activating the PTZ mechanism so as to keep the object in the middle of the image. Since the camera is constantly moving, background modeling is impossible, hence the term 'light tracking' to underline the fact that the tracker operates without much information except a color signature of the object to track. Light tracking is implemented using a particle filtering approach [12] based on color histogram information. Such information may come from two sources: the user can manually select an object to track, in which case the color content of the selected region of interest provides the reference signature, or surveillance system may automatically "get interested" in an object being tracked in wide FOV mode, in which case the reference signature is the color signature that comes from the associated track (Figure 3).

The selected flavor of particle filter is that of Pérez et al. [12] where color and motion are cues for tracking an object. In short, simple frame difference provides the motion cues and contributes to the proposal function according to which particles are drawn. Particle weighting is done by comparing the color histogram associated to the particle to the reference signature discussed earlier. Since many histogram comparisons are required, efficient histogram building is made possible by using an optimization "Integral histograms" technique [14]. Computing integral histograms requires a lot of memory (basically an entire histogram per pixel), so histogram size must be kept low. A comparison with standard histogramming shows that the approach yields a significant gain if the number of particles is high (~500) and the regions of interest bound to the particles have a significant size (in the thousands of pixels). In situations where the size of these regions of interest is small with respect to frame size (e.g. Figure 3), tracking is close to real time despite the potentially high number of particles used.

Figure 3 : Tracking of a pedestrian in "light" mode. The green box around the pedestrian represents the best particles of the particle filter.

3. OPERATIONAL RESULTS

The background segmentation component runs in real time (30 fps) on a desktop Pentium 4. The whole module runs at about 20-22 fps when there is low activity in the scene but can slow down when there are many objects to follow.

However, empirical observations in real situations have shown that:

- Specular reflections of the sunlight may generate false object hypotheses but they are easily rejected by the track management component.

- Dynamic background objects such as flags or trees moving in the wind end up being ignored by the surveillance system because of their motion incoherence during tracking.
- Pedestrians and vehicles in simple situations are well recognized by the system.
- The system has been tested in various weather conditions with good success: sunny/cloudy days, with or without rain, during summer or winter.

Figure 4 shows some tracking results using a typical surveillance camera in a typical setting. Yellow and red traces illustrate the path followed by pedestrians and vehicles respectively. The dense set of paths in the upper part of the image is explained by the presence of a highway in this area. Note the object reflections on building windows, which are either classified as vehicles or unknown objects (green lines).

Figure 4: Two examples of detection, tracking and object recognition in the wide field mode on a parking lot. Human trajectories are in yellow, vehicles in red.

Figure 5 shows additional results using a standard Handycam camcorder. Again, yellow and red lines correspond to tracked pedestrians and vehicles, while green lines are shown when objects have not been recognized reliably.

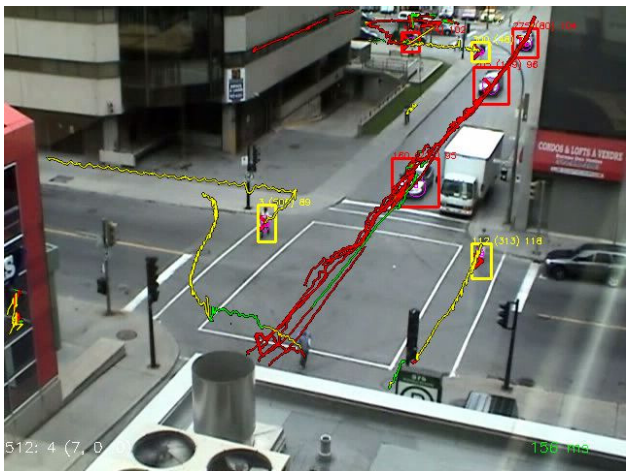

Figure 5: Example of detection, tracking and object recognition in the wide field mode

Figure 6 shows examples of operational results after the final integration within the VideoStream's VST OneTrack system. The main system communicates with the module through an API that enables on-line modifications of detection and tracking characteristics during the operational mode for tests and tuning.

Figure 6: Two examples of fully operational uses after system integration. On the left-hand side, three moving objects detected and tracked: 2 vehicles and 1 pedestrian in the far field by nearby the building. On the right-hand side, a far field moving car detected in the parking (even through glass reflection) and tracked in light mode.

4. CONCLUSION

We have presented an overview of a practical object tracking modules that was developed for a commercial PTZ video surveillance system. The module allows autonomous real-time object detection, tracking and recognition in uncontrolled outdoor environments. The final system detects moving objects, identify the presence of pedestrians or vehicles, track them, and optionally zoom on them. The system achieves good performances for most operational uses despite the numerous sources of complexity present in uncontrolled environments.

Obviously, there is still room for improvements as fully automatic video surveillance systems in uncontrolled environment will always be a challenge. For instance,

- Low-cost cameras may generate unstable color patterns on edges of man-made structures such as building windows, which may fool the background segmentation component even though temporal filtering helps attenuate such noise.
- Foreground/background segmentation may be very sensitive to slight camera movement or vibration if the outdoor camera is badly shielded from the wind.
- Complex scenarios such as groups of pedestrians or vehicles as well as shadows are still major issues (especially without any prior information about the geometry of the scene and with real-time constraints).

ACKNOWLEDGMENTS

This work was supported in part by Precarn Inc. (www.precarn.ca) through the Precarn-CRIM Alliance Financing Program and the "Ministère du Développement Économique de l'Innovation et de l'Exportation (MDEIE)" of "Gouvernement du Québec".

This project wan the award of the "2006 Best Technological Innovation" of the "Fédération Informatique du Québec (http://www.fiq.qc.ca/FRANCAIS/octas-2006/2006_Laureats_innovation_techno.html).

REFERENCES

1. B. Bose, "Classification of Tracked Objects in Far-Field Video Surveillance", MSc Thesis, MIT, 2004
2. A. Branzan, D. Laurendeau, S. Comtois, D. Ouellet, P. Hébert, A. Zaccarin, M. Parizeau, R. Bergevin, X. Maldague, R. Drouin, S. Drouin, N. Martel-Brisson, F. Jean, H. Torresan, L. Gagnon, F. Laliberté, "MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras", In Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, August 20-24, 2006
3. L. Brown, "View independent vehicle/person classification", VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks, 2004, p. 114-123
4. R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, L. Wixson, "A system for video surveillance and monitoring", Technical Report, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2000
5. R. Collins, A. Lipton, H. Fujiyoshi, T. Kanade, "Algorithms for cooperative multisensor surveillance", Proceedings of IEEE, Vol. 89, No. 10, pp. 1456-1477, 2001
6. J. Fan, D. K. Y. Yau, A. K. Elmagarmid, W. G. Aref, "Automatic Image Segmentation by Integrating Color-edge Extraction and Seeded Region Growing", IEEE Transactions on Image Processing, Vol. 10, No. 10, p. 1454-1466, 2001
7. L. Gagnon, F. Laliberté, S. Foucher, A. Branzan Albu, D. Laurendeau, "A System for Tracking and Recognizing Pedestrian Faces using a Network of Loosely Coupled Cameras", SPIE Defense & Security: Visual Information Processing XV (SPIE #6246), pp. 0N-1-0N-9, 2006
8. I. Haritaoglu, D. Harwood, L.S. Davis, "W4: Real-Time Surveillance of People and Their Activities", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.8, pp. 809–830, 2000
9. O. Javed, M. Shah, Tracking and Object Classification for Automated Surveillance, ECCV`2002, European Conference on Computer Vision, Copenhagen, Denmark, 2002
10. O. Javed, Z. Rasheed, O. Alatas, M. Shah, "Knight^{M}: A Real Time Surveillance System for Multiple Overlapping and Non-Overlapping Cameras", IEEE conf. on Multimedia and Expo, Special Session on Multi-Camera Surveillance Systems, Baltimore, 2003
11. A. Mittal, N. Paragios, "Motion-Based Background Subtraction Using Adaptive Kernel Density Estimation". Proc. CVPR, pp. 302-309, 2004
12. P. Pérez, J. Vermaak, A. Blake, "Data fusion for visual tracking with particles", Proc. IEEE, 92(3):495-513, 2004
13. C. Stauffer, W.E. Grimson, "Learning patterns of activity using real-time tracking", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp. 747-757, 2000
14. F. Woelk, I. Schiller, R. Koch, "An Airborne Bayesian Color Tracking System", IEEE Intelligent Vehicles Symposium, Las Vegas, USA, June 6-8, 2005
15. T. Xiang, S. Gong, D. Parkinson, "Autonomous visual events detection and classification without explicit object-centred segmentation and tracking", In Proceedings of British Machine Vision Conference, volume 1, pages 233-242, Cardiff, UK, September 2002