MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras

Alexandra Branzan Albu[1], Denis Laurendeau[2], Sylvain Comtois[2], Denis Ouellet[2], Patrick Hebert[2], Andre Zaccarin[2], Marc Parizeau[2], Robert Bergevin[2], Xavier Maldague[2], Richard Drouin[2], Stephane Drouin[2], Nicolas Martel-Brisson[2], Frederic Jean[2], Helene Torresan[2], Langis Gagnon[3], France Laliberte[3]

[1]*University of Victoria,* [2]*Laval University,* [3]*Centre de Recherche Informatique de Montréal (*CRIM*), Canada*

aalbu@ece.uvic.ca, laurend@gel.ulaval.ca, langis.gagnon@crim.ca

Abstract

MONNET is a visual surveillance system for tracking pedestrians over extended premises. The MONNET system is composed of intelligent nodes, which exchange information on the individually tracked pedestrians in an asynchronous manner. Each node in MONNET builds an appearance model for every observed pedestrian and compares it with models received from other nodes. The compact appearance models based on colour cues and face biometrics are stored locally on each node. The system is dynamically reconfigurable since its design allows for adding/removing nodes in a simple manner, comparable to the 'plug and play' technology. MONNET also contains an optional 'observer' node for interactive data visualization. This node displays a user interface which allows a human operator to observe and to interact in real-time with the distributed tracking process. MONNET was extensively tested with and without user input, and it is able to function correctly in both modes.

1. Introduction

Motion analysis and representation have received lately increased attention from researchers in Computer Vision. Their interest is motivated by the underlying theoretical challenges specific to video understanding, and also by the wide spectrum of potential applications in surveillance. The study of human motion from video sequences is mostly driven by applications in security, such as gait recognition [1], and detecting suspicious activities [2]. Other related applications have also emerged, such as the visual monitoring of senior wellbeing [3].

Regardless of the particular application context, gathering reliable information about the multiple moving objects in the scene is a key issue in the design of intelligent systems for visual motion analysis. Therefore, prior to implementing any high-level activity understanding mechanisms, a multiple-camera surveillance system is to 'watch' over each pedestrian moving across the fields of view of its cameras by establishing a correct inter-camera correspondence.

Several challenges are critical to the accomplishment of this primary task, such as inter-sensor communication, temporary occlusion, variable pose and depth, as well as the simultaneous tracking of several pedestrians. To address these challenges, several system architectures, tracking approaches, and motion representations have been proposed recently.

A system allowing a human operator to monitor activities over a large area using multiple calibrated cameras and a geospatial site model was proposed in [4]. Their approach was based on image correlation mapping, and on the computation of the 3D location on the site model. Inter-sensor communication consisted in a simple 'handing off' mechanism between sensors situated along the object's trajectory.

The decentralized architecture described in [5] used multiple calibrated cameras ('a forest of sensors') for learning patterns of activities from motion observation. A basic assumption for learning consisted in the preservation of the object identity throughout the tracking process.

A wide area surveillance system implemented via a client-server architecture was proposed in [6]. They used uncalibrated cameras with overlapping and/or non-overlapping fields of view (FOVs), and trained

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

their system for learning 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. The use of an appearance model is suggested as a possible improvement for the tracker performance.

One may observe that most of the recently reported surveillance systems require overlapping FOVs, off-line camera calibration, and spatial site models. In addition, the client-server architecture is preferred to a decentralized architecture since it controls better the sequence of processes involved in motion analysis. However, overlapping FOVs, camera calibration and spatial site models may not be realistic system features when monitoring extended premises. Also, the client-server architecture, although reliable, is rigid and not resilient with respect to the failure of its critical components (i.e. server failure).

To address such current limitations in surveillance systems, this paper proposes a new, decentralized systemic approach. The MONNET system is composed of intelligent nodes, which exchange information on the individually tracked pedestrians in an asynchronous manner. Any node in MONNET builds an appearance model for each observed pedestrian and compares it with models received from all other nodes. The compact appearance models, based on colour, shape, and face information, are stored locally on each node. The system is dynamically reconfigurable since its design allows for adding/removing nodes in a very simple manner, comparable to the 'plug and play' technology. After a node is added to the network and begins tracking, its appearance models become immediately available to the other nodes. The removal of a node has no effect on the other nodes in the network, which simply continue their observations. The MONNET system contains also a special 'observer' node dedicated to interactive data visualization. This node consists of a user-friendly interface which allows a human operator to observe and to interact in real-time with the distributed tracking process. It is worth mentioning that the presence of the 'observer' node is optional; thus MONNET can also function without any human intervention. As a matter of fact, any MONNET node can act as an observer node.

The rest of the paper is structured as follows. Section 2 describes the proposed systemic approach for pedestrian surveillance. The experimental results are discussed in section 3. Section 4 draws the conclusions and describes some future work directions.

2. Proposed approach

2.1 Task description for processing nodes

The MONNET system is composed of several processing nodes and one optional 'observer' node. The minimal configuration of a processing node consists of a computation unit connected to a video camera and an optional infrared camera. This section provides the task description for a typical processing node. The sequence of the main tasks to be performed in real-time by each processing node is as follows.

a) Video acquisition. For a specified node, the user is able to select among the following options: i) acquisition in the visible spectrum only, and ii) synchronized acquisition of infrared and visible data. It is thus possible to customize each node according to the particular environmental conditions related to the FOVs of its cameras. IR sequences provide enhanced contrast between the human body and the background in a low-lit room, while visible sequences convey rich information about the colour, texture and shape of the pedestrians when proper lighting conditions are available. Moreover, there are situations where a fusion between IR and visible data improves significantly the performance of figure-ground separation. Frame samples from a simultaneous IR/visible spectrum acquisition are shown in Fig. 1.

Figure 1. A pair of 'visible' (left) and IR (right) frames

A sequence of preprocessing steps is necessary to correct vignetting, fixed pattern noise and dead pixels and also to perform temperature calibration [7]. For simultaneous IR-visible acquisition, temporal synchronization is also necessary.

b) Figure-ground segmentation.

If acquisition provides only 'visible' data, then a pixel-based statistical background subtraction algorithm [8] is used. The background and the moving cast shadows are represented with Gaussian mixture models (GMMs). This representation is robust with respect to complex and changing illumination patterns.

In case of simultaneous IR-visible acquisition, a data fusion algorithm is implemented [7]. First, figure-ground segmentation is performed independently on IR and visible sequences and results in objects

corresponding to pedestrians. The quality of the segmentation is measured by confidence ratios attached to each object on a frame-by-frame basis. The fusion algorithm uses the confidence ratios to establish a dynamic *master-slave* relationship between IR and visible segmentation results. The *master* object is defined by the highest confidence ratio in the current frame, and it is used to refine and correct the segmentation of the corresponding *slave*.

c) Tracking is based on a five-point model (head, hands and feet) [9]. Feature points are first detected from shape, colour and motion information and then tracked independently. The issue of body-part self-occlusion during walking is handled by integrating motion correspondence and optical flow techniques.

d) *Pedestrian-representative appearance models* are essential for establishing a correct inter-node correspondence in the global tracking process. Each node builds and updates regularly one appearance model for each currently tracked pedestrian. These models are broadcast to all other nodes in the network for comparison and matching. Two approaches for generating appearance models are currently implemented in the MONNET prototype. The first one allows for building a global, silhouette-based model [10], while the second describes face biometrics and is applicable only for frontal face poses. To build the global appearance model, the bounding box of the silhouette is split into three regions according to user-specified proportions with respect to the total height. Default values are: head region 20%, trunk region 50%, and legs region 30 %. Each region is described by a feature vector, which contains information about the spatial and spectral distribution of dominant colours [10]. The comparison of two global appearance models is performed region-wise by using a quadratic colour histogram distance measure [10].

e) Inter-node communication protocol.

The MONNET system is configured as a wireless network of collaborative nodes, able to exchange information about the tracked pedestrians in an asynchronous manner. The open source Bonjour[tm] protocol from Apple Computer Inc. is used for a dynamic discovery of the collaborating nodes. Working with Bonjour[tm] is ideal for ad-hoc, zero-configuration networking based on the standard IP protocol.

A Bonjour[tm] information sharing process typically involves two partners: one requesting a certain service (i.e. 'the client'), and one offering that service (i.e. 'the server'). The novelty in MONNET is the dual role assumed by each of its processing nodes, which act as both 'client' and 'server'. Therefore, they are able to broadcast/receive pedestrian-related information to/from other nodes. Moreover, nodes can easily be added or removed from the system, without temporarily affecting the global performance of the system. MONNET can be defined as a private network of nodes. Every new node (i.e. a 'server') is to register via Bonjour[tm] as a new member of the MONNET subnet; thus, it becomes able to broadcast its appearance models to all other MONNET members by using multicast addresses. A new node in MONNET has also 'client' privileges, and it is thus provided with a list of all active, broadcasting 'servers'. The removal of a node in MONNET is signaled to all other members by an update of the above-mentioned list.

f) Intra-node software management

Node-specific data processing is organized into a modular structure, as shown in Fig. 2. Moreover, the computational complexity and the diversity of the processing tasks (background subtraction, tracking, and pedestrian modeling) require a flexible integration strategy. This strategy is also responsible for the coordination of parallel processes and for the management of computational resources.

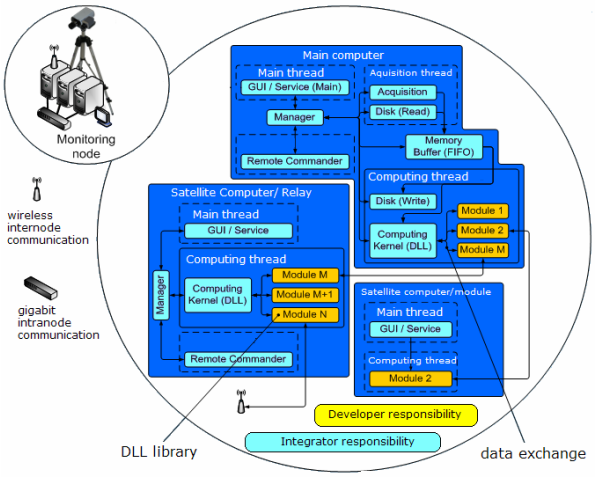


Figure 2. Intra-node software management

As shown in Fig. 2, the physical configuration of a processing node may consist in several computing units. The main computer controls the video acquisition process and the optional graphical user interface (GUI). The computing thread may be distributed across the main computer and additional satellite computers, if the task complexity leads to the overloading of the main computing unit. Satellite computers may be used for the distributed implementation of a single module (see Fig. 2, lower right) or for implementing several modules (see Fig. 2, lower left). Since intra-node communication involves real-time video data exchange between modules, the

computing units in a node are connected via gigabit communication links.

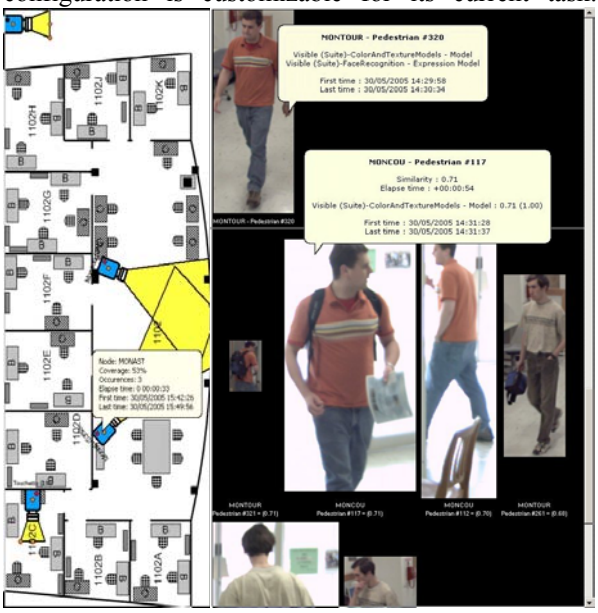
The modules are dedicated to processing algorithms, which are implemented as Dynamic Link Libraries (DLLs). For software management purposes, the algorithmic implementation is performed and controlled by the developer of the respective module. The three basic processing tasks, namely background subtraction, tracking, and model generation may have more than one algorithmic solution, and thus more than one corresponding module. For example, the appearance model may be built using global colour cues [10], or face recognition techniques when face detection is feasible. While default modules are specified in the initial system configuration, the user is able to select the most appropriate algorithms (modules) for the task at hand via the visualization interface. The remote commander block allows for updating the node configuration according to user specifications received from the visualization node.

2.2 Interactive visualization interface

The MONNET system is equipped with one optional 'observer' node connected to all processing nodes via the customized Bonjour[tm] protocol. This node allows for the interactive visualization of the monitoring process, which is a novelty in automated surveillance systems. Since MONNET is a multi-task surveillance system, the user is allowed to specify a task of interest, and to visualize specific information related to this task. For example, if the user wants to gather information about a certain person already detected by a node, this person is selected to be 'the active pedestrian'. Interactive visualization gives access to on-line, real-time monitoring data (e.g. 'the *active* pedestrian enters *now* in the FOV of node *n*'), to a log of events having already occurred (e.g. 'the *active* pedestrian has previously been seen by node *i*, *j*, and *k*; his estimated trajectory and temporal information about time spent in each node's FOV are available'), and to a database of static images representative for the active pedestrian (see Fig. 3).

A second task performed by MONNET is related to the global outcome of the distributed monitoring process (number of pedestrians tracked in a certain time interval by each processing node), and allows for displaying information about pedestrian traffic.

The visualization interface also allows the user to select the desired algorithms and their parametric configuration for the task of interest. As specified in section 2.1, several algorithms are available for each processing step: background subtraction, tracking, and pedestrian modeling. Thus, the MONNET configuration is customizable for its current task.



a) b)

Figure 3. Visualization interface: a) 'Room map' window showing node location and activity; b) 'Pedestrian viewer' window

3. Experimental results

A prototype of MONNET was built using four nodes equipped with infrared and visible cameras and observing an extended indoor environment (see Fig. 3a). The system used Pulnix TMC6700CL video cameras with a 12 mm lens, and Radiance 1 Amber Engineering thermal infrared cameras with a 25 mm lens. During the acquisition process, the infrared and video streams were synchronized, and the FOVs of each pair of infrared and visible cameras were almost overlapping. The final configuration of a processing node was distributed on two PCs with Pentium IV processors at 3.4 GHz and 1GB of RAM (from which only 200 MB were in use by the application). Wireless communication between nodes adopts the 802.11b standard. The main computer (see Fig.2) performed the background subtraction and tracking, while the satellite computer was used for pedestrian modeling.

The global four-node system functioned in real-time at a frame rate of 10 Hz, with input IR and visible images of size 640x480. Most of the computation time (approx. 95%) on the main computer was required by the background subtraction algorithm. Pedestrian modeling is performed on the satellite computer and consists of two independent approaches: silhouette-based modeling using colour cues [10], and face

biometrics [11]. A face model is built only when the face pose and distance to the camera are satisfactory. Building face models is a computationally expensive process, functioning at a 1Hz frame-rate. Therefore, if face analysis is feasible, the face model is built and broadcast to other nodes at every 10 seconds. However, colour-based appearance models can be built and broadcast at 10 Hz.

Comparison of colour-based appearance models broadcast by two nodes gives good results (average precision of 91% for a recall of 80%) when lighting conditions are similar in the FOVs of both nodes. The inter-node comparison of models based on face biometrics scored surprisingly lower than the comparison of colour-based models. The most probable reasons for this result were acquisition noise, the small size and the low resolution of face images.

The system functionality has been tested extensively over a two-month period. Experiments were carried out over several days without interruption on a wireless MONNET system composed of 5 nodes (4 processing nodes and 1 observer). In addition, the functionality of a two-node setup has been successfully tested in a crowded indoor public environment (hotel lobby). These experiments proved that MONNET is able to operate continuously and to provide reliable data on more than 200 tracked pedestrians.

4. Conclusions

The wireless MONNET system demonstrates the feasibility of deploying a complex vision system that processes visual data efficiently so as to distribute this data on a network without requiring high performance communication links. The modular design of the system framework allowed for conducting autonomous research on a variety of aspects of human motion analysis: background subtraction, data fusion between invisible and infrared image, tracking, and pedestrian modeling. Moreover, the implementation of the modules via DLLs (Dynamic Link Libraries) helped to protect proprietary information for collaborators, since no source code was needed for the system integration.

The design of an 'observer' node dedicated to an interactive visualization interface represents a novelty in surveillance systems. This visualization interface allows for a rapid retrieval of information related to a specified pedestrian, for gathering information related to the pedestrian traffic distribution, and also for customizing the system configuration. MONNET was extensively tested with and without user intervention, and it is able to function correctly in both modes. Ongoing work is focused towards improving the performances of the comparison of pedestrian models. A motion model describing the gait of the tracked pedestrian is currently under development.

Acknowledgements

The authors acknowledge the financial support of PRECARN Inc., NSERC, and DRDC-RDDC-Valcartier.

References

[1] J. Boyd and J. Little. "Biometric Gait Recognition" in Summer School on Biometrics, Alghero Italy, 2003, Eds: M. Tistarelli, J. Bigun, E. Grosso, Lecture Notes in Computer Science, Vol. 3161/2005, Springer, pp. 19-42, 2005.

[2] N. Rota and M. Thonnat, "Video sequence interpretation for visual surveillance", in *Proc. of 3rd IEEE Int. Workshop on Visual Surveillance*, pp. 59-68, July 2000.

[3] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts and shadows in video streams", *IEEE Trans. Patt. Anal. Mach. Intell*., 25(10), pp. 1337–1342, 2003.

[4] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance", Proceedings of IEEE, 89(10), pp. 1456-1477, 2001.

[5] C. Stauffer and W.E. Grimson, "Learning patterns of activity using real-time tracking", *IEEE Trans. Patt. Anal. Mach. Intell. ,* 22(8), pp. 747-757, 2000.

[6] O. Javed, Z. Rasheed, O. Alatas and M. Shah, "M-Knight: A Real Time Surveillance System for Multiple Overlapping and Non-Overlapping Cameras", IEEE Conf. on Multimedia and Expo, Baltimore, 2003.

[7] H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hebert, and X. Maldague, "Advanced surveillance systems: combining video and thermal imagery for pedestrian detection",Proc. of SPIE, vol. 5405, pp. 506-515, 2004.

[8] N. Martel and A. Zaccarin, "Moving Cast Shadow Detection from a Gaussian Mixture Shadow Model", Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), (San Diego, CA), pp. 643-648, 2005.

[9] F. Jean, R. Bergevin and A. Branzan Albu, "Body Tracking in Human Walk from Monocular Video Sequences", in IEEE *Canadian Conference on Computer and Robot Vision (CRV)*, (Victoria, BC), pp. 144-151, 2005.

[10] M. Lantagne, M. Parizeau and R. Bergevin, "VIP: Vision tool for comparing Images of People", Proc. of the 16th IEEE Conf. on Vision Interface, pp. 35-42, 2003.