TOWARDS MULTIMODAL SURVEILLANCE FOR SMART BUILDING SECURITY

G. Amato, P. Barsocchi, F. Falchi, E. Ferro, C. Gennaro, G. R. Leone, D. Moroni, O. Salvetti, C. Vairo

Institute of Information Science and Technologies (ISTI)
National Research Council of Italy (CNR)

ABSTRACT

The main goal of a surveillance system is to collect information in a sensing environment and notify unexpected behavior. Information provided by single sensor and surveillance technology may not be sufficient to understand the whole context of the monitored environment. On the other hand, by combining information coming from different sources, the overall performance of a surveillance system can be improved. In this paper, we present the Smart Building Suite, in which independent and different technologies are developed in order to realize a multimodal surveillance system.

Index Terms— smart building, sensor networks, smart cameras, surveillance, pervasive computing, face recognition

1. INTRODUCTION

The concept of "smart building" is very wide, as it embraces all the building's components, from technology to services, in order to fully support the people living or working there [1]. Optimizing spaces utilization, building life costs and surveillance system is part of the process of making "smart" a building. The life costs are mostly related to operations and maintenance; as an example, the tendency to be unconcerned about energy saving is more prominent in office than at home; thus, in a smart building it is necessary to provide automatic energy saving services. Surveillance is another important aspect in order to both avoid waste of money or destruction of properties due to thefts and for the safety of the people [2]. Unfortunately, today the need of good surveillance is getting higher due to the high level of criminality, especially in sensitive areas, government offices, public places, and even at home.

Security and intrusion control concerns are the motivating factors for the deployment on the market of many video surveillance systems, such as surveillance cameras, which are closed-circuits that need a command and a control center to monitor all the activities using cameras, or radio frequency identifications (RFID), which use radio waves to automatically identify persons or objects by means of RFID transponders and readers. In this context, the information provided by a dedicated surveillance technology may not be sufficient to automatically identify the intrusions in a monitored environment. Combining several sources of information coming, for example, from both a wireless sensor network and cameras, improves the overall performance of a surveillance system. Indeed, ambient sensors (like RFID, PIR, noise, etc...) may generate false positive readings that can be reduced by fusing these measurements with data coming from surveillance cameras.

In this paper, we present an envision enhanced surveillance system for smart building, in which independent and different technologies, currently implemented and deployed, are integrated and coordinated by a high-level central reasoner in order to provide a more robust and efficient surveillance system for a smart building. Such an enhanced surveillance system will be able both to monitor what happens outside the buildings and to detected if a violation occurs. We also describe the different technologies, based on sensors and cameras, and the software solutions we implemented and that we aim to integrate in order to realize the surveillance scenario described above. The paper is organized as follows. Section 2 describes the system and its functionalities. Section 3 reports the technologies used in developing the system; Section 4 presents the methodologies adopted and describes the algorithm developed. Finally, Section 5 concludes the paper.

2. SYSTEM DESCRIPTION AND FUNCTIONALITIES

The Smart Building Suite is a distributed smart surveillance system with the capabilities to notify real time alerts, to store and to analyze big amount of data coming from a large number of IP nodes. The infrastructure is constituted by a network of wireless sensors installed in the offices of a building in the CNR Research Area in Pisa and the embedded smart cameras (ESC) installed both on its roof and in front of the offices' entrance, thus monitoring the outside environment and inside the offices. The wireless sensor network (WSN) is also used to monitor the energy consumptions of the offices, taking recovery actions to contain the energy waste, while the ESCs are embedded computers equipped with camera modules. The main functionalities offered by the system are:

Presence detector. Data collected from the various sensors in an office, crossed each other, allow to understand whether or not the office is occupied, thus creating a sort of

"virtual presence sensor";

Real-time alert. To each office, a set of authorized people images is associated; when a non-authorized person is detected, a notification, and possibly an alarm, is raised. Moreover, the system also provides motion detection capabilities, which allow monitoring predesignated regions of the video frame;

Video-recording. Video streams captured by cameras are recorded and stored. It is possible to define strategies to control the recording process both in time and space (e.g., on a per-camera basis);

Analysis and understanding. The video streams acquired by cameras are analyzed to extract information useful for searching video shots and taking decisions. Two types of information are be extracted: (1) low level information, consisting of visual features (global and local visual descriptors of areas of the frames, face features, etc.); (2) high level information, consisting, for instance, in tracking information of objects, people, or occurrence of unexpected visual anomalies;

Data Management. It is possible to access the faces' data base, to browse all videos and historical data (motion detection events, unknown persons, etc) and to search for video shots to retrieve the occurrence of an entity (e.g., a car or a person), thereby determining the paths followed by the entity in the past or its current location. Search is based on the information extracted from the video streams during the analysis and the understanding phase.

To achieve high scalability and efficiency we adopted an event-based communication where the events are generated by peripheral smart subsystems that locally elaborate the raw data and feed the database only if something happens.

The use case diagram depicted in Fig.1 shows the interaction among the different actors involved in the system: the ESC and the WSN are the peripheral smart subsystems that send events of interest (face, motion or presence detection); every time an event occurs, the Activity Monitor (AM) decides if it is the case to trigger an alarm depending on the security parameters active in that moment. The Operator (typically a person of the security staff) receives the alarm via monitor, email or text; he can browse all the historical data (events, video, etc.), manage the faces' database and do searches according to the information provided by the High Level Reasoner (HLR), which is an A.I. agent that performs the analysis and understanding functionality described before.

3. INVOLVED TECHNOLOGIES

Raw sensor data and video streams are processed by two smart peripheral subsystems: the embedded smart cameras and the wireless sensor network, which are in charge to detect the events of interest and to communicate with the central server. In the following paragraphs we briefly describe the technologies involved in these tasks.

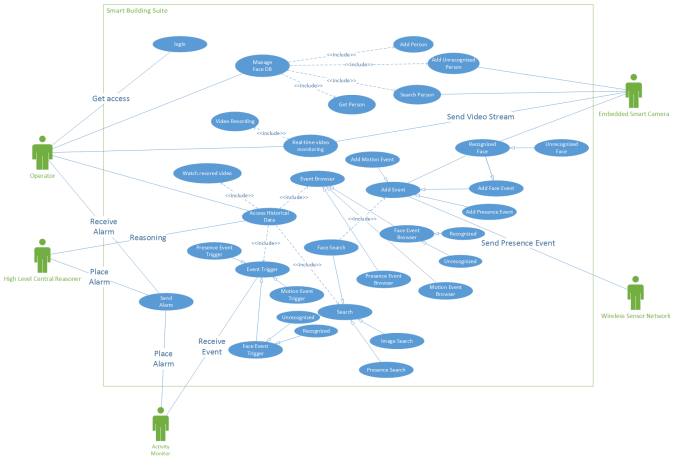

Fig. 1. The use case diagram.

3.1. Embedded smart camera

The Embedded Smart Camera features an embedded Linux architecture for providing enough computational power and ease of programming. The custom printed-circuit board has been designed in order to have the maximum flexibility of use while maximizing the performance/consumption ratio. A good trade-off has been achieved by using a Freescale CPU based on the ARM architecture, which integrates a Power Management Unit (PMU), in addition to numerous peripheral interfaces, thus minimizing the complexity of the board. The average consumption is less than 500mW, measured at the highest speed (454MHz). The image sensor is a low cost device compliant with USB Video Class device (UVC), which offers good image quality also with the very low illumination, as during the night. The board and imaging sensor are housed into an IP66 shield.

Several computer vision algorithms have been envisaged and ad hoc ANSI C implementations are particularly suitable for the embedded camera prototype and have shown to provide good results in traffic flow monitoring [3]. Nevertheless, a porting of OpenCV[1] libraries has been produced, which makes directly available on the embedded camera a plethora of state of the art computer vision methods for disparate applications.

3.2. Wireless Sensor Network

We developed a long-term monitoring system, composed by: i) a distributed ZigBee WSN, ii) a middleware communication platform, and iii) an occupancy detection algorithm able to infer the occupancy status of the monitored room. Each sensor node of the WSN can aggregate multiple transducers,

[1] http://opencv.org/

such as humidity, temperature, current, voltage, Passive Infrared (PIR), and noise detector [4, 5]. Each node of the WSN is connected to a ZigBee Sink, which provides IoT connectivity through the IPv6 addressing. The ZigBee choice is driven by several technology characteristics, such as ultra low power consumption, use of unlicensed radio bands, cheap and easy installation, flexible and extendable networks, integrated intelligence for network set-up and message routing. In order to measure the energy consumption in a room, we evaluate the values of current and voltage waveforms at the same time instant; for this scope, driven by the need to operate within existing buildings without the possibility of changing the existing electrical appliances, we used a current and voltage transformer. We also installed a PIR and a noise detector into a single node. All these measured values help in deciding whether or not someone is in the office, in order to apply appropriate decisions on the energy saving in that specific room or on surveillance policies. As an example, in an office where nobody is present, lights and any other electric gear, a part from the computer, are automatically switched off, if on. Sensor data collected by the deployed WSN are stored in a NoSQL document-oriented local database, such as MongoDb. In order to provide both a real-time view of the sensor data and a configuration interface, we implemented a web interface, named WebOffice, that runs on JBoss 7.1; it implements the JavaEE specification, is free and open-source (Fig.2).

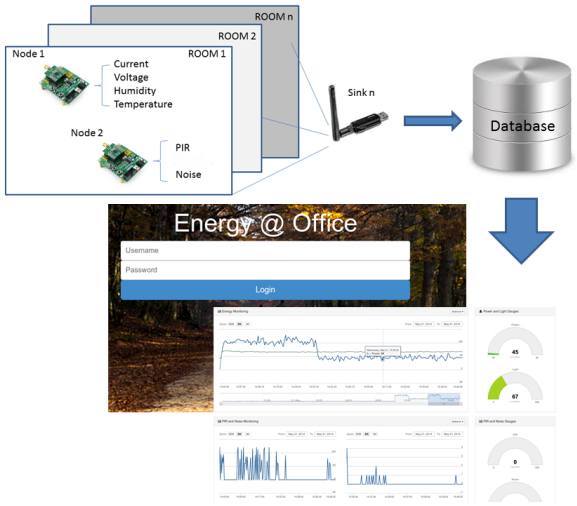

Fig. 2. The proposed long-term monitoring system.

4. METHODOLOGIES

The algorithms used to run with the hardware introduced in previous Section are shortly described in the following paragraphs.

4.1. Motion detection and tracking

In this paragraph, we are about to briefly discuss the outdoor surveillance of the building, which relies on the Embedded Smart Cameras presented in Section 3.1. The main characteristic of a smart surveillance system is the ability to autonomously understand if someone or something is present in the observed scene and where the subject is going. Therefore, conventional pipeline always starts with *motion detection*, which is the ability to understand if anything is moving in the target area, then it moves forward to *classification* to identify what type of moving item has been detected and, if it is of any interest, a *tracking* phase is started to understand where the detected object is going. Motion detection with a visual sensor is based on the analysis of the difference in the frames of the video stream. This technique is called background subtraction. The simplest algorithm takes into account only the intensity of two adjacent frames and triggers a motion if somewhere in the picture there is a change that overcomes some predefine threshold. This is very fast but also inaccurate on outdoor scenario because of the false positive due to sudden illumination change, shadows, trembling trees, clouds, etc. Many different approaches have been presented and evaluated [6, 7, 8]. The key point of all the techniques is to build a background reference model that is robust to errors resulting from anything that is not a real item moving in the scene. Many algorithms are computationally expensive and not suitable for our lightweight hardware. We adopted the solution successfully tested in urban traffic scenario [3] and railways surveillance [9]: a background model updated with a dynamic learning factor, which can be set to zero to prevent main errors due to sudden illumination changes. This is a good compromise between robustness and computational efficiency.

The classification step simply analyzes the size of the detected object. It discerns between vehicles and not-vehicles: this is because, for security reasons, we are mainly interested in walking people that could enter the facility overstepping the fence, while vehicles should mandatory enter and exit through the main entrance.

The tracking of moving objects is achieve by using a Multi Object Tracking Kalman Filter, which is commonly used in vision based tracking applications as it is an optimal recursive algorithm, easy to implement, and computationally not demanding. Such a filter represents an efficient solution to the general problem of estimating the state of a discrete-time controlled process. Details of the implementation can be found in [10].

4.2. Face re-identification

The visual person re-identification system has the purpose of monitoring the access to sensible or restricted areas of the building. In order to recognize the face of the person entering the controlled area we apply Deep Learning techniques combined with k-nearest neighbor (kNN) [11] classification

algorithms to perform the face re-identification task and to determine whether or not the person is allowed to get access to that area. Deep Learning [12] is a branch of Machine Learning that allows a neural network, composed of large number of layers, to learn representation of input data with increasing levels of abstraction. It provides near-human level accuracy in performing tasks like image classification [13], object detection [14], object recognition [15], speech recognition [16], parking monitoring [17], face recognition [18], and more.

In order to determine whether or not a person is allowed to enter into an specific area, we built a training set that contains ten pictures of the faces of all the people authorized to access that office. These pictures have been taken off-line before deploying the system and they are linked to a label identifying the person. The pictures used to build the training set; the person is captured in different positions (face rotation of a few degrees with respect to the central axes, both horizontally and vertically) and with different facial expressions.

The system uses a special type of neural network called Convolutional Neural Network (CNN), which was pre-trained to recognize people faces. In particular, we used the output of the neurons (i.e., deep features) of the VGG_Face [18] as a representation of the face. VGG_Face is a CNN provided by the Visual Geometry Group of the University of Oxford, composed of 16 convolutional layers and trained to recognize faces with a dataset composed of 2,6 million faces belonging to 2,6 thousands different persons. We take the output of the fully-connected 7th layer of the CNN (fc7), which is a high level representation of the input face, and we compute the L2 distance between the query face deep feature and all the deep features in the training set, in order to perform the kNN classification and to obtain the identity of the person whom the query face belongs to.

The person re-identification system is composed of two hardware components: an embedded camera connected to the network and a PC that processes the acquired images and performs the person re-identification task. The camera is placed in front of the office entrance, in order to have a clear view of the face of the entering person, and sends the acquired images to the PC that executes the re-identification algorithm. In particular, for each captured image, a first phase of face detection is executed in order to determine and crop the face from the whole image. We used the OpenCV implementation of the Viola-Jones [19] algorithm for the face detection. Each detected face is then processed by the VGG_Face network in order to extract the deep feature that will be used to perform the kNN similarity search. If the distance of the person returned by the kNN search is under a given threshold, then the entering person is among the allowed ones. Otherwise, the system raises a notification of an unauthorized person entering a monitored office.

4.3. Virtual occupancy sensor

With the term "occupancy", we refer to the possibility of determining whether or not a room is occupied by a person along with a time interval. We determine the occupancy by means of a fusion strategy that combines the sensing information gathered from the WSN described in Section 3.2.

The WSN we designed periodically gathers data from the motion sensor, the noise sensor, and the power consumption sensor. The environmental sensors, namely noise and motion sensors, are characterized by a binary output. Differently, the output of the power consumption sensor is constituted by scalar values. According to [20], a high value of the power usage, combined with a high load of the variability, is often correlated with the presence of a person within the environment monitored. To this purpose, we consider two statistics, namely the mean value and the standard deviation of the power consumption along the time. Such statistics are used to infer the occupancy of a room. The design of the proposed algorithm is inspired by the stigmergy concept [21]. The stigmergy approach compares the value sampled by a sensor to the intensity of a pheromone. Similarly to the decay of an ant's pheromone released when it finds some food, the value of a sensor changes along the time and it decays as the time progresses. Stigma determines in real-time the room occupancy by observing the output of the motion, noise and power supply sensors. Detailed information about the used algorithm can be found in [22].

5. CONCLUSIONS

In this work, we presented the Smart Building Suite, where independent and different technologies are developed in order to realize a multimodal surveillance system. As a future work we plan to integrate all the technologies described in the paper and to enhance the current surveillance system with a high-level reasoner that crosses all the selected information coming from smart cameras, the sensors and the face recognition system, in order to better determine whether or not a non-authorized access to the building has occurred and, in case, to raise a series of alarms, according to the severity of the intrusion.

6. ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research, and the "Renewed Energy" project of the DIITET Department of CNR for supporting the Smart Area activity, in whose framework this research is carried on.

7. REFERENCES

[1] P. Barsocchi, E. Ferro, L. Fortunati, F. Mavilia, and F. Palumbo, "EMS@ CNR: An energy monitoring sensor network infrastructure for in-building location-based services," in *High Performance Computing & Simulation (HPCS)*. IEEE, 2014, pp. 857–862.

[2] T. He, S. Krishnamurthy, J. A. Stankovic, T. Abdelzaher, L. Luo, R. Stoleru, T. Yan, L. Gu, J. Hui, and B. Krogh, "Energy-efficient surveillance system using wireless sensor networks," in *Proceedings of the 2nd international conference on Mobile systems, applications, and services*. ACM, 2004, pp. 270–283.

[3] M. Magrini, D. Moroni, G. Palazzese, G. Pieri, G. Leone, and O. Salvetti, "Computer vision on embedded sensors for traffic flow monitoring," in *Intelligent Transportation Systems (ITSC)*. IEEE, 2015, pp. 161–166.

[4] C. Vairo, G. Amato, S. Chessa, and P. Valleri, "Modeling detection and tracking of complex events in wireless sensor networks," in *Systems Man and Cybernetics (SMC)*. IEEE, 2010, pp. 235–242.

[5] G. Amato, S. Chessa, C. Gennaro, and C. Vairo, "Efficient detection of composite events in wireless sensor networks: design and evaluation," in *Computers and Communications (ISCC), 2011 IEEE Symposium on*. IEEE, 2011, pp. 821–823.

[6] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[7] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Trans on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[8] S. Brutzer, B. Höferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *CVPR*. IEEE, 2011, pp. 1937–1944.

[9] G. R. Leone, M. Magrini, D. Moroni, G. Pieri, O. Salvetti, and M. Tampucci, "A smart device for monitoring railway tracks in remote areas," in *Computational Intelligence for Multimedia Understanding (IWCIM)*. IEEE, 2016, pp. 1–5.

[10] D. D. Bloisi, L. Iocchi, G. R. Leone, R. Pigliacampo, L. Tombolini, and L. Novelli, "A distributed vision system for boat traffic monitoring in the venice grand canal." in *VISAPP (2)*, 2007, pp. 549–556.

[11] G. Amato, F. Falchi, and C. Gennaro, "Fast image classification for monument recognition," *J. on Computing and Cultural Heritage*, vol. 8, no. 4, p. 18, 2015.

[12] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans on PAMI*, vol. 38, no. 1, pp. 142–158, 2016.

[15] G. Amato, F. Falchi, and L. Vadicamo, "Visual recognition of ancient inscriptions using convolutional neural network and Fisher vector," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 4, p. 21, 2016.

[16] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[17] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo, "Deep learning for decentralized parking lot occupancy detection," *Expert Systems with Applications*, vol. 72, pp. 327–334, 2017.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[19] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[20] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy, "Non-intrusive occupancy monitoring using smart meters," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*. ACM, 2013, pp. 1–8.

[21] P. Barsocchi, M. G. Cimino, E. Ferro, A. Lazzeri, F. Palumbo, and G. Vaglini, "Monitoring elderly behavior via indoor position-based stigmergy," *Pervasive Mob. Comput.*, vol. 23, no. C, pp. 26–42, Oct. 2015.

[22] P. Barsocchi, A. Crivello, M. Girolami, F. Mavilia, and E. Ferro, "Are you in or out? monitoring the human behavior through an occupancy strategy," in *Computers and Communication (ISCC)*. IEEE, 2016, pp. 159–162.