KNOWLEDGE-BASED SITUATIONAL ANALYSIS OF UNUSUAL EVENTS IN PUBLIC PLACES

David Münch, Stefan Becker, Hilke Kieritz,
Wolfgang Hübner and Michael Arens

david.muench@iosb.fraunhofer.de

Fraunhofer Institute for Optronics, System Technologies and Image Exploitation IOSB,
Gutleuthausstraße 1, 76275 Ettlingen (Germany)

Abstract

Combining appropriate methods from computer vision and artificial intelligence enables further progress in smart video surveillance. In this work, an Interacting Multiple Model (IMM) filter is used for person tracking due to the fact that a single motion may not capture the complex dynamics from persons. In addition, context information from the IMM is used for controlling the background model to detect left luggage. The combination of this processing chain serves as input for the situation recognition in addition to person detection and tracking.

The computer vision components are integrated in the distributed Cognitive Vision System (dCVS) architecture, which is applied up to now to Traffic, Robotics, Smart Homes, and Video Surveillance. For this work, we cope with situations dealing with unusual events in public places.

Keywords: left luggage detection, video surveillance, situation recognition.

1 MOTIVATION

The combination of suitable methods from both computer vision and artificial intelligence enables further progress in smart video surveillance. Supporting the surveillance operator with significant cues of unusual events, such as left luggage, is a huge step towards more situational awareness.

Methods establishing situational awareness are dependent on lower level computer vision components providing the necessary raw input data such as person detection, person tracking, change detection, or object detection, see e.g. Figure 1.

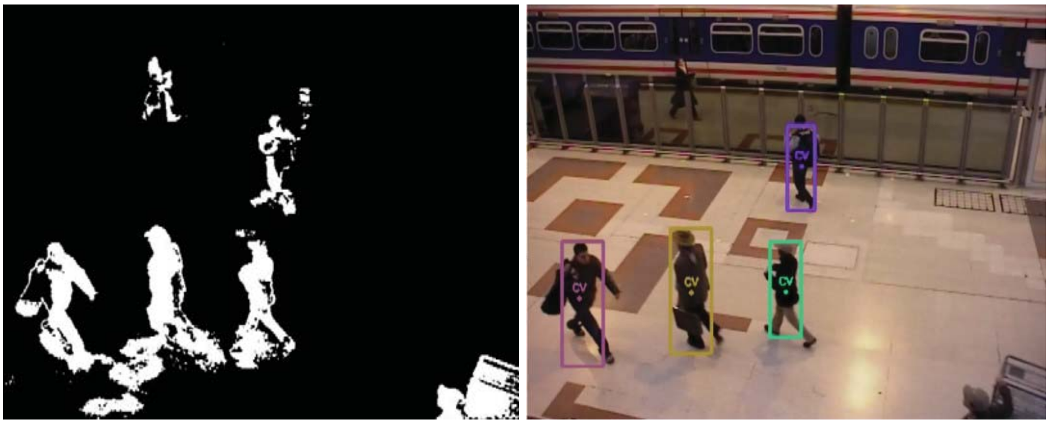

Fig. 1. Background subtraction (left) and person detection & tracking (right). All people are moving and are in a “constant velocity”-state.

In this paper, we integrate reliable computer vision components: person detection, person tracking, and left luggage detection with high-level semantic video understanding. The components are plugged into a distributed architecture to satisfy real practical needs in a working environment of video surveillance.

1.1 Related Work

The survey papers [1], [2], [3], [4], and [5] provide a broad overview about the recent field of situation recognition in video surveillance. The strategies to deal with the problem to extract meaningful semantic information from raw image sequences can be divided into different approaches. On the one hand, there are the direct approaches making use of massive machine learning support. They work in principal as a black box where the input image sequence is fed in, and without exactly knowing how these images are processed, the black box results the recognized situations. On one side, these methods work quite good in limited scenes, on the other side their success is mostly dependent on a huge amount of training data.

On the other hand, there are hierarchical approaches. They divide the problem into several layers in which different subproblems are addressed. Among them, there are statistical approaches [6] making use of graphical models and other probabilistic methods. It is a smart theory, in contrast to complex modelling and expensive inference. Syntactical approaches, such as formal grammars [7], are easier to model but less flexible for complex situations. Description-based approaches do not rely on training data; instead, they are based on expert knowledge and background knowledge. Their basis is mostly formal, such as higher order formal logic etc. Knowledge about spatial, temporal, and abstract properties is formulated and is available for inference during the process of recognizing situations [8], [12].

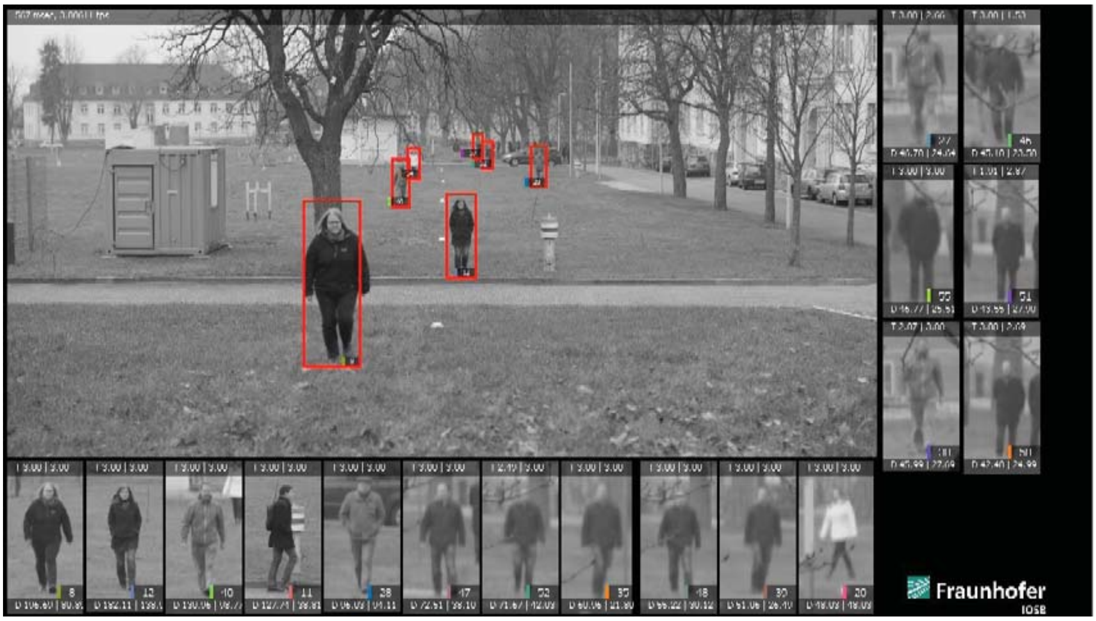

Fig. 2. Person detection & tracking: One challenge is to detect and track the corresponding height resolution of each person.

2 COMPUTER VISION COMPONENTS

In the following section, the computer vision components used in this work are explained.

2.1 Person detection

In any person centric scenario, a person detection method is a core component. As every other subsequent step is based on the person data, it has to provide robust, reliable, and coherent results. The person detection is based on [9], [10], and [11].

Person detection consists of three steps:

1. Pre-processing for generation and transformation of meaningful features.

   In our case, ten feature maps are generated, including colour maps, gradient, and gradient orientations.

2. Classification with a soft-cascade and sliding window multi-scale approach.

   For this work, the classifier instead of the image is scaled. This has two advantages: First, no recalculation of features, second, learning the classifier on different scales lets the classifier learn to deal with unpredictable artefacts at different scales.

3. Post-processing to identify the detected person with a non-maxima suppression.

Training of the person detection method is done with two disjoint data sets to avoid the problem of overfitting. One data set (training) is used for setting the parameters of the weak classifier and the other data set (validation) is used for defining the soft-cascade thresholds.

Figure 1 (right) and Figure 2 depict the result of person detection (and tracking). Challenges in the person detection step are different view angles of the person (Figure 1 (right)) and a large variance of the distance of persons to the camera and consequently their resolution in image space (Figure 2).

Directly from person detection (and tracking) simple events can be inferred. There are e.g. person pass-through, person counting, and intrusion detection.

2.2 Person Tracking & Left Luggage Detection

In recent years, many methods have been proposed for automatically detect abandoned objects in video surveillance. These methods not only differ in their desired application, but in addition, how such an event is defined. In this work, we use the approach of Becker et al. [13] for automatically detecting left luggage in surveillance scenarios. Similar to [14] a left luggage sets a split from a person as a prerequisite. One processing stage is the detection and tracking of persons in order to detect a drop-off event from a non-human, static object. One way for detecting a change of static

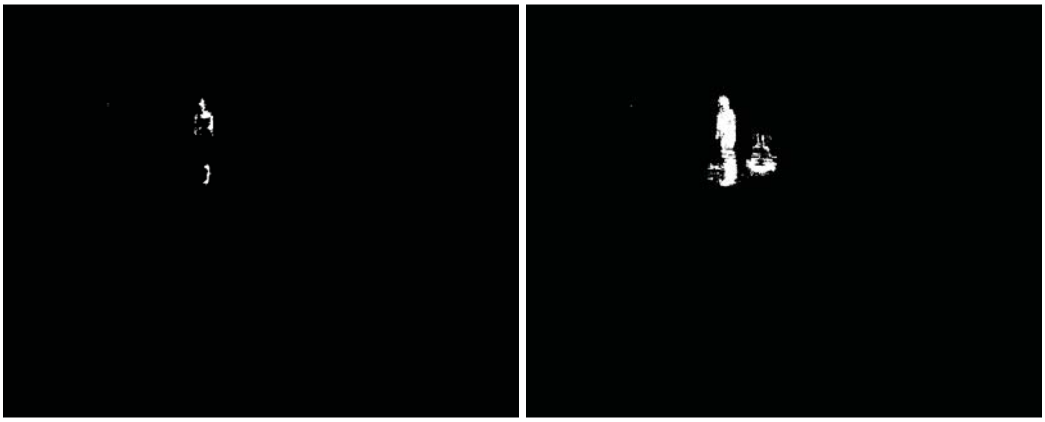

Fig. 3. Foreground masks obtained from the background subtraction algorithm with (right) and without context information (left).

object in the scene caused by such an event is to compare long-term and short-term background models with different learning rates. Instead of using such dual background models, the used approach relies on the state estimate of the dynamics of a tracked person for controlling the pixel-wise update probabilities of one non-parametric background model. The approach of Hofmann et al. [15] serves as basis for the background model. There, the background is modelled by a history of recently observed pixel values. Not only a dynamic controller controls the foreground decision, also the background update is based on a learned per-pixel state variable. As mentioned, here the background update is in addition combined with the person state estimate. This combination prevents that objects carried by person are too fast integrated in the background model. Hence, the per-pixel update probability depends on a measure of the pixel dynamics and on the person dynamics. In a region close to a person, the update probability is decreased, when the person is in a standing or loitering state. In contrast, the update probability is not changed in cases where the person is for example in a fast moving state.

Figure 3 shows an example image of a resulting foreground masks in case of a slowly moving or rather standing person with and without context information. On the left the original image is shown, in the middle the resulting foreground mask obtained with no tracking information is shown, and on the right with context information. When the tracking state is not used, the person and the carried object are almost fully integrated in the background model. Another major part of this processing pipeline consists of determining the context information for the used context aware pixel-based adaptive segmenter. In order to differ between persons that stand still, walk, or run is important to use several motion model to describe such varying characteristic and to consider that the motion model of a person can change over time. For modelling the different motion states of person and simultaneously better capture the complex dynamics of a person, an Interacting Multiple Model (IMM) filter [16] is used. IMM filter are a popular choice for estimating systems, whose model changes according to a finite-state, discrete-time Markov chain. IMM filter can also be used in situations, in which its parameters are estimated from a set of candidate models, and hence it can be also used as a method for model comparison [17]. In our experiments, we used a set of three different motion model. A constant position, a constant velocity, and a constant acceleration model. The IMM filter consists of three major steps: interaction (mixing), filtering, and combination. Under the assumption that a particular model is the right model at current time step, the initial conditions for this model is obtained by mixing the state estimates produced by all filters. Then a standard Kalman filtering is applied for each model. Followed by computing the weighted combination of all updated state estimates. This yields to the final state estimate and covariance in that particular time

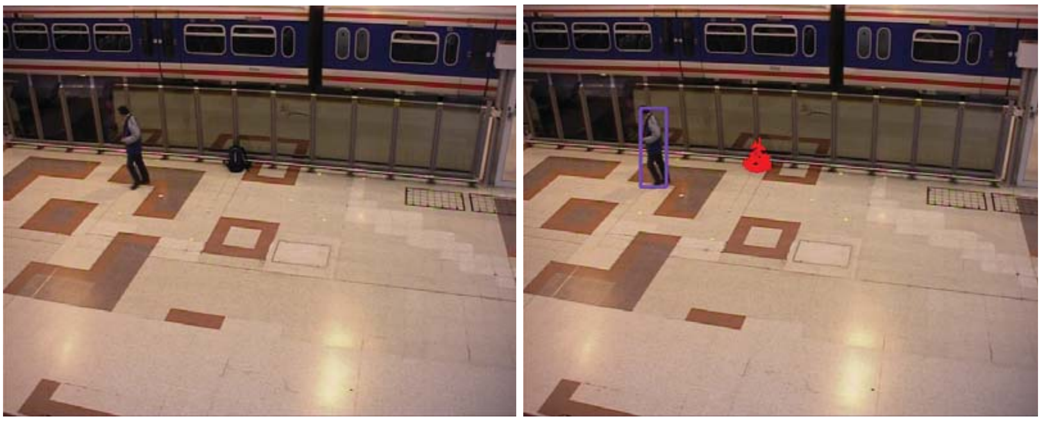

Fig. 4. (Left) Sample input image. (Right) Example of detected abandoned object. (PETS2006 dataset; S1-T1).

step. The weights are chosen according to the probabilities of the models, which are computed in filtering step of the algorithm. For more details on the IMM Filter, we refer to the work of Bar-Shalom et al. [16] or Blom et al. [18]. The states of our IMM filter are the image space coordinates and the height of the person bounding box. In case that the constant position model is the best fitting model for a current time step. The current states is used to define a region of interest and conduce as context information for a background model. Further, the tracking id is assigned to this region. Hence, a unique assignment between possible left object and the responsible person is achieved. As mentioned above, a too fast integration of a standing person and his carried object in to the background model is avoided. When a person leaves its assigned region of interest, which is associated with a switching from a standing state to a walking or running state, the detected foreground pixels are used to trigger an alarm event. In our experiments, an alarm event is triggered when a person leaves its assigned region of interest for a defined time interval and the number of detected foreground pixels inside this region exceeds a threshold. Figure 4 shows exemplary the result of detected left luggage event in PETS2006 dataset [19]. The detected left luggage is highlighted with red. The tracked persons are marked with a bounding box.

3 ARTIFICIAL INTELLIGENCE MODULES

The computer vision components above are integrated in the distributed Cognitive Vision System (dCVS) architecture [12], which is applied up to now to Traffic, Robotics, Smart Homes, and Video Surveillance. The dCVS consists of different levels; in the Contextual Level (CL) reasoning, situational analysis, visualization, and evaluation is done. Expert knowledge encoded as Situation Graph Trees and Fuzzy Metric Temporal Logic is used as knowledge base. In between of the CL and the Quantitative Level (QL) a persistence layer cares about communication between the different computer vision modules from the QL and the CL. At the bottom in the Interactive Subsystem raw sensor data, such as video data from cameras, is processed and passed to the QL.

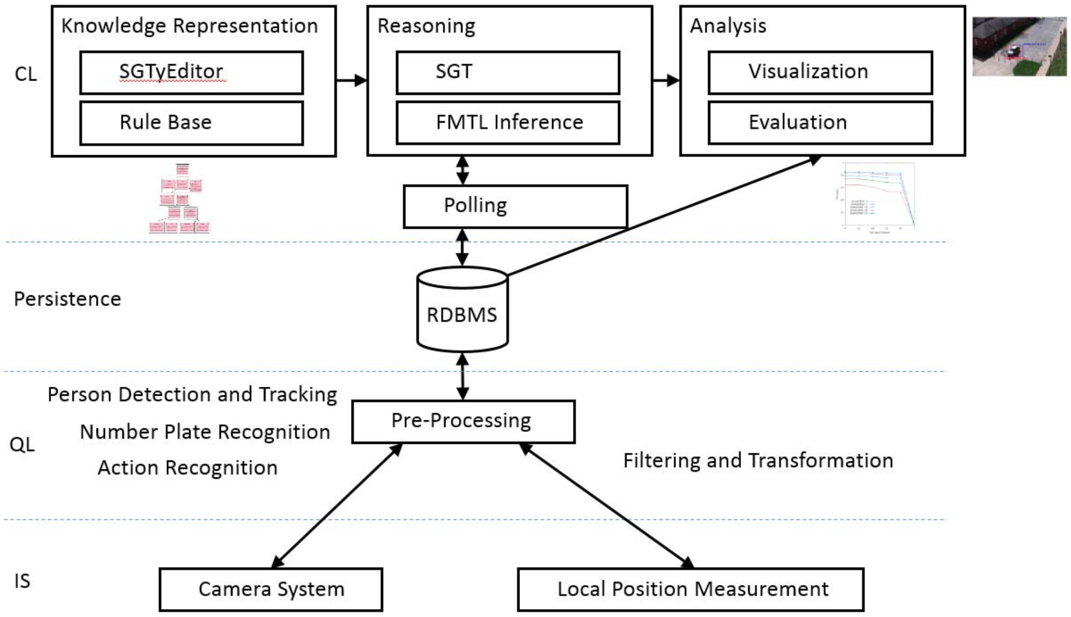

Fig. 2. Concrete instantiation of the dCVS.

Some simpler situations can be detected with the computer vision modules. These situations are usually single person/object related and based on the position and its derivatives. Such situations are walking, running, pass-through, counting, intrusion detection, left luggage detection. Whereas using high-level semantic support, such as the dCVS, raises the capability to recognize much more complex situations. In the following, we will discuss the components of the dCVS and its consequences.

The dCVS is a modular architecture for (computer-vision-based) situation recognition. This property makes it universally applicable in any kind of surveillance task. E.g., any of the computer vision modules in the QL can be exchanged, added, or changed, such that the computer vision modules fit to the application at hand. In a practical environment (with an existing infrastructure), there is need for a consistent information basis for the situation recognition in CL. This is achieved with a persistence layer, where all results from QL are stored and subsequently passed to CL.

The logic based inference system in CL can use background knowledge in different kind of formats. On the one hand, there is basic knowledge "the physics of this world" which is provided as logical formulas. On the other hand, there is higher-level knowledge, which is provided via situation graph trees. In this structure, knowledge of the expected situations of the agents in an observed scene can be exhaustively modelled.

The combination of both allows inferring about the completely observed scene with all its persons and objects. At this point interactions and group situations can be detected because every single-person property can be put in context.

4 CONCLUSION

This work focuses on combining suitable computer vision components for high-level situation recognition. The contribution of this paper is the concrete set of computer vision modules into the dCVS architecture to deal with indoor surveillance scenarios. The computer vision modules in conjunction with the dCVS allow an exhaustive situational analysis for better supporting surveillance operators.

REFERENCES

[1] Turaga, P., R. Chellappa, V. S. Subrahmanian, and O. Udrea (2008). *Machine Recognition of Human Activities: A Survey*. IEEE Transactions on Circuits and Systems for Video Technology.

[2] Lavee, G., E. Rivlin, and M. Rudzsky (2009). *Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video*. IEEE Transactions on System, Man, and Cybernetics, Part C: Applications and Reviews.

[3] Aggarwal, J. K. and M. S. Ryoo (2011). *Human Activity Analysis: A Review.* ACM Computing Surveys.

[4] Vishwakarma, S. and A. Agrawal (2012). *A survey on activity recognition and behavior understanding in video surveillance*. The Visual Computer.

[5] Ye, Juan, Simon Dobson, and Susan McKeever (2012). *Situation identification techniques in pervasive computing: A review*. Pervasive Mob. Comput.

[6] Fischer Y. and J. Beyerer (2012). *Defining dynamic Bayesian networks for probabilistic situation assessment.* Proceedings of the 15th International Conference on Information Fusion (FUSION).

[7] Aloimonos, Y., G. Guerra, and A. Ogale (2009). *The language of action: a new tool for human-centric interfaces*. In Human-Centric Interfaces for Ambient Intelligence.

[8] Ryoo, M. and J. Aggarwal (2009). *Semantic Representation and Recognition of Continued and Recursive Human Activities*. International Journal of Computer Vision.

[9] Dollár, Piotr, Serge Belongie, and Pietro Perona (2010). *The Fastest Pedestrian Detector in the West.* BMVC.

[10] Benenson, Rodrigo, et al (2013). *Seeking the strongest rigid detector*. Computer Vision and Pattern Recognition (CVPR).

[11] Kieritz, H., W. Hübner, and M. Arens (2013). *Learning transmodal person detectors from single spectral training sets.* SPIE 8901A Optics and Photonics for Counterterrorism, Crime Fighting and Defence.

[12] Münch D., A.-K. Grosselfinger, H. Kieritz, W. Hübner, M. Arens (2014). *Architecture for and Evaluation of Situational Analysis in the Real World*. Proc. of the 9th Future Security Research Conference, Berlin, Germany.

[13] Becker, S., D. Münch, H. Kieritz, W. Hübner, and M. Arens (2015). *Detection of abandoned objects based on interacting multiple models* Proc. SPIE Volume 9652 Optics and Photonics for Counterterrorism, Crime Fighting, and Defence.

[14] Ferrando, S., G. Gera, and C. Regazzoni (2006). *Classification of unattended and stolen objects in videosurveillance system.* International Conference on Advanced Video and Signal based Surveillance (AVSS).

[15] Hofmann, M., P. Tiefenbacher, and G. Rigoll (2012). *Background segmentation with feedback: The pixel-based adaptive segmenter.* Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[16] Bar-Shalom, Y., T. Kirubarajan, and X.-R. Li (2002). *Estimation with Applications to Tracking and Navigation.* John Wiley & Sons, Inc., New York, NY, USA.

[17] Li, X. and V. Jilkov (2005). *Survey of maneuvering target tracking. part v. multiple-model methods*. IEEE Transactions on Aerospace and Electronic Systems.

[18] Blom, H. and Y. Bar-Shalom (1988). *The interacting multiple model algorithm for systems with markovian switching coefficients.* Transactions on Automatic Control.

[19] *PETS, International workshop on Performance Evaluation of Tracking and Surveillance*. (2006). (see http://www.cvg.rdg.ac.uk/PETS2006/index.html).