# Tracking The Invisible Man
## *Hidden-object Detection for Complex Visual Scene Understanding*

Joanna Isabelle Olszewska
*School of Computing and Technology, University of Gloucestershire, The Park, Cheltenham, GL50 2RH, U.K.*

Keywords: Surveillance Application, Visual Scene Analysis, Automated Scene Understanding, Knowledge Representation, Spatio-temporal Visual Ontology, Symbolic Reasoning, Computer Vision, Pattern Recognition.

Abstract: Reliable detection of objects of interest in complex visual scenes is of prime importance for video-surveillance applications. While most vision approaches deal with tracking visible or partially visible objects in single or multiple video streams, we propose a new approach to automatically detect all objects of interest being part of an analyzed scene, even those entirely hidden in a camera view whereas being present in the scene. For that, we have developed an innovative artificial-intelligence framework embedding a computer vision process fully integrating symbolic knowledge-based reasoning. Our system has been evaluated on standard datasets consisting of video streams with real-world objects evolving in cluttered, outdoor environment under difficult lighting conditions. Our proposed approach shows excellent performance both in detection accuracy and robustness, and outperforms state-of-the-art methods.

## 1 INTRODUCTION

The growth of video-surveillance in daily life applications (Albanese et al., 2011) has opened the door to the development of automatic systems for multiple-object tracking (Bhat and Olszewska, 2014), suspicious object detection (Ferryman et al., 2013), or unusual activity recognition (Chen et al., 2014), in a single or multiple views of a recorded scene (Dai and Payandeh, 2013).

In particular, the efficient detection and tracking of objects of interest in a multi-camera environment (Fig. 1) is still a challenging task. Indeed, it implies the understanding of the camera network in terms of visual coverage of the cameras (Mavrinac and Chen, 2013), calibration of the cameras (Remagnino et al., 2004), etc. It also requires the design of computer vision techniques being robust to varying lighting conditions, or to objects occlusions of different nature such as object-to-object occlusions and object-to-scene occlusions (Yilmaz et al., 2006). Moreover, it usually involves the modelling of the knowledge about the scene, e.g. the number of the persons evolving in the scene, their location within the scene, or the direction of their trajectory.

In the computer-vision literature, works performing multi-object tracking in multi-camera environment apply techniques such as synergy map (Evans et al., 2013), probabilistic occupancy maps (Fleuret et al., 2008), or K-shortest path (Berclaz et al., 2011) to model such knowledge. Despite being widely used, these statistical methods are limited in terms of scalability with respect to the number of considered, contextual data (Riboni and Bettini, 2011).

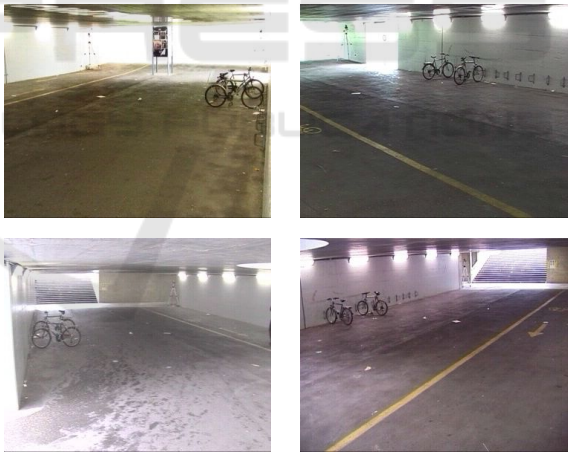

Figure 1: Samples of real-world, outdoor views acquired by static, synchronized cameras in context of multi-camera video surveillance of a passageway scene.

On the other hand, symbolic representation has been used to codify knowledge about visual scenes in context of video content analysis (Bai et al., 2007), video summarization (Park and Cho, 2008),



Olszewska, J.
Tracking The Invisible Man - Hidden-object Detection for Complex Visual Scene Understanding.
DOI: 10.5220/0005852302230229
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 223-229
ISBN: 978-989-758-172-4

or video annotation (Natarajan and Nevatia, 2005), (Jeong et al., 2011), especially modelling the video-surveillance domain (Vrusias et al., 2007). In these works, ontologies have been developed to describe the studied visual scenes, but not to deduce new information.

Recently, some papers propose to integrate structured symbolic knowledge into computer-vision systems for event recognition (Sridhar et al., 2010), event prediction (Lehmann et al., 2014) or tracking estimation (Gomez-Romero et al., 2011). These approaches have been proven to be efficient. However, these context-aware methods are mainly deductive rather than inductive, and are designed for single camera views only.

In this paper, we propose to incorporate symbolic description of the scene together with ontological reasoning into a vision system to infer knowledge about the scene in order to detect and track objects of interest which may be hidden in some/most of the views. Hence, the designed system features a multi-camera, knowledge-based, detector and tracker of both visible and non-visible objects of the scene, and generates a complete, semantic description of the scene as well as its visual annotation in all views.

The analysed scene is assumed to be acquired in outdoor or indoor environment, captured by multiple, synchronized cameras with overlapping field of views (FOVs). Our system supports both static or mobile cameras, and does not require the specific knowledge of the parameters of the cameras.

Objects evolving in the scene could present occlusions in one or several views. Occlusions could be of object-to-object type, when two or more objects of interest overlap each other in a ratio from 0.1 to 1 (or full occlusion); or of object-to-scene type, when an object of interest is partially or totally not visible due to objects present in the background.

The developed intelligent vision system allows a computationally efficient and accurate analysis of objects of interest evolving in one or multiple views of a scene. It provides both qualitative and quantitative answers to the following questions: How much objects are in the scene? Where are the objects in each view? Is there a hidden object in a view? Which object is hidden in that view? Where about it is hidden in that view?

Hence, the contributions of this paper are twofold:

- the design of an automated vision system to detect both visible and invisible objects of interest evolving in real-world scenes captured by multiple, synchronized cameras;
- the use of symbolic knowledge representation and qualitative spatial relations for information induction rather than deduction, in context of automated detection and tracking of objects of interest in multi-view scenes.

The paper is structured as follows. In Section 2, we describe our system for multi-camera stream analysis (see Fig. 2) based on both computer vision techniques to compute quantitative data, and on artificial intelligence methods to process qualitative knowledge in order to induce information. Our approach performance have been assessed on standard, real-world video-surveillance dataset as reported and discussed in Section 3. Conclusions are drawn up in Section 4.

## 2 PROPOSED APPROACH

The proposed approach consists of seven steps as summarized in Fig. 2.

At first, frames of the different views of the scene are extracted from the videos acquired by cameras which could have the same or different calibration parameters (see Fig. 4, 1st row).

Secondly, these visual views are processed in order to be synchronized both in time and space. The temporal synchronization consists in matching the time stamp of each of the video frame with this of a frame related to another view. Spatial matching (Ferrari et al., 2006) of temporally synchronized views is performed by matching local descriptors extracted in both frames of each of the background view. If there is more than two views, the matching process is repeated for each of the pair of views. It is worth to note that this second step could be done offline or partially skipped in case of synchronized videos or previously aligned views.

Thirdly, the visible objects of interest are detected in each of the visual views as described in (Olszewska, 2015) by means of active contours (see Fig. 4, 2nd row). Active contours are initialized based on blobs obtained by combining both frame difference and background subtraction techniques. Considering a color image $I(x,y)$ with M and N, its width and height, respectively, and RGB, its color space, blobs are computed in parallel by, on one hand, the difference between a current frame $I_k^v(x,y)$ in the view $v$ and the precedent one $I_{k-1}^v(x,y)$, and by, on the other hand, the difference between the current frame $I_k^v(x,y)$ and a background model of the view $v$, and afterwards, by adding both results in order to extract the foreground in the corresponding view. The background itself is modeled using the running Gaussian average (RGA), characterized by the mean $\mu_b^v$ and the variance $(\sigma_b^v)^2$, as the RGA method suits well for real-time tracking.

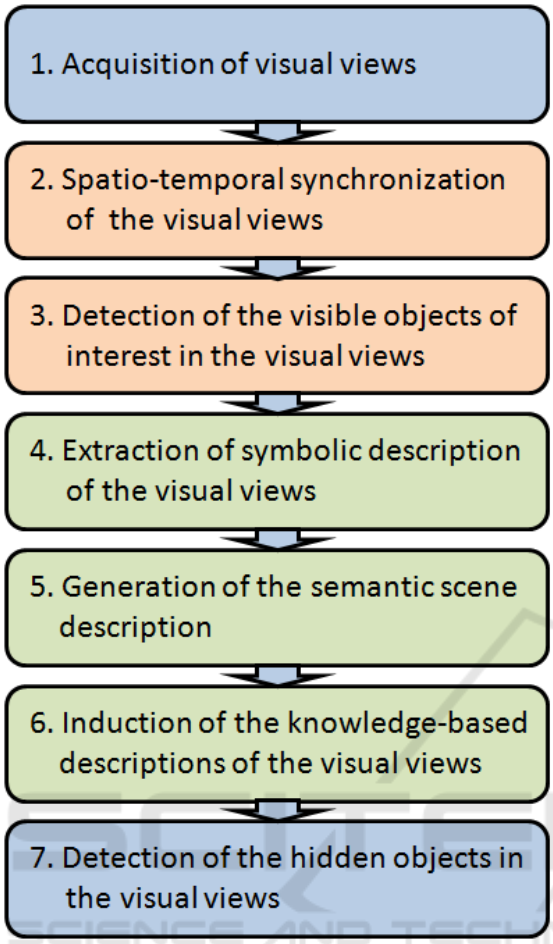


Figure 2: Overview of our symbolic-based approach for hidden-object detection in a multi-camera environment.

Hence, the foreground is determined by (Olszewska, 2015)

$$F^v(x,y) = \begin{cases} 1 & \text{if } \left|F_f^v(x,y) \cup F_b^v(x,y)\right| = 1, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

with

$$F_f^v(x,y) = \begin{cases} 1 & \text{if } \left|I_k^v(x,y) - I_{k-1}^v(x,y)\right| > tf, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

and

$$F_b^v(x,y) = \begin{cases} 1 & \text{if } \left|I_k^v(x,y) - \mu_b^v\right| > n \cdot \sigma_b^v, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where $tf$, is the threshold, and $n \in \mathbb{N}_0$.

To compute a final blob defined by labeled connected regions, morphological operations such as opening and closure are applied to the extracted foreground $F^v$, in order to exploit the existing information

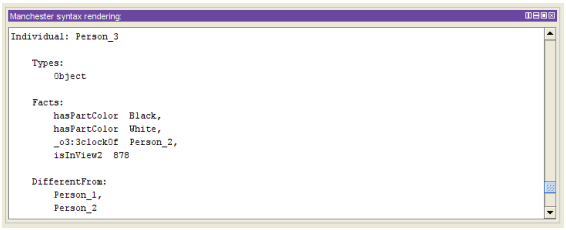


Figure 3: Sample of auto-generated scene description in OWL.

on the neighboring pixels, in a view $v$,

$$f^v(x,y) = Morph(F^v(x,y)). \quad (4)$$

Then, an active contour is computed for each frame $k$ in each view $v$ separately, and for each targeted object. In this work, an active contour (Olszewska, 2012) is a parametric curve $\mathcal{C}(s) : [0,1] \rightarrow \mathbb{R}^2$, which evolves from its initial position computed by means of Eq. (4) to its final position, guided by internal and external forces as follows:

$$\mathcal{C}_t(s,t) = \alpha\,\mathcal{C}_{ss}(s,t) - \beta\,\mathcal{C}_{ssss}(s,t) + \Xi, \quad (5)$$

where $\mathcal{C}_{ss}$ and $\mathcal{C}_{ssss}$ are respectively the second and the fourth derivative with respect to the curve parameter $s$; $\alpha$ is the elasticity; $\beta$ is the rigidity; and $\Xi$ is the multi-feature vector flow (Olszewska, 2013) .

After the detection of the visible objects of interest in the different views is performed using active contours as explained above, the symbolic description of the visual views is extracted automatically using the framework set in (Olszewska and McCluskey, 2011) and repeated for each of the view.

The generation of the semantic scene description, as illustrated in Fig. 3 for a snippet of the scene presented in Fig. 4., followed by the generation of the views' knowledge-based descriptions (Fig. 4, 3rd row), is automatically induced by the reasoner. As each view is related to another one because of the overlapping fields of view of the same scene and because of views' synchronisation, logic rules have been defined in DL such as

$$\begin{aligned} hasInScene \sqsubseteq\ & Scene\_Property \\ & \sqcup hasInView1 \\ & \sqcup \ldots \\ & \sqcup hasInViewN, \end{aligned} \quad (6)$$

where $N \in \mathbb{N}$ is the number of views, and *hasInViewn* is set as a sub-property of *hasInScene*, with $n = 1,\ldots,N$.

Then, the detection of the hidden objects in the visual views is based on this induced knowledge. Moreover, qualitative spatial relations such as RCC-8 and the o'clock model applied to objects of each of the views allow the definition of the potential regions

where could appear hidden objects. Finally, the non-visible objects detected in the last step as well as the visible objects detected in the third step are all localized in the views by means of bounding boxes (see Fig. 4, 4th row). The latter ones are computed to surround the detected objects in order to use standard metrics for sake of comparison with other methods, when tracking over the time the target objects, as detailed in Section 3.

# 3 EXPERIMENTS AND DISCUSSION

To validate our approach, we have applied our system on the publicly available CVLAB dataset (Berclaz et al., 2011) called *Passageway*. It contains four video-surveillance sequences, recorded each by one of the four corresponding DV cameras at a rate of 25 fps, and encoded with Indeo 5. All cameras were synchronized and located about 2 meters from the ground. They were filming the same area under different angles, and their fields of view were overlapping (Fig. 1). The resulting four videos are made of 2500 frames each, with a frame resolution of 360x288 pixels. The chosen location for the data acquisition was an outdoor environment consisting of a dark underground passageway to a train station, where were evolving objects of interests, i.e. pedestrians.

This database of 10,000 images in total owns challenges such as handling variations of the persons in quantity, pose, motion, size, appearance, and scale. In particular, the area covered by the system is wide, and people get very small on the far end of the scene, making their precise localization challenging.

This series of multi-camera video sequences involving several people passing through a public underground passageway also presents large lighting variations, which is typical in real-world surveillance situations. Indeed, scene's lighting conditions are very poor, since a large portion of the images is either underexposed or saturated.

Most importantly, this dataset requires the processing of multi-view video streams where many parts of the scenario were filmed by only two or even a single camera, with some people partially occluded or not visible over significant numbers of frames in the related views. All these difficulties make the dataset challenging and interesting to test our approach.

All the experiments have been run on a computer with Intel Core 2 Duo Pentium T9300, 2.5 GHz, 2Gb RAM, using MatLab and OWL languages as well as HermiT reasoner.

To evaluate the performance of our system, we adopt the standard CLEAR metrics, i.e. Multiple Object Detection Accuracy (MODA) and Multiple Object Detection Precision (MODP), as well as Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP). These metrics have become standard for the evaluation of detection and tracking algorithms, and are convenient to compare our approach with other works such as (Berclaz et al., 2011). The detection precision metric (MODP) assesses the quality of the bounding box alignment in case of correct detection, while its accuracy counterpart (MODA) evaluates the relative number of false positives and missed detections (Kasturi et al., 2009). The tracking precision metric (MOTP) measures the alignment of tracks compared against ground truth, while the tracking accuracy metric (MOTA) produces a score based on the amount of false positives, missed detections, and identity switches (Bernardin and Stiefelhagen, 2008).

Table 1: Multiple Object Detection Accuracy (MODA) and Multiple Object Detection Precision (MODP) in CVLAB *Passageway* video frames, using approaches such as (Fleuret et al., 2008), (Berclaz et al., 2011), and our.

| | ◇ | □ | **our** |
|---|---|---|---|
| MODA | 63% | 72% | 96% |
| MODP | 66% | 70% | 94% |

Table 2: Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) in CVLAB *Passageway* video streams, using approaches such as (Berclaz et al., 2011) and our.

| | □ | **our** |
|---|---|---|
| MOTA | 73% | 95% |
| MOTP | 68% | 94% |

Our approach has been tested for detection and tracking of both visible and hidden objects of interest on the four multi-camera video streams of the CVLAB dataset.

Samples of our results are presented in Fig. 4. This scene presents difficult situations such as strong patterns, e.g. the yellow line on the floor in views 2 and 3 or the staircases in views 3 and 4, poor foreground/background contrast, light reflections, or illumination changes. Moreover, some target objects could only be seen in one of the views as per configuration illustrated in Fig. 1. Hence, in Fig. 4, the three persons present in the scene are only visible in one of the four views (view 2). Views 1 and 4 only show two of the three persons, whereas only one of them appears in the view 3. Our system copes well with these situations as discussed below.

In Table 1, we have reported the Multiple Object Detection Accuracy (MODA) and Multiple Ob-


Figure 4: Examples of results obtained with our approach for a scene captured by four synchronized cameras with overlapping field of views. First column: view 1; Second column: view 2; Third column: view 3; Fourth column: view 4. First row: step 1 - raw data images. Second row: step 3 - visible persons detected in each of the camera views (e.g. Person 1 by yellow active contour; Person 2 by blue active contour; Person 3 by red active contour). Third row: step 6 - snippet of the automatically generated description for each view (extracted knowledge in blue, inducted knowledge in yellow). Fourth row: step 7 - detected persons in all views (bounding boxes), including the hidden persons (dotted bounding boxes). Best viewed in color.

ject Detection Precision (MODP) rates of our method against the rates achieved by (Fleuret et al., 2008) and (Berclaz et al., 2011), while in Table 2, we have displayed the Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) scores of our method against the rate obtained by (Berclaz et al., 2011).

From Tables 1-2, we can observe that our system provides reliable detection of objects of interest in multi-camera environment, and that our multiple-object tracking method is also very accurate, outperforming state-of-the-art techniques. Indeed, methods relying, e.g. on detection maps which can get very noisy due to the difficult real-world, outdoor environment conditions, have thus their performance greatly affected (Fleuret et al., 2008), (Berclaz et al., 2011), unlike our approach.

Furthermore, state-of-the-art methods only deal with partial occlusions of the objects, whereas our system allows the detection of hidden objects, i.e. objects of interest fully occluded by either other foreground objects or by background objects. Our system performs the invisible object detection and tracking by means of the conjunction of effective vision techniques with knowledge induction and integration of qualitative spatial relations. It is worth noting that strong occlusions are an additional difficulty for tracking systems to keep the tracks of the objects of interest.

For all the dataset, the average computational speed of our approach is in the range of milliseconds, thus our developed system could be used in context of real-world, video surveillance.

## 4 CONCLUSIONS

Detecting and tracking both visible and invisible objects in multi-camera environment is a challenging task in video surveillance. For this purpose, we have developed a system incorporating symbolic knowledge, including spatial relations, into a computer-vision framework. Our approach outperforms the ones found in the literature for both object detection

and tracking as demonstrated on outdoor real-world scenes, while the proposed conceptual reasoning contribute to the visual processing, allowing the location of hidden objects through knowledge induction.

## REFERENCES


Albanese, M., Molinaro, C., Persia, F., Picariello, A., and Subrahmanian, V. S. (2011). Finding unexplained activities in video. In *Proceedings of the AAAI International Joint Conference on Artificial Intelligence*, pages 1628–1634.

Bai, L., Lao, S., Jones, G. J. F., and Smeaton, A. F. (2007). Video semantic content analysis based on ontology. In *Proceedings of the IEEE International Machine Vision and Image Processing Conference*, pages 117–124.

Berclaz, J., Fleuret, F., Tueretken, E., and Fua, P. (2011). Multiple object tracking using K-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1806–1819.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10.

Bhat, M. and Olszewska, J. I. (2014). DALES: Automated Tool for Detection, Annotation, Labelling and Segmentation of Multiple Objects in Multi-Camera Video Streams. In *Proceedings of the ACL International Conference on Computational Linguistics Workshop*, pages 87–94.

Chen, L., Wei, H., and Ferryman, J. (2014). ReadingAct RGB-D action dataset and human action recognition from local features. *Pattern Recognition Letters*, 50:159–169.

Dai, X. and Payandeh, S. (2013). Geometry-based object association and consistent labeling in multi-camera surveillance. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(2):175–184.

Evans, M., Osborne, C. J., and Ferryman, J. (2013). Multicamera object detection and tracking with object size estimation. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 177–182.

Ferrari, V., Tuytelaars, T., and Gool, L. V. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2):159–188.

Ferryman, J., Hogg, D., Sochman, J., Behera, A., Rodriguez-Serrano, J. A., Worgan, S., Li, L., Leung, V., Evans, M., Cornic, P., Herbin, S., Schlenger, S., and Dose, M. (2013). Robust abandoned object detection integrating wide area visual surveillance and social context. *Pattern Recognition Letters*, 34(7):789–798.

Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282.

Gomez-Romero, J., Patricio, M. A., Garcia, J., and Molina, J. M. (2011). Ontology-based context representation and reasoning for object tracking and scene interpretation in video. *Expert Systems with Applications*, 38(6):7494–7510.

Jeong, J.-W., Hong, H.-K., and Lee, D.-H. (2011). Ontology-based automatic video annotation technique in smart TV environment. *IEEE Transactions on Consumer Electronics*, 57(4):1830–1836.

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Boonstra, M., Korzhova, V., and Zhang, J. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336.

Lehmann, J., Neumann, B., Bohlken, W., and Hotz, L. (2014). A robot waiter that predicts events by high-level scene interpretation. In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages I.469–I.476.

Mavrinac, A. and Chen, X. (2013). Modeling coverage in camera networks: A survey. *International Journal of Computer Vision*, 101(1):205–226.

Natarajan, P. and Nevatia, R. (2005). EDF: A framework for semantic annotation of video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, page 1876.

Olszewska, J. I. (2012). Multi-target parametric active contours to support ontological domain representation. In *Proceedings of the RFIA Conference*, pages 779–784.

Olszewska, J. I. (2013). Multi-scale, multi-feature vector flow active contours for automatic multiple-face detection. In *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*.

Olszewska, J. I. (2015). Multi-camera video object recognition using active contours. In *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pages 379–384.

Olszewska, J. I. and McCluskey, T. L. (2011). Ontology-coupled active contours for dynamic video scene understanding. In *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, pages 369–374.

Park, H.-S. and Cho, S.-B. (2008). A fuzzy rule-based system with ontology for summarization of multi-camera event sequences. In *Proceedings of the International Conference on Artificial Intelligence and Soft Computing. LNCS 5097.*, pages 850–860.

Remagnino, P., Shihab, A. I., and Jones, G. A. (2004). Distributed intelligence for multi-camera visual surveillance. *Pattern Recognition*, 37(4):675–689.

Riboni, D. and Bettini, C. (2011). COSAR: Hybrid reasoning for context-aware activity recognition. *Personal and Ubiquitous Computing*, 15(3):271–289.

Sridhar, M., Cohn, A. G., and Hogg, D. C. (2010). Unsupervised learning of event classes from video. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, pages 1631–1638.

Vrusias, B., Makris, D., Renno, J.-P., Newbold, N., Ahmad, K., and Jones, G. (2007). A framework for ontology enriched semantic annotation of CCTV video. In *Proceedings of the IEEE International Workshop on Image Analysis for Multimedia Interactive Services*, page 5.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object Tracking: A Survey. *ACM Computing Surveys*, 38(4):13.