Multi-camera Video Object Recognition Using Active Contours

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

Keywords: Active Contours, Multi-camera Detection, Unsupervised Segmentation, Video-object Recognition, Semantic Colors, Multi-feature Vector Flow, Information Fusion, Video Surveillance, Scene Understanding.

Abstract: In this paper, we propose to tackle with multiple video-object detection and recognition in a multi-camera environment using active contours. Indeed, with the growth of multi-camera systems, many computer vision frameworks have been developed, but none taking advantage of the well-established active contour method. Hence, active contours allow precise and automatic delineation of entire object's boundaries in frames, leading to an accurate segmentation and tracking of video objects displayed into the multi-view system, while our late fusion approach allows robust recognition of the detected objects in the synchronized sequences. Our active-contour-based system has been successfully tested on video-surveillance standard datasets and shows excellent performance in terms of computational efficiency and robustness compared to state-of-art ones.

1 INTRODUCTION

The growing use of multi-camera networks for video surveillance (Kumar et al., 2010), (Bhat and Olszewska, 2014) and its related applications such as robotics (M. Kamezaki, 2014), intelligent transport (Spehr et al., 2011), monitoring (Remagnino et al., 2004), event detection (Zhou and Kimber, 2006), or tracking (Fleuret et al., 2008) stimulates the development of computer-vision approaches which aim to efficiently analyse the resulting big amount of visual data to extract meaningful information.

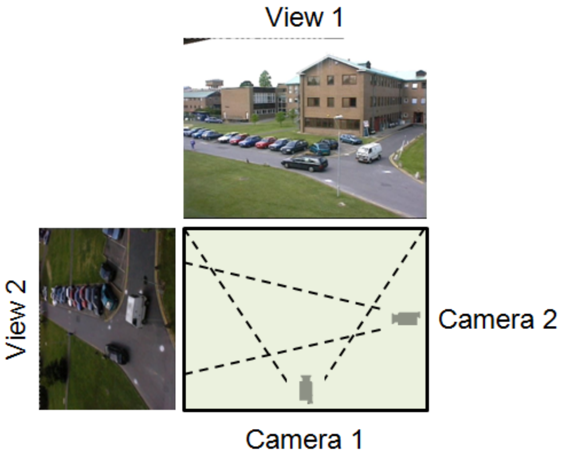

Figure 1: Outlook of the field of view (FOV) of each camera, as in the used scenario to test our approach.

In particular, the design of multi-view video recognition systems is of prime importance. Such systems need to process multi-view video streams, i.e. video sequences of a dynamic scene captured simultaneously by multiple cameras, for detecting and recognizing objects of interest in order to automatically understand the acquired, complex scene. For this purpose, visual data should be processed through three main stages, namely, object-of-interest detection, segmentation, and recognition.

Most of the existing works about the analysis of multi-camera video streams are focused on tracking multiple, moving objects and apply approaches such as background subtraction (Diaz et al., 2013), local descriptors' matching (Ferrari et al., 2006), occupancy map based on motion consistency (Fleuret et al., 2008), Bayesian framework (Hsu et al., 2013), particle filter (Choi and Yoo, 2013), or Cardinalized Probability Hypothesis Density (CPHD) based filter (Lamard et al., 2013).

In this paper, we propose to introduce the active contour method (Olszewska, 2012), (Olszewska, 2013), (Bryner and Srivastava, 2014), which is efficient both for precisely segmenting and tracking meaningful objects of interest, into a full and automatic system which takes multi-camera video stream inputs and performs visual data processing to recognize multi-view video objects, in context of outdoor video-surveillance.

Our system does not require any camera calibra-

Isabelle Olszewska J..
Multi-camera Video Object Recognition Using Active Contours.
DOI: 10.5220/0005334303790384
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (MPBS-2015), pages 379-384
ISBN: 978-989-758-069-7

tion parameter, since in real-world situations, camera characteristics are not often readily available (Guler et al., 2003), (Mavrinac and Chen, 2013) and/or camera calibration is computationally expensive (Black et al., 2002), (Farrell and Davis, 2008), (Lee et al., 2014). Moreover, our approach does not use any 3D model as in (Sin et al., 2009), and thus reduces the computational burden.

Our system deals with dynamic scenes recorded by standard pan-tilt-zoom (PTZ) static cameras with partially overlapping and narrow fields of view (FOV). In fact, this configuration (Fig. 1) captures a rich variety of real-world situations, where target objects could be seen in both views, leading to a full coverage of some areas like with omnidirectional cameras (Guler et al., 2003), or where targets could be visible in only one view, this latter case being similar to records of a non-overlapping camera network (Kettnaker and Zabih, 1999), (Chen et al., 2008).

On the other hand, the acquired multi-view sequences usually contain noise, complex backgrounds and blurred, moving objects, called objects of interest or foregrounds. Video frames could be subject to illumination variations or poor resolution (Fig. 2).

Hence, the contribution of this paper is threefold:

- the use of active contours for multi-camera video stream analysis;
- the color categorization algorithm for object-of-interest recognition purpose;
- the development of an automatic system based on active contours for multiple, visual target detection and recognition in multi-camera environment.

The paper is structured as follows. In Section 2, we describe our multi-camera stream analysis system (see Fig. 2) based on active contour computation in each view and on the late fusion of the resulting information for fast, multiple video-object recognition. Our approach performance have been assessed on standard, real-world video-surveillance dataset as reported and discussed in Section 3. Conclusions are presented in Section 4.

2 PROPOSED APPROACH

To detect multiple objects of interest in video scenes, we use active contours (Olszewska and McCluskey, 2011; Olszewska, 2011; Olszewska, 2012) which present the major advantage to quickly and precisely delineate an entire targeted object and thus to segment the object as a whole, rather than only disparate pieces as in (Travieso et al., 2014). Hence, we adopt multi-target, multi-feature vector flow active contour approach (Olszewska, 2012) which has been proven to be efficient to detect objects of interest accurately and robustly.

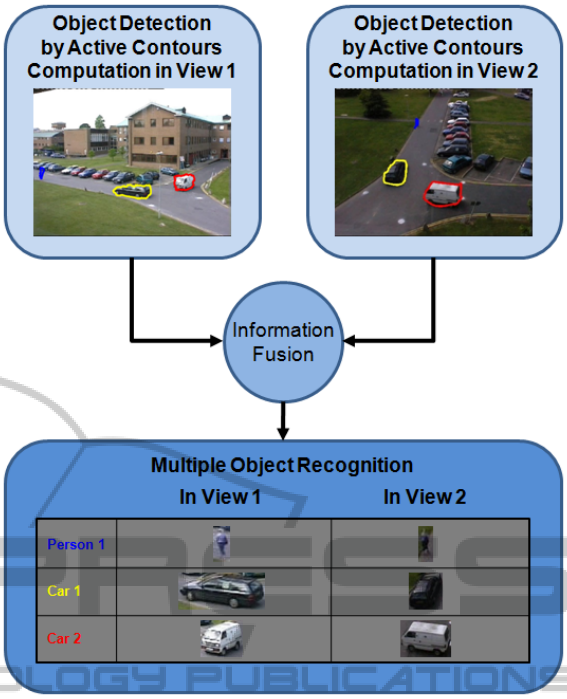
Object Detection by Active Contours Computation in View 1
Object Detection by Active Contours Computation in View 2
Information Fusion
Multiple Object Recognition
In View 1
In View 2
Person 1
Car 1
Car 2

Figure 2: Overview of our active-contour-based approach for multiple-object detection and recognition in a multi-camera environment.

In order to initialized these active contours, we first use the background subtraction method. This could be computed by difference between two consecutive frames (Archetti et al., 2006), by subtracting the current frame from the background (Toyama et al., 1995; Haritaoglu et al., 2000), or combining both frame difference and background subtraction techniques (Huang et al., 2007; Yao et al., 2009).

The latter technique consists in computing in parallel, 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 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, to combine both results in order to extract the foreground in the corresponding view.

To model the background, we adopt the running Gaussian average (RGA) (Wren et al., 1997), characterized by the mean μ_b^v and the variance $(\sigma_b^v)^2$, rather than, for example, the Gaussian mixture model (GMM) (Stauffer and Grimson, 1999; Friedman and Russell, 1997; Zivkovic and van der Heijden, 2004), since the RGA method is much more suitable for real-time tracking.

Hence, the foreground is determined by

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

with

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

and

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

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

Finally, to compute a blob defined by labeled connected regions, morphological operations such as opening and closure (Haralick, 1988) are applied to the extracted foreground F^v, in order to exploit the existing information on the neighboring pixels, in a view v,

$$f^v(x,y)=Morph(F^v(x,y)). \tag{4}$$

Then, an active contour is computed for each frame k in each view v separately, and for each targeted object (see Fig. 2). In this work, an active contour 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

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

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

Once the object has been detected as per previous steps, the colors of the objects are extracted for recognition purpose. In this work, the color concept is defined by 16 basic color keywords defined in SVG standard (Olszewska and McCluskey, 2011). Moreover, the object's color is not the average value over the whole detected object (Parker, 2010), but a set of colors of the different parts of the object.

To compute object's colors, an object is considered to have p parts according to intra-object relations using the *o'clock* concept (Olszewska and McCluskey, 2011), which does not induce an arbitrary division of the target but a partition taking automatically into account object's concavities and convexities. The color in each object's part c_p is found by associating the extracted numeric (R,G,B) value in the red (R), green (G), blue (B) color space to the related semantic name.

Algorithm 1: Inhomogeneous/Homogeneous Color.

```
Given $C=\{c_b\}$, the set of the object's colors such as
$C=CV_1\cup CV_2$ and $b\in\mathbb{N}$;
$L=C\setminus\{c_1\}$ with $c_1=head(C)$; $G=\{c_1\}$;
and $th$, the threshold;

do
repeat
$c_j=head(L)$;
    if $c_j\notin G$ then $G=G\cup\{c_j\}$
    end if
$L=L\setminus\{c_j\}$;
until $L=\emptyset$

return $G$

if $(\#G>th)$
then the color of object is inhomogeneous
else the color of object is homogeneous
end if

end do
```

Next, for object recognition purpose, our system performs a late fusion, approach proven to be more efficient than early fusion where all the cameras are used to make a decision about the detection of the objects of interest (Evans et al., 2013). Indeed, in our system, objects of interest are detected in individual cameras independently. Then, the results are combined on the majority voting principle based on the semantic consistency of the color across multiple camera views (see Fig. 2) and not on the sole geometrical correspondences of objects as in (Dai and Payandeh, 2013). Hence, in our recognition approach, the object's color sets CV_1 and CV_2 in view 1 and 2, respectively, are matched using the Hausdorff distance $d_H(CV_1,CV_2)$, which is computed as follows (Alqaisi et al., 2012):

$$d_H(CV_1,CV_2)=max\bigg(d_h(CV_1,CV_2),d_h(CV_2,CV_1)\bigg), \tag{6}$$

where $d_h(CV_1,CV_2)$ is the directed Hausdorff distance from CV_1 to CV_2 defined as

$$d_h(CV_1,CV_2)=\max_{cv\in CV_1}\min_{cw\in CV_2}d_P(cv,cw), \tag{7}$$

with $d_P(cv,cw)$, the Minkowski-form distance based on the L_P norm, and defined as

$$d_P(cv,cw)=\left(\sum_k(cv_k-cw_k)^P\right)^{1/P}. \tag{8}$$

After the recognition step, objects of interest could be categorized further, in context of video surveillance. Indeed, based on real-world observation, assumption is made that an homogeneous color is associated to an object such as 'car', while an inhomogeneous color corresponds to a 'person' type object. Homogeneous and inhomogeneous colors of detected objects are thus distinguished using the Algorithm 1. Moreover, in the studied camera framework, camera devices are relatively far from the scene, so foregrounds' close-up are not likely to occur, and a car area is usually perceived as greater than a person's one. Thus, the area $\mathcal{A}_o$ inside the active contour of the detected object o could be used to validate the classification of objects of interest into 'car' and 'person' categories by comparing the area defined as $\mathcal{A}_o = max_v\{\mathcal{A}_o^v\}$ against a threshold ta, i.e if $\mathcal{A}_o > ta$ the object is a car, otherwise it is a person.

Hence, the semantic color values of any object detected and segmented with active contours in a frame are automatically compared within views and additionally checked against inhomogeneous and homogeneous criterion in order to achieve a precise target recognition. This technique as well as the use of metrics, such as object's area directly provided by the active contours, ensures the robustness of the system.

3 EXPERIMENTS AND DISCUSSION

To assess our approach, we have applied our system on the standard dataset (PETS, 2001) consisting of video-surveillance dynamic scene recorded by two PTZ cameras whose fields of view are overlapping (Fig. 1) as illustrated e.g. in Figs. 3 (a)-(b). Furthermore, the FOVs do not necessarily end neatly at the edge of a camera's field of vision as observed e.g. in Fig. 3 (d). The resulting, two synchronized videos were captured in outdoor environment and contain 2688 frames each, with an image average resolution of 576x768 pixels.

This database owns challenges of multi-view video stream, as well as quantity, pose, motion, size, appearance and scale variations of the objects of interest, i.e. of the people and cars.

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

Some examples of the results of our system are presented, in Fig. 3, for detection and recognition of multiple objects of interest, which could be either moving persons and cars. These frames present difficult situations such as poor foreground/background contrast, light reflection, or illumination changes. Moreover, some targeted objects could only seen in one of the views as per configuration depicted in Fig. 1. Hence, in Figs. 3 (a)-(b), two objects of interest, one person and one car, respectively, are present in both views. On the other hand, in Figs. 3 (c)-(d), there are six objects of interest, i.e. five persons and one car, in the first view, whereas only four persons and one car are visible in the second view, bringing the number of observed objects of interest to five. Our system copes well with these situations as discussed below.

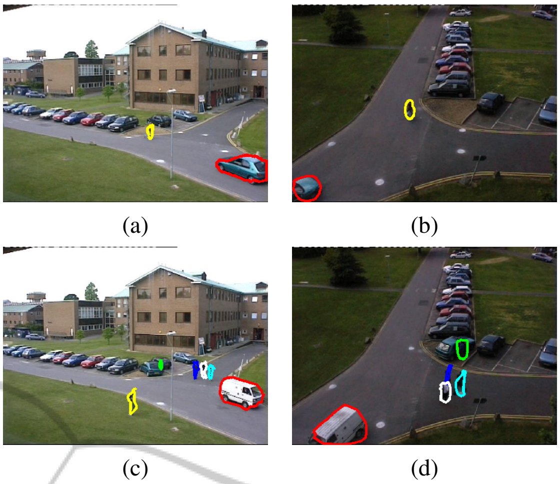

Figure 3: Examples of results obtained with our approach for same scenes in both views. First column: view from the first camera. Second column: view from the second camera.

To measure the detection accuracy of our system, we adopt the standard criteria (Izadi and Saeedi, 2008) as follows:

$$detection\ rate\ (DR) = \frac{TP}{TP+FN}, \tag{9}$$

$$false\ detection\ rate\ (FAR) = \frac{FP}{FP+TP}, \tag{10}$$

with TP, true positive, FP, false positive, and FN, false negative.

The recognition accuracy of our system could be assessed using the following standard criterion:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{11}$$

with TN, true negative.

In Table 1, we have reported the average detection and false alarm rates of our method against the rates achieved by (Izadi and Saeedi, 2008) and (Bhat and Olszewska, 2014), while in Table 2, we have

Table 1: Average detection rate (DR) and average false alarm rate (FAR) of object-of-interests in video frames, using approaches of ◊(Izadi and Saeedi, 2008), □(Bhat and Olszewska, 2014), and our.

	◊	□	**our**
DR	91.3%	91.6%	95.2
FAR	9.5%	4.9%	3.1%

Table 2: Average accuracy of object-of-interest recognition in video frames, using approaches of △(Athanasiadis et al., 2007), □(Bhat and Olszewska, 2014), and our.

	△	□	**our**
average accuracy	85%	95%	96%

displayed the average accuracy of object-of-interest recognition of our method against the rate obtained by (Athanasiadis et al., 2007) and (Bhat and Olszewska, 2014).

From Tables 1-2, we can conclude that our system provides reliable detection of objects of interest in multi-camera environment, and that our multiple-object recognition method is very accurate as well, outperforming state-of-the art techniques.

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

In this paper, we focus on the reliable detection and recognition of multiple objects of interest in multi-stream visual data such as surveillance videos. For this purpose, we have incorporated active contours in the process of automatically analyzing multi-camera, synchronized video sequences with narrow, partially overlapping fields of view. Our approach outperforms the ones found in the literature for both object detection and object recognition.

REFERENCES

Alqaisi, T., Gledhill, D., and Olszewska, J. I. (2012). Embedded double matching of local descriptors for a fast automatic recognition of real-world objects. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'12)*, pages 2385–2388.

Archetti, F., Manfredotti, C., Messina, V., and Sorrenti, D. (2006). Foreground-to-ghost discrimination in single-difference pre-processing. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 23–30.

Athanasiadis, T., Mylonas, P., Avrithis, Y., and Kollias, S. (2007). Semantic image segmentation and object labeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):298–312.

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.

Black, J., Ellis, T., and Rosin, P. (2002). Multi View Image Surveillance and Tracking. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pages 169–174.

Bryner, D. and Srivastava, A. (2014). Bayesian active contours with affine-invariant, elastic shape prior. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 312–319.

Chen, K.-W., Lai, C.-C., Hung, Y.-P., and Chen, C.-S. (2008). An adaptative learning method for target tracking across multiple cameras. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Choi, J.-W. and Yoo, J.-H. (2013). Real-time multi-person tracking in fixed surveillance camera environment. In *Proceedings of the IEEE International Conference on Consumer Electronics*.

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.

Diaz, R., Hallman, S., and Fowlkes, C. C. (2013). Detecting dynamic objects with multi-view background subtraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 273–280.

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.

Farrell, R. and Davis, L. S. (2008). Decentralized discovery of camera network topology. In *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–10.

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.

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.

Friedman, N. and Russell, S. (1997). Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the 13th Conference on Uncertainty in AI.*

Guler, S., Griffith, J. M., and Pushee, I. A. (2003). Tracking and handoff between multiple perspective camera views. In *Proceedings of the 32nd IEEE Workshop on*

Applied Imaginary Pattern Recognition, pages 275–281.

Haralick, R. M. (1988). Mathematical morphology and computer vision. In *Proceedings of the IEEE Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 468–479.

Haritaoglu, I., Harwood, D., and Davis, L. (2000). Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 77(8):809–830.

Hsu, H.-H., Yang, W.-M., and Shih, T. K. (2013). People tracking in a multi-camera environment. In *Proceedings of the IEEE Conference Anthology*, pages 1–4.

Huang, W., Liu, Z., and Pan, W. (2007). The precise recognition of moving object in complex background. In *Proceedings of 3rd IEEE International Conference on Natural Computation*, volume 2, pages 246–252.

Izadi, M. and Saeedi, P. (2008). Robust region-based background subtraction and shadow removing using colour and gradient information. In *Proceedings of the 19th IEEE International Conference on Pattern Recognition*, pages 1–5.

Kettnaker, V. and Zabih, R. (1999). Bayesian multi-camera surveillance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1–5.

Kumar, K. S., Prasad, S., Saroj, P. K., and Tripathi, R. C. (2010). Multiple cameras using real-time object tracking for surveillance and security system. In *Proceedings of the IEEE International Conference on Emerging Trends in Engineering and Technology*, pages 213–218.

Lamard, L., Chapuis, R., and Boyer, J.-P. (2013). CPHD Filter addressing occlusions with pedestrians and vehicles tracking. In *Proceedings of the IEEE International Intelligent Vehicles Symposium*, pages 1125–1130.

Lee, G. H., Pollefeys, M., and Fraundorfer, F. (2014). Relative pose estimation for a multi-camera system with known vertical direction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 540–547.

M. Kamezaki, Y. Junjie, H. I. S. S. (2014). An autonomous multi-camera control system using situation-based role assignment for tele-operated work machines. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 5971–5976.

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

Olszewska, J. I. (2011). Spatio-temporal visual ontology. In *Proceedings of the 1st EPSRC Workshop on Vision and Language (VL'2011)*.

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. 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.

Parker, J. R. (2010). *Algorithms for Image Processing and Computer Vision*. John Wiley and Sons, 2nd edition.

PETS (2001). PETS Dataset. Available online at: ftp://ftp.pets.rdg.ac.uk/pub/PETS2001.

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

Sin, M., Su, H., Savarese, S., and Fei-Fei, L. (2009). A multi-view probabilistic model for (3D) object classes. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1247–1254.

Spehr, J., Rosebrock, D., Mossau, D., Auer, R., Brosig, S., and Wahl, F. M. (2011). Hierarchical scene understanding for intelligent vehicles. In *Proceedings of the IEEE International Intelligent Vehicles Symposium*, pages 1142–1147.

Stauffer, C. and Grimson, W. (1999). Adaptive background mixture model for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Toyama, K., Krumm, J., Brumitt, B., and Meyers, B. (1995). Wallflower: Principles and practice of background maintenance. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 255–261.

Travieso, C. M., Dutta, M. K., Sole-Casals, J., and Alonso, J. B. (2014). Detection and tracking of the human hot spot. In *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pages 325–330.

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785.

Yao, C., Li, W., and Gao, L. (2009). An efficient moving object detection algorithm using multi-mask. In *Proceedings of 6th IEEE International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 354–358.

Zhou, H. and Kimber, D. (2006). Unusual event detection via multi-camera video mining. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 1161–1166.

Zivkovic, Z. and van der Heijden, F. (2004). Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):651–656.