European Review, Vol. 25, No. 2, 246–259 © 2016 Academia Europæa
doi:10.1017/S1062798716000582

Object Extraction and Classification in Video Surveillance Applications

MUHSIN CIVELEK* and ADNAN YAZICI**

*Department of Computer Engineering, Turkish Military Academy, Ankara, Turkey. E-mail: mcivelek@kho.edu.tr
**Department of Computer Engineering, Middle East Technical University, Ankara, Turkey. E-mail: yazici@ceng.metu.edu.tr

In this paper we review a number of methods used in video surveillance applications in order to detect and classify threats. Moreover, the use of those methods in wireless surveillance networks contributes to decreasing the energy consumption of the devices because it reduces the amount of information transferred through the network. In this paper we focus on the most popular object extraction and classification methods that are used in both wired and wireless surveillance applications. We also develop an application for identification of objects from video data by implementing the selected methods and demonstrate the performance of these methods on pre-recorded videos using the outputs of this application.

1. Introduction

Benefiting from multimedia data along with scalar sensor data for surveillance purposes is very popular for both indoor and outdoor security applications. In general, several cameras having different specifications are set up and connected to monitors that are watched by security personnel on a 24/7 basis within the scope of traditional surveillance applications. Together with advances in technology and reduction of camera prices it is now almost impossible to come across a security system at any scale without video.

Multimedia is the most valuable type of data that is produced by the surveillance systems as it is the most effective and easiest way of identifying events and threats. Use of audio-visual content is not only popular in small scale private security systems, but also it is the inevitable part of many large scale civil or military surveillance applications.[1] The main issue regarding the use of multimedia in surveillance applications is the amount of data that are needed to be stored, analysed and transferred. Especially for critical and large-scale applications such as border/homeland security systems, this process must be performed continuously and carefully. For that purpose, organizations establish shifts that are composed of several operators

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

or officers, and those shifts must examine all the multimedia data carefully on a real-time basis in order to detect events and take the necessary actions on time.

Unfortunately, a security staff assignment to track surveillance systems and analyse their outputs may not always be feasible because of the hard or dangerous environmental conditions.[2–4] When the area to be monitored is vast, such as a country border, or there exist life-threatening situations, traditional surveillance applications are not feasible to provide adequate and real-time protection. In such a case the first solution coming to mind is setting up a system that is capable of transferring produced information to a remote headquarters by using wired links. As it can be foreseen easily, this adds a considerable amount of setup and operation cost to the surveillance system.

In some cases, geographical conditions, scalability/mobility concerns or cost-related factors may make the use of wired surveillance systems infeasible. Under those circumstances wireless solutions may be preferred. But there are natural constraints of wireless systems that have to be dealt with, such as limited energy, less processing power, memory and storage spaces and limited communication capability. While using such a constrained system, new issues with regard to production, collection and transfer of multimedia data arise. Since wireless multimedia devices use battery power in order to operate, the main concern in designing and implementing such systems is to prolong device lifetime. For such a device, it is inefficient to capture multimedia data continuously and transfer it to a remote location using a wireless communication interface because those operations consume significant power, which shortens the lifetime of the wireless sensors.

A number of researchers have been working on intelligent wireless surveillance systems that require less human intervention to operate in a reasonable lifetime.[1,5–8] With current processing, memory and sensor technologies, it is possible to implement such smart security devices that involve scalar and multimedia sensors as well as internal processing units. Rather than providing an enormous amount of surveillance data that should be examined by the operators in real time, researchers are focused on smart security systems that process their own data and produce brief results. In this way, the semantic information related with the detected events can be extracted by a security system itself in a faster and finer manner thus, resulting in less human resource usage and more reliable output as well as faster reactions to events. In addition, by using this processing capability in wireless surveillance networks, the amount of data to be transferred decreases, which prolongs the lifetime of the overall wireless sensor networks.

There are various studies on processing audio-visual and scalar sensor data produced by surveillance systems in order to extract the semantic information.[1,9,10] Those studies propose different approaches, methods and architectures in order to produce accurate results in reasonable times. Some common tasks implemented within the scope of those studies are object and event extraction and classification from the collected data. Those tasks are important in order to make the system capable of making its own decision related with threats and then taking necessary actions. The general approach that proposes execution of object detection and

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

classification operations in the sensor nodes is called in-network processing. Audio-visual data produced by sensors are processed by the sensor node itself and specific features of this data are extracted in order to draw conclusions. While doing this, scalar data, if available, may be fused with multimedia data in order to increase the accuracy of results.

Training is another common task that is applied as a pre-phase of object extraction and classification process. The performance of the system is directly related to the selection of an appropriate training set and a classification method. Features that will be used for object identification are determined first and then the system is trained by using the same features of similar data. When new data arrive, the system is able to classify them according to the information that it learns from the training dataset.

In this study we use the most popular methods and approaches that are preferred for object extraction and identification in surveillance networks. We evaluate those methods and approaches according to their compliance with wired and wireless surveillance networks. We also implement an application by using the most common methods that are used in wireless surveillance systems for object extraction and classification. We record outdoor videos and extract ground truth of those videos. Then, we experiment the accuracy and performance of the application experiments by using ground truth data tagged from videos of the application.

The rest of the document is organized as follows. In the second section some methods used at each phase for object identification process are reviewed. In the third section implementations of those methods along with their performance results are given. Finally, we draw some conclusions.

2. Object Extraction and Classification

2.1. Foreground Extraction

An object or event identification process in video starts with segmentation of the captured frame in order to detect the areas of interest. Those areas of interest (named as foreground) involve objects/threats except for the background. Performance of this step is crucial in achieving the desired accuracy at the end of the identification process. There are various methods including lightweight image processing operations that are suitable for devices with low processing capabilities, as well as more complex statistical methods that require more processing power and memory space. In this section we review the most popular methods that are preferred for the object extraction phase of the identification process.

Background Subtraction (BS) is the general name of the most popular foreground extraction methods that are used for moving object detection in video streams. BS is widely used in surveillance applications in order to distinguish the foreground objects from the background model before performing further tasks such as classification, localization or tracking.[5–8] BS process requires keeping a model of the background and using this model in order to extract the foreground pixels from the video frame as shown in Figure 1. There are various approaches to construct this background model, like keeping a stationary background frame, building of an adaptive background

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

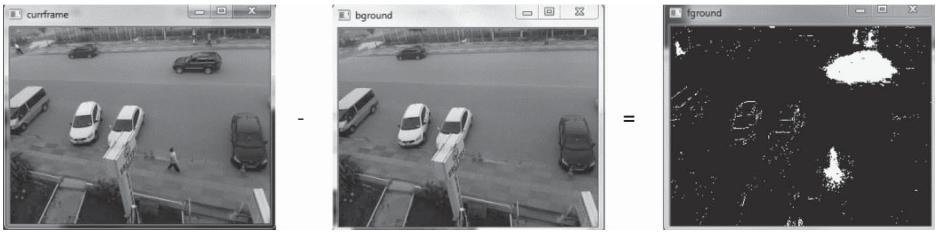


Figure 1. BS operation.

using statistical techniques or temporal modelling. In order to extract the foreground, the difference between the background model and the current video frame is calculated.

The simplest BS approach is getting the difference of the pixel values of the current frame and a stationary reference frame, and then applying a threshold to the result in order to detect the non-background areas.[11] This approach is not suitable for outdoor surveillance applications, in particular, because it is not possible to assume the background model as static. Light changes, moving background objects such as rustling leaves or changes in camera viewing angle require updates in the background model. For that reason statistical methods such as the application of Gaussian Average or Mixture of Gaussians,[12] are used effectively in order to keep the background updated. Using these methods, a pixel's background value is calculated periodically by applying statistical formulas such as averaging or probability density functions.

Temporal frame differencing is another effective approach that is considered as an adequate method for background modelling in the study of Cucchiara *et al.*[13] This method proposes keeping the last few frames as the background model, which requires additional space. A median value is calculated by using those frames, and the foreground is detected according to this median value. This approach is very successful while detecting moving objects but it fails when the foreground object stops moving.

One of the most preferred background modelling methods is multimodal background modelling, which deals with multiple background objects at the same location at different time frames.[14] A pixel is modelled with a weighted combination of several Probability Density Functions (PDF) rather than a single PDF. Because of that approach, this method is also called as Mixture of Gaussians. In practice, the number of PDFs is set to be between 3 and 5. Simple methods such as Gaussian average offer acceptable accuracy while achieving a high frame rate and having limited memory requirements. But more complex methods such as Mixture of Gaussians prove very good model accuracy.[12]

2.2. Post-processing

As stated above, the success of the identification process is closely related with the success of the object extraction phase. Although BS methods provide a foreground

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press



Figure 2. Application of morphological operations on foreground.

view that contains objects of interest, this view most probably contains noise that decreases the accuracy and performance of the classification process. To prevent this, the residue and noise in the foreground image should be removed before classification. Besides noise cleaning, foreground objects need to be concretized for a better classification. For that purpose the primary morphological operations erosion-$\ominus$ and dilation-$\oplus$ are applied on the binary image.[15] Application of those operations one after another is called opening and closing operations. Performing those operations on the binary image means probing the image with a structuring element in order to enlarge the foreground areas and shrinking the holes or vice versa.

The median filter is one of the most popular nonlinear filters for removing impulse noise because of its good de-noising power and computational efficiency.[16] It is widely used for removal of salt and pepper type of noises in an image. The idea of the median filter is to replace the pixel value with the median of the neighbouring values that forms a window. Application of those morphological operations on a binary foreground image, which is extracted by BS, is shown in Figure 2.

2.3. Object Extraction/Segmentation

Image segmentation means partitioning the image into meaningful regions. All of the pixels in a region are similar with respect to some characteristics or computed property, such as colour, intensity, or texture.[17] The foreground image that is extracted by using BS may contain several distinct objects. Each of those objects is processed separately during classification. Segmentation algorithms are applied on foreground images in order to extract the regions of actual objects.

Fu and Mui categorize image segmentation techniques into three classes: thresholding, edge detection, and region extraction.[18] In the thresholding approach, an image is partitioned based on one or more threshold values. This method can be used as a simple segmentation approach for partitioning an image into background and foreground areas or as a more complicated approach for extracting multiple objects using multiple threshold values.

Edge detection techniques transform images to edge images benefiting from the changes of grey tones in the images.[17] There exist three main edge detection

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press



Processed foreground | Segmentation

Figure 3. Segmented foreground.

techniques in literature: Roberts, Prewitt and Sobel. Those techniques detect gradient changes by calculating grey level differences among neighbouring pixels. A segmented foreground picture using edge detection algorithm is shown in Figure 3.

Another effective technique for segmentation of images is the iterative pixel classification in which decisions related to pixels are made according to its neighbours iteratively. One application of this method is the connected component labelling operation in binary images. For a binary image, represented as an array of d-dimensional pixels or image elements, the connected component labelling is a process of assigning labels to the BLACK image elements in such a way that adjacent BLACK image elements are assigned to the same label.[19]

Even just object extraction can contribute to the smartness of surveillance applications. For example, studies such as Refs. 9,10,20–23 are focused on automatic detection of moving or abandoned objects in surveillance videos using those methods. Those methods introduced up to now facilitate the work of operators by a considerable amount. Rather than dealing with real-time, continuous, raw multimedia data, the operators receive and analyse only a small portion of data that consists of the threats. In addition, for a wireless multimedia sensor network, the significant decrease in the amount of transferred data results in prolonging the device lifetime. For classification or tracking purposes, further operations are needed to be performed. These operations are explained in the following sections.

2.4. Features of Objects

For building a smart surveillance system, the next step is to identify the extracted object. For that purpose, the specific features of the object are used. These features help to identify the type of objects. The features may be simple structures such as point, corner, and edge as well as more complicated structures such as texture, blob or the object itself.

Features can be categorized as shape-based features, such as aspect (width/height) ratio, complexity of the shape (perimeter^2/area), edges, corners; texture-based features, such as Gabor features; and motion-based features, such as speed.[24] Öztarak *et al*. use the width/height ratio of the bounding rectangle of the object as

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

well as speed in their study in order to classify whether the detected object is a person, a vehicle or an animal.[25] Lin and Wei also use width/height ratio and walking rhythm in order to classify the objects as a pedestrian, vehicle or motorcycle in a street surveillance system.[26] Brown utilizes compactness and dispersedness features in order to distinguish humans from vehicles for an arbitrary scene.[27] Since the shape-based features are easy to extract, they are preferred in power-constrained applications such as wireless surveillance systems.

Although shape-based features are very effective in object identification processes, most of the features are affected by illumination changes, scaling, camera position/angle, or the direction of the object. In order to extract more stable and invariant features, several methods are proposed. One of those methods is Scale Invariant Feature Transformation (SIFT), which is proposed by Lowe in his study.[28] In the SIFT method; an image is transformed into local feature vectors by applying a staged filtering technique. Speeded Up Robust Features (SURF) is a similar approach proposed in the study of Bay *et al*.,[29] which claims to be more efficient in terms of feature extraction. SURF uses Hessian matrix approximation on an integral image rather than the original image, in order to detect the feature points.[30] An example of similarity matching based on SURF key points is shown in Figure 4. Histograms of Oriented Gradient (HOG) descriptors are also scale invariant features that are originally proposed for human detection in images.[31] The HOG method is based on the idea that an object can be characterized well enough by the distribution of local intensity gradients or edge directions.

Kim *et al*. use SURF in order to recognize the objects in surveillance videos.[32] In their study, besides object recognition, they also perform queries based on the local feature descriptors of the objects. Zhou *et al*. propose an algorithm in their study to track multiple types of objects in video.[33] They use SIFT features and similarity matching methods to match the object among video frames. Since it takes time and requires space and processing power to extract and store, usage of local features in constrained environments such as wireless multimedia sensors is a challenging task.

2.5. Classification

Object recognition is the most critical phase that realizes the goals of building a smart video surveillance system. In this phase the type or identity of the foreground object in the video is determined by using its extracted features. There are various classification categories such as Minimum Distance, Maximum Likelihood, Fuzzy Sets, Neural Networks and Decision Trees. Those classifiers are mainly categorized based on the existence of training data or distribution of collected information.

One of the most popular classification approaches is to use datasets that contain a sufficient number of samples for each object type, in order to train the surveillance system.[34] The system is able to determine the class of the new sample by using statistical techniques such as maximum likelihood, means, etc. This type of classification is called as supervised classification. Other work uses different methods of supervised classification for real-time object recognition or tracking purposes.[27,35–41] When training data are not available or sufficient to make a decision, an unsupervised

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

classification method is preferred. The unsupervised classification methods involve usage of clustering techniques, such as k-means, centroid method, and group average method. By using these techniques, spectrally separable classes are determined. Other studies[42,43] cover unsupervised classification methods for moving object classification and event identification.

When the distribution of the collected data fits into a certain pattern such as a Gaussian distribution, it is possible to guess classification parameters such as mean vector or covariance matrix. Under these circumstances, the classes of objects are determined by applying statistical methods that use those parameters. These types of classifiers are called as parametric classifiers. If it is not possible to fit the distribution of the data into any pattern, then parameter assumption cannot be performed. The classification is done by applying clustering techniques such as k-Nearest Neighbourhood (kNN) or Parzen window. These types of classifiers are called non-parametric classifiers.

One of the popular classification methods is kNN, which is a non-parametric (does not make any assumption on the distribution of data) and lazy learning (does not use training data points for generalization) method for classification. The classification is based on the distance between the feature descriptors and the descriptors of the samples. Another effective classifier is support vector machine (SVM),[44] which is also a non-parametric and supervised classification algorithm. The classification is based on the hyperplane with the maximum distance to training data points. These classifiers, shown in Figure 5, are preferred for real-time classification applications in resource-constrained environments because of lower computation complexity, simplicity and less space requirements.

An Artificial Neural Network (ANN) is also a popular classifier, and has a computational model similar to the human brain. The classifier has a layered network architecture and each layer is composed of nodes (neurons) connected to each other with a few inputs and outputs (edges). During the training phase, the edges are weighted to produce the correct output label. After that, the network forms a function, and then using that function the network is able to predict the output for a given input. Due to their computational complexity, the ANN type classifiers do not fit well to resource-constrained applications, as shown in the study of Munder and Gavrila.[45]

3. Application

For testing the performance of various automatic detection and classification methods, we use an application and implement it using the C++ programing language. We use OpenCV open source library for image processing operations.[46] We record a number of low resolution surveillance videos and produce ground truth for those videos. Then, we execute the application in order to extract and classify the moving objects from videos.

3.1. Object Extraction and Classification Procedure

We use the BS MoG method to extract foreground objects. Morphologic operations that are stated in the previous part are applied into foreground images. Then we split

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

Table 1. Summary of the features and classification methods used in this study.

Approach	Features	Method
1	Width/Height, Compactness, Dispersedness, Blob Ratio	KNN
	SURF	ANN
2	Width/Height, Compactness, Dispersedness, Blob Ratio	SVM
3	HOG	ANN

the images into the moving objects by using the Canny edge detection algorithm, which was developed by John F. Canny.

After retrieval of the foreground objects, we perform feature extraction operations on those objects. The shape-based features (width/height ratio, compactness, dispersedness and blob ratio) as well as the low level features (SURF, HOG) are obtained for classification at the end of the feature extraction process. We train the system at the beginning of the application. As the training dataset we use images that are cropped from the first part of the videos that we work on. We extract and store the same features of the training dataset images and label those images in three different categories, person, group and vehicle.

We perform a classification by using three different approaches. In the first approach we apply a cascade classification method that uses the shape-based features and SURF together. We execute the kNN algorithm first, through which we find the five nearest training dataset items by using the shape-based features. At the end of the kNN algorithm, if four or five of the neighbours belong to the same class, then we consider that kNN has a classification result with a certainty above a threshold. Otherwise a second level classification is performed using the SURF features of the nearest neighbours and the extracted object, as shown in Figure 4. The label of the most similar object is considered as the result of classification.

As the second approach, we perform classification by using the shape-based features only. We do not include any low-level feature. We use the SVM classifier because of its good efficiency and performance. And, as the third approach, we extract the HOG features of objects and match them with the HOG features of the training set using Euclidian distance (nearest neighbour-ANN method). The summary of the methods and features that are used for the three approaches are given in Table 1. The output of the application is shown in Figure 6.

3.2. Results

The performance of the application is measured for each type of object separately. The results are given in Table 2. The results show that classification with SVM using the shape-based features is quite satisfactory. Meanwhile, by using cascade classification such as kNN + SURF, it is possible to achieve a 10% contribution to the accuracy of classification in recognizing a group of people objects. The performance results show that HOG itself is not appropriate for this type of an application because

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

Table 2. Performance results

App.	Person	Group	Vehicle
1	70%	79%	96%
2	84%	53%	98%
3	40%	35%	96%

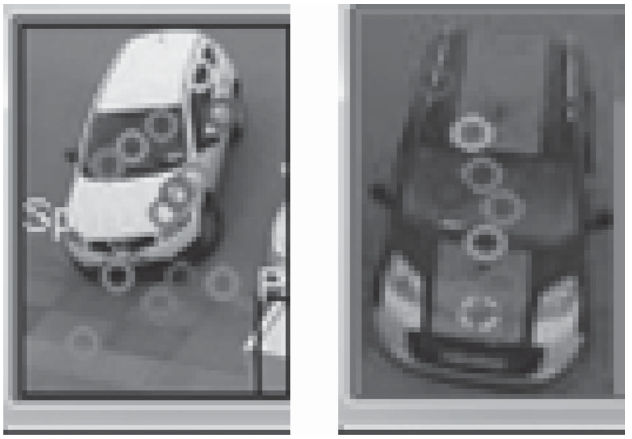


Figure 4. Similarity matching using SURF key points.

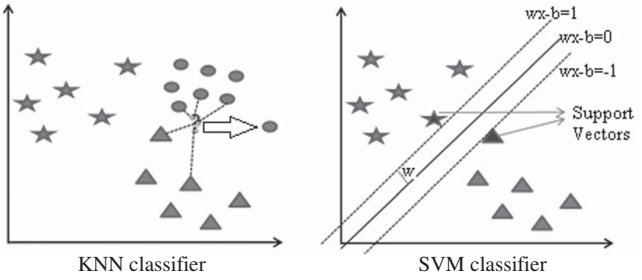


Figure 5. Popular classifiers.

of its poor performance. We also expect that the performance of cascade classification and scalability of the proposed system increases quite a bit when the SVM method is used, instead of the kNN method.

4. Conclusion

In this study we present a review of the most popular approaches used in video surveillance applications. The classification methods are discussed, implemented and

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press



Figure 6. C++ application for object extraction and classification from surveillance videos.

tested by using them for object extraction and recognition in surveillance videos. Those methods are shown to be effective for building smart security systems, requiring little or no human intervention to operate. These systems may be used for achieving security in dangerous and/or safety critical areas where human staff assignment would be infeasible. For wireless multimedia sensor network applications, the methods reviewed in this paper, such as SVM using the shape-based features and kNN with SURF features, introduce efficient solutions and greatly decrease the amount of transferred data on the wireless multimedia sensor network, thus prolonging the network lifetime.

Acknowledgement

This work was supported by a research grant from TUBITAK with the grant number 114R082.

References

1. V. Sharif and E.C. Potdar (2009) Wireless multimedia sensor network technology: a survey. *Seventh IEEE International Conference on Industrial Informatics*, pp. 606–610.
2. A. Filippopoulitis, L. Hey, G. Loukas, E. Gelenbe and S. Timotheou (2008) Emergency response simulation using wireless sensor networks. *Proceedings of the First International Conference on Ambient Media and Systems (ICST)*, 21.
3. N. Dimakis, A. Filippopoulitis and E. Gelenbe (2010) Distributed building evacuation simulator for smart emergency management. *The Computer Journal*, **53**(9), pp. 1384–1400.
4. E. Gelenbe and F.-J. Wu (2012) Large scale simulation for human evacuation and rescue. *Computers and Mathematics with Applications*, **64**(12), pp. 3869–3880.
5. S. Vasuhi, A. Fathima, S. Shanmugam and S.V. Vaidehi (2012) Object detection and tracking in secured area with wireless and multimedia sensor network. *Networked Digital Technologies* (Berlin, Heidelberg: Springer), pp. 356–367.
6. H. Öztarak, K. Akkaya and A. Yazici (2013) Efficient tracking of multiple objects in wireless multimedia sensor networks. *Ad-hoc & Sensor Wireless Networks*, **19**(3-4), pp. 241–276.

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

7. K. Joshi and T. Darshak (2012) A survey on moving object detection and tracking in video surveillance system. *International Journal of Soft Computing and Engineering (IJSCE)*, pp. 2231–2307.
8. K. Kim and D. Larr (2011) Object detection and tracking for intelligent video surveillance. *Multimedia Analysis, Processing and Communications* (Berlin, Heidelberg: Springer), pp. 265–288.
9. M. Magno, F. Tombari, D. Brunelli, L. Di Stefano and L. Benini (2009) Multimodal abandoned/removed object detection for low power video surveillance systems. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 188–193.
10. W.T. Chen, P.Y. Chen, W.S. Lee and C.F. Huang (2008) Design and implementation of a real time video surveillance system with wireless sensor networks. *IEEE Vehicular Technology Conference*, pp. 218–222.
11. L. Wei, W. Xiaojuan, M. Koichi and Z.H. An (2010) foreground detection based on optical flow and background subtract. *International Conference on Communications, Circuits and Systems (ICCCAS)*, pp. 359–362.
12. M. Piccardi (2004) Background subtraction techniques: a review. *IEEE International Conference on Systems, Man and Cybernetics*, pp. 3099–3104.
13. R. Cucchiara, C. Grana, M. Piccardi and A. Prati (2003) Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(10), pp. 1337–1342.
14. C. Stauffer and W.E.L. Grimson (1999) Adaptive background mixture models for real-time tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**.
15. R.M. Haralick, S.R. Sternberg and X. Zhuang (1987) Image Analysis using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**(4), pp. 532–550.
16. R.H. Chan, C.W. Ho and M. Nikolova (2005) Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Transactions on Image Processing*, **14**, pp. 1479–1485.
17. N. Senthilkumaran and R. Rajesh (2009) Edge detection techniques for image segmentation – a survey of soft computing approaches. *International Journal of Recent Trends in Engineering*, **1**(2).
18. K.S. Fu and J. K. Mui (1981) A survey on image segmentation. *Pattern Recognition*, **13**(1), pp. 3–16.
19. M.B. Dillencourt, H. Samet and M. Tamminen (1992) A general approach to connected-component labeling for arbitrary image representations. *Journal of the Association for Computing Machinery*, **39**, pp. 253–280.
20. A. Singh, S. Sawan, M. Hanmandlu, V.K. Madasu and B.C. Lovell (2009) An abandoned object detection system based on dual background segmentation. *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 352–357.
21. X. Zheng, Y. Xiaoshi, N. Li and H. Wu (2009) An automatic moving object detection algorithm for video surveillance applications. *International Conference on Embedded Software and Systems*, pp. 541–543.
22. C. Poppe, S. De Bruyne, T. Paridaens, P. Lambert and R. Van de Walle (2009) Moving object detection in the H. 264/AVC compressed domain for video surveillance applications. *Journal of Visual Communication and Image Representation*, **20**(6), pp. 428–437.
23. A. Bayona, J.C. Sanmiguel and J.M. Martínez (2010) Stationary foreground detection using background subtraction and temporal difference in video

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

surveillance. *17th IEEE International Conference on Image Processing (ICIP)*, pp. 4657–4660.
24. T. Tuytelaars and K. Mikolajczyk (2007) Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, **3**(3), pp. 177–280.
25. H. Öztarak, K. Akkaya and A. Yazici (2013) Efficient tracking of multiple objects in wireless multimedia sensor networks. *Ad-hoc & Sensor Wireless Networks*, **19**(3-4), pp. 241–276.
26. H.Y. Lin and J.Y. Wei (2007) A street scene surveillance system for moving object detection, tracking and classification. *IEEE Intelligent Vehicles Symposium*, pp. 1077–1082.
27. L.M. Brown (2004) View independent vehicle/person classification. *Proceedings of the ACM Second International Workshop on Video Surveillance & Sensor Networks*, pp. 114–123.
28. D. Lowe (1999) Object recognition from local scale-invariant features. *International Conference on Computer Vision*, **2**, pp. 1150–1157.
29. H. Bay, T. Tuytelaars and L.V. Gool (2008) SURF: speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, **110**(3), pp. 346–359.
30. P. Viola and M. Jones (2001) Rapid object detection using a boosted cascade of simple features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **1**, pp. 511–518.
31. N. Dalal and B. Triggs (2005) Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **1**, pp. 886–893.
32. D. Kim, S. Rho and E. Hwang (2012) Local feature-based multi-object recognition scheme for surveillance. *Engineering Applications of Artificial Intelligence*, **25**(7), pp. 1373–1380.
33. H. Zhou, Y. Yuan and C. Shi (2009) Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding*, **113**(3), pp. 345–352.
34. D. Lu and Q. Weng (2007) A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, **28**(5), pp. 823–870.
35. G.L. Foresti (1998) A real-time system for video surveillance of unattended outdoor environments. *IEEE Transactions on Circuits and Systems for Video Technology*, **8**(6), pp. 697–704.
36. M.H. Zaki, T. Sayed and M. El Esawey (2015) A mixed urban traffic road-users classification based on automated video data analysis. *Advances in Transportation Studies*, **35**.
37. M. Tschentscher, C. Koch, M. Konig, J. Salmen and M. Schlipsing (2015) Scalable real-time parking lot classification: An evaluation of image features and supervised learning algorithms. *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.
38. S. Boragno, B. Boghossian, D. Makris and S. Velastin (2008) Object classification for real-time video-surveillance applications. *Fifth International Conference on Visual Information Engineering,* pp. 192–197.
39. L. Zhang, S.Z. Li, X. Yuan and S. Xiang (2007) Real-time object classification in video surveillance based on appearance learning. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
40. C. Cramer and E. Gelenbe (2000) Video quality and traffic QoS in learning-based subsampled and receiver interpolated video sequences. *IEEE Journal on Selected Areas in Communications*, **18**(2), pp. 150–167.

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press

41. C. Cramer, E. Gelenbe and H. Bakircioglu (1996) Low bit-rate video compression with neural networks and temporal sub-sampling. *Proceedings of the IEEE*, **84**(10), pp. 1529–1543.
42. A. Naftel and S. Khalid (2006) Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space. *Multimedia Systems*, **12**(3), pp. 227–238.
43. Z. Zhang, Y. Cai, K. Huang and T. Tan (2007) Real-time moving object classification with automatic scene division. *IEEE Conference on Computer Vision and Pattern Recognition*, **5**, pp. 149–152.
44. C. Cortes and V. Vapnik (1995) Support-vector networks. *Machine Learning*, **20**(3), pp. 273–297.
45. S. Munder and M.D. Gavrila (2006) An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(11), pp. 1863–1868.
46. OpenCV Website. URL:http://opencv.org, last accessed May 2015.

About the Authors

Muhsin Civelek received his BSc degree in Computer Science Engineering from Marmara University in 1998 and MSc degree in Software Management from Middle East Technical University in 2006. He worked as a wide area network administrator and telecommunication network administrator for 11 years. He is currently a PhD student at Computer Engineering Department in Middle East Technical University. He works as an instructor in the Computer Engineering Department of the Turkish Military Academy where he also teaches Computer Network courses.

Adnan Yazici received a PhD degree in Computer Science from Department of EECS, Tulane University, New Orleans, LA, USA, in 1991. He was a visiting professor with Tulane University, New Orleans, LA, USA, between 1998 and 2000. He is currently chair of the Department of Computer Engineering, Middle East Technical University (METU), Ankara, Turkey, and is the Director of the Multimedia Database Laboratory of the Computer Engineering Department. He has published more than 180 international technical papers and co-authored two books, *Fuzzy Database Modelling* (New York: Springer, 1999) and *Uncertainty Approaches for Spatial Data Modelling and Processing: A Decision Support Perspective* (New York: Springer, 2010). His current research interests include intelligent database systems, fuzzy database modelling, index structures, spatio-temporal, multimedia and video databases, and wireless multimedia sensor networks. Dr Yazici received the 2011 IBM Faculty Award and the Parlar Foundation 2001 Young Investigator Award. He was a Conference Co-Chair of the 23rd IEEE International Conference on Data Engineering (ICDE-2007), Conference Co-Chair of the 38th International Conference on Very Large Data Bases (VLDB 2012), Program Co-Chair of the International Conference on Flexible Query Answering Systems (FQAS 2011), and the General Chair of the IEEE Conference on Fuzzy Systems (FUZZ-IEEE 2015).

https://doi.org/10.1017/S1062798716000582 Published online by Cambridge University Press