Human Object Classification Based on Nonsubsampled Contourlet Transform Combined with Zernike Moment

Luu The Phuong and Nguyen Thanh Binh(✉)

Faculty of Computer Science and Engineering,
Ho Chi Minh City University of Technology, VNU-HCM, Vietnam
phuong74l0@gmail.com, ntbinh@hcmut.edu.vn

Abstract. The surveillance systems are more and more popular because of the security needs, but the traditional ones do not meet human's expectation. This paper proposes the algorithm to classify objects mainly based on their contour property which are represented by the amplitude of zernike moment on non-subsampled contourlet transform of a binary contour image. This feature shows promising results by just a simple association with the aspect ratio but gives high accuracy. The aspect ratio helps contour feature in case that the image is too blurred to extract the object's contour. It also plays as a weak filter with nearly no more computational cost except for a division to support contour feature when applying gentle boost algorithm.

Keywords: Object classification · Zernike moment · Nonsubsampled contourlet transform

1 Introduction

Nowadays, the surveillance systems are more and more popular because of the security needs, but the traditional ones do not meet human's expectation. Human must spend money not only on such devices as cameras, monitoring rooms but also on those who sit in front of the monitors and check every second if something happens. That's why we need the intelligent surveillance systems, which can automatically do monitoring tasks. In that tendency, many researchers focus on image and video processing to answer that question. Building an intelligent surveillance system can be split into four main challenges: moving object detection, object classification, object tracking and event or behavior recognition.

Object classification can be applied in many purposes such as security surveillance systems for airports, train stations, schools, or buildings of government etc.; or traffic control systems like automatic traffic signal systems, street-crossing safety systems, traffic density statistics systems etc. Object classification brings the class information of objects for a system to assess the situation. For example, if a situation in which an object moves fast toward another one happens, even you cannot guess what it means without their class information. If that's a man and an airplane, it may be a late coming. If that's a car and a crossing human, it may potentially be a traffic accident. Many expected jobs of

P.C. Vinh and V. Alagar (Eds.): ICCASA 2015, LNICST 165, pp. 212–222, 2016.
DOI: 10.1007/978-3-319-29236-6_21

the intelligent surveillance systems just cannot be done without object classification problem solved.

The overview of object classification has three stages: moving object detection, feature extraction and classification. A lot of research can be found in recent years such as Setitra [1] and Karasulu [2] some of which are to review about proposed methods for each stage, and their strong and weak points. The successive research will improve and/or merge one or more methods to reach better results. For the moving object detection, Lin [3] use Gaussian Mixture Model (GMM) and Elhoseiny [4] use non-parametric Kernel Density Estimation (KDE), which are both quite old now (more than ten years). GMM is quite fast that it can meet real time requirement for outdoor scenes, but it's not robust to many types of noise like illumination changes, removed object – ghost effect. KDE is quite complex with many sub algorithms for each step that brings it robustness for many types of noise like illumination changes, occlusion, shadow of object etc., and the slow speed, too. The proposal uses the recently proposed algorithm FTSG which is shown better than many state-of-the-art algorithms in [5]. It gives a medium speed (10 fps for a 320 × 240 video) and the robustness to many types of noise mentioned.

For feature extraction and classification, Lin [3] use four features like: speed, width, RMI and CAR. The speed and RMI features are more complex because of requiring tracks of previous frames for calculating. The classification algorithm bases on thresholds, which may be less accurate in case of different video resolution, camera distance, and perspective. Vishnyakov [6] bases on statistics and Logistic Regression training algorithm, which is tested with a large number. The high accuracy is quite promising but the algorithm uses only statistics as feature so it does not exploit much information from the object image. Elhoseiny [4] uses many features which exploit much information from the object image, so it gives a better result with the support of Adaboost algorithm. But many features cost high computation and the result shows that it requires a lot of training (80 % dataset), which may be scene-dependent. It may be less accurate for completely new scene. In this paper, we proposed a new method for object classification. The proposed method uses contour property as the main feature, which is quite robust to video resolution, distance or perspective because the object can be rotated and characterized the shape of its contour to make training data. This main feature is associated with the aspect ratio to build a better classifier with gentle boost, which is simple and promising to be effective to classify objects. The rest of the paper is organized as follows: we described the background of selecting a new generation wavelet transform and zernike moment in Sect. 2; the proposed method is shown in Sect. 3; the result and conclusion of the paper are presented in Sects. 4 and 5 respectively.

2 Background

2.1 Select a New Generation Wavelet Transform

The wavelet transform (WT) is a good tool to provide time - frequency representation of the signal, so its applications appear in many fields. But WT is not really perfect; it's

only good at representing point singularities. If they are lines, edges or textures, WT has a lot of redundant coefficients and is computationally expensive. Ridgelet transform [9] is proposed with better representation for linear singularities. But most of the real world applications do not show straight - line singularities, especially the object's contour. Curvelet transform [10] takes its place to solve the problem by partitioning the image then applying Ridgelet to each part. The idea is a curve split into many parts which correspond to lines. But curvelet also has two drawbacks: first, not optimal for sparse approximation of curve beyond C^2 singularities and second, highly redundant. Contourlet transform [11] is built from a discrete domain first, then extend to the continuous domain, it has lower redundancy and a faster discrete implementation version than curvelet. But contourlet is just multidirectional and multi-scale but not shift - invariant, which causes pseudo-Gibbs phenomena visible on the decoded image by high compression ratio. Nonsubsampled Contourlet transform (NSCT) [12] brings shift - invariance for contourlet with the trade-off of more redundancy. As NSCT is used for contour detection, this redundancy is not a drawback but even gives better results.

The important feature of the proposal bases on object's contour and NSCT is chosen to extract the object's contour. NSCT belongs to the family of the new generation wavelet transform. The NSCT development through its main predecessors may start at wavelet transform (WT) [7] as in [8].

NSCT comprises two parts: a Nonsubsampled Pyramid structure (NSP) which gives multi-scale property and a Nonsubsampled Directional Filter Bank (NSDFB) structure that brings directional property. Both of them are shift - invariant due to Nonsubsampled filter banks. NSP comprises four filters: $H_0(z)$ and $H_1(z)$ are low and high pass decomposition filters, and $G_0(z)$ and $G_1(z)$ are low and high pass reconstruction filters. In an ideal case, the passband supported by low filter at the jth stage is $[(-\pi/2^j), (\pi/2^j)]^2$ and the corresponding high filter supports the complement region $[(-\pi/2^{j-1}), (\pi/2^{j-1})]^2 \backslash [(-\pi/2^j), (\pi/2^j)]^2$. The filters of first stage are upsampled to obtain filters for next stages, so it gives multi-scale property without any more filter design. NSDFB has filters constructed from fan filter banks. The upsampled fan filters of the second stage supports checker-board frequency, when combined with first stage filters will give the four directions in frequency decomposition. The properties of NSCT like multi-scale, multi-direction, shift-invariant make it very suitable for contour detection.

2.2 Zernike Moment

In image processing, moments can be used as object recognition feature which is invariant under translate, scale and/or rotation. There are many types of moments, which can be categorized in two groups [13]:

Non orthogonal moments: Cartesian moments, rotational moments and complex moments are some of this type. This moment type has non orthogonal basis, so as the high redundancy. The invariance is up to the specific moments.

Orthogonal moments: legendre moments, zernike moments and pseudo - zernike moments are some examples. With orthogonal basis, they have no redundancy between

moments. Legendre moments are highly sensitive to noise. Zernike moments have better reconstructed image than other moments like Cartesian Moments with same computational accuracy. It is easier to choose the order of moments based on the contribution of each moment due to orthogonality. This property also brings better performance. Zernike is only rotationally invariant, so it will need to be normalized by another moment to be invariant under translation and scale too.

Pseudo - zernike moments are an improved version of Zernike with most of the advantages of Zernike and a higher noisy tolerance, but they are not popular because of high computational cost.

3 Object Classification Based on NSCT Combined with Zernike Moment

In this section, we proposed the method for object classification based on NSCT combined with zernike moment. The overall of the proposed method is presented as Fig. 1.

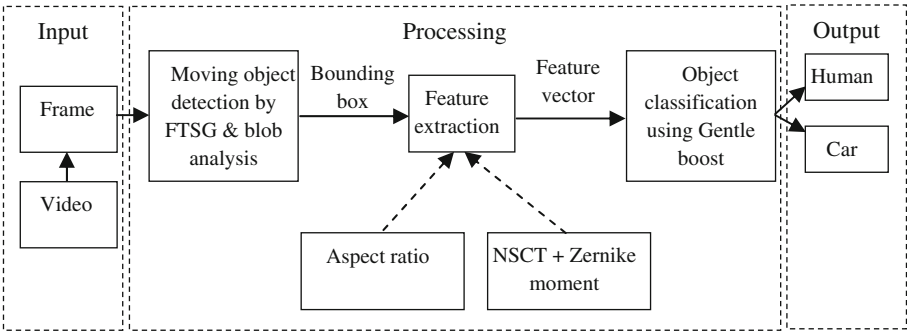

Fig. 1. The overall of the proposed method.

The proposed method has three stages: moving object detection, features extraction and object classification. The input data are videos which have the same serial frames. The outputs are classification of a human or car.

3.1 Moving Object Detection

In this stage, we use Flux Tensor with Split Gaussian models (FTSG) combined with blob analysis for moving object detection. The FTSG algorithm has three main steps: (i) Flux Tensor (FT) and Split Gaussian Model (SG), (ii) Fusion of FT and SG, (iii) Stopped and removed objects classification. Most of detailed implementation can be found in [5].

Flux Tensor is a temporal variation of optical flow field in a local 3D spatiotemporal image volume. It can classify pixels as foreground or background based on

temporal gradient changes information incorporated at each step without calculating eigenvalue. Detailed implementation can be found in [14, 15].

Split Gaussian Model is based on GMM [16], but has the model with variable K Gaussian for background and a separated model with one Gaussian only for foreground. This separation prevents background from corrupted by foreground pixels, brings better adaption for complex background environment (static and dynamic). Some not mentioned default values can be used from [16]. In [5], the parameter σ_t^2 is the variance at the time t and defined as:

$$\sigma_t^2 = (1-\alpha)M\sigma_{t-1}^2 + M\alpha(I_t - \mu_t)^T\alpha(I_t - \mu_t). \quad (1)$$

In here, we defined

$$\sigma_t^2 = (1-\alpha)M\sigma_{t-1}^2 + M\alpha(I_t - \mu_t)^2. \quad (2)$$

This variance plays as standard to check if a pixel value is considered to be matched to the model or not. T in (1) is from [16], which is the minimum proportion of data that should be accounted for the model. Higher T allows multi-modal distribution to adapt to small motions like leaves, grass or flags in the wind. α is the learning rate at each time a new pixel is updated to the model. Here the proposal chooses to use the version in (2). This is equivalent to T = 1 and just α multiplied to variance of the pixel. By this way, the model adapts to small motions more than just a small portion at each time like in (1). But this way also means that we completely lie on α to update this variance. A value of 0.004 for background and 0.5 for foreground is small enough to avoid this defect.

Fusion of FT and SG in step 2 is complementary. FT is robust to illumination changes, but fails to detect stopped objects. SG is sensitive to illumination changes, but can handle stopped objects due to its background model.

The last step is to help SG to recognize revealed background pixels to incorporate into its model after object moving to another location. By chamfer matching [17] the contour of input image and background model image of SG with foreground mask, it can detect the old object in background model image and foreground mask but not input image.

The algorithm to get moving objects as a bounding box from foreground mask is from Matlab's vision blob analysis [18]. This built-in class will return bounding box information (x, y, width, height) of moving objects and ignore small blobs which are false alarm if any.

3.2 Feature Extraction

Each bounding boxed object will be calculated two values as feature vector: Aspect ratio and Zernike NSCT value. Aspect ratio (AR) is:

$$AR = W/H \tag{3}$$

where W, H are width and height of the bounding box.

AR feature is very simple but quite efficient. The rule of it is that the bounding box of human usually has less in width than height, and the opposite for car. That assumption is usually correct in outdoor scenes, because human usually appears in the standing pose like walking, running. For the car, most cars in real world are more in width than in height, and they appear in one pose only. This feature may fail in cases that human sits down and his height is just a half of the standing pose. Or poses with two raising arms may increase the width of the bounding box and break the assumption. For the car, that is the perspective of the camera. If the camera view is in the same line of a moving car, it will show the longer dimension in height not width. But many of the outdoor cases, the assumption are true.

AR feature is chosen to associate with main feature – contour because it is fast and its nature is completely different so that the union of failed case set of them is smaller. When the AR fails by the human pose or car's perspective, the contour fills in the gap with the training of many poses and perspective of object. And if the contour is blurred because of surrounding color, or bad contour detection algorithm, AR does the job.

Zernike NSCT value is a value that represents the contour property of objects and used to differentiate between a human and a car. It is calculated as the amplitude of Zernike moment on contour binary image of the bounding boxed object as followings: first, the bounding boxed object image is contour detected by applying NSCT [12] decomposition on it. The n levels parameter is [0, 1, 3]. This parameter is has 3 numbers, means using 3 pyramidal levels (from coarser to finer scale). The first number – zero means at level 1 of pyramid, the level of directional filter bank decomposition will be 2 exponent zero to 1. It is the contour image received which is synthesized from two next levels. The second level of pyramid is 2 exponents 1 to 2. It's $\pi/2^1$ so we may see it like the two images with horizontal and vertical ways of energy. Similarly, third level is $\pi/2^3$, which is 8 images with rotation of energy. The synthesis of from third through second and to first level gives us contour image with energy at all ways keeping. The number of levels and direction number at each level is chosen to be computationally efficient and good enough to reflex the contour of object. In our experiment, NSCT is quite slow, but three levels are also enough to do the job. Not all contour images are perfectly detected but it is good enough for the classification result. Detailed implementation of NSCT can be found in [12].

Second, the contour image is converted to binary image based on threshold which is just simply the mean of the pixel values in image. This is just a preparatory step for Zernike, which is done due to [19] that the binary image with just contour point is faster and more accurate than the original image for Zernike.

Third, the binary image is passed through Zernike moment to get amplitude [20]. Zernike moment is rotationally invariant, so it is suitable for characterizing contour of object, which may be changed because of the various activities and this property reduces effort in training many poses which are just the rotations of another. Detailed implementation of Zernike and its amplitude can be found in [12].

3.3 Object Classification

Classifier of the proposal uses Gentle Boost (or Gentle AdaBoost) algorithm [21]. First, the classifier is trained with labeled-by-human-eyes data. Then, it will be used to classify based on object's feature vector.

The training stage is as follows: N objects in training data give us N feature vectors which are merged into a matrix called X with N rows and two columns (each vector has two numbers as in Sect. 3.2). N class values of these objects which are classified by human-eyes are merged into a matrix called Y with N rows and one column. Gentle Boost implementation in Matlab [22] will take X and Y matrixes as parameters to initialize and train the classifier. Three more parameters required by the classifier are (i) method, which is filled by "GentleBoost", (ii) nlearn, which is 2, and (iii) learners which is "Tree".

Boost algorithm family is used because of better performance than SVM [23], and the ability to associate two weak classifiers (two features in Sect. 3.2) to a better classifier that show better accuracy for recognizing objects [4]. The defect of Boost is sensitive to noise in training data, but there is no problem with our labeled-by-human-eye training data. Gentle Boost is chosen from many Boost candidates because of its better performance than Real AdaBoost and LogitBoost [21] and *gentle* property. By using weighted least-squares regression to update, Gentle Boost does not cause large update like Real AdaBoost and LogitBoost when a weak learner shows a perfect classification.

Learners parameter is an array of weak learner templates. The proposal chooses only one template for better performance (the less templates, the less work) and the "Tree" template because in [22], two other options are (i) KNN, which is just for Subspace ensemble and (ii) Discriminant, which requires an assumption about Gaussian distributions to ensure its accuracy that may be not the case for our data.

Nlearn parameter is the number of times that each weak leaner will be trained for every template in Learners. The value of 2 means that two weak learners corresponding to two feature values. By experiment, it also shows high accuracy and speed.

The classification stage is just simple that the feature vector is passed to the predicted method of classifier to get the class value, 0 for human and 1 for car in the proposal. Gentle Boost just aggregates results from weak learners for each class as scores. The class with higher score is returned.

4 Experiments and Results

Experiments are developed in Matlab 2013a and carried out on computer of Intel i7 4700MQ 2.4 GHz CPU and 16 GB DDR3 memory. The proposal focuses on the outdoor scenes so PETS 2001 [24] is chosen. Videos are at the resolution of 768×576 and the rate of 25 frames per second. Dataset 1, 2 and 3 are used which include testing and training videos in different folders. Training data comprise 4710 human and 2714 car objects. Testing data contain 5830 human and 2093 car objects. Some test scene images and sample bounding boxed images of human and car are shown in Figs. 2 and 3.

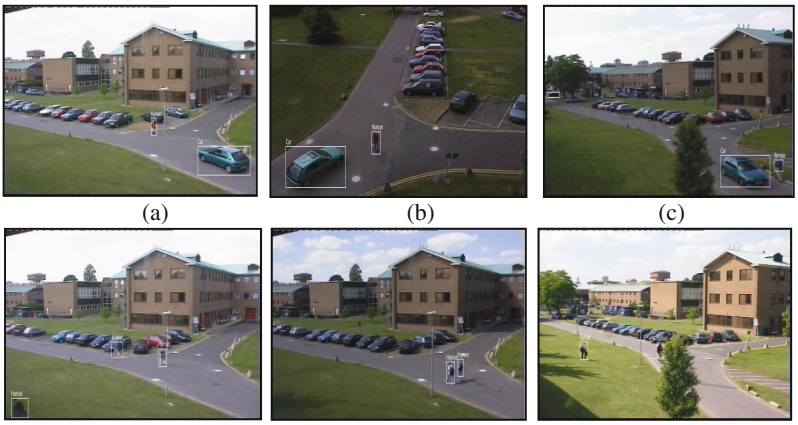

Fig. 2. Scenes in test dataset. (a) and (b) are scenes in test dataset 1. (c) and (d) are scenes in test dataset 2. (e) and (f) are scenes in test dataset 3.

Fig. 3. Sample images. (a) A car in training data. (b) A human in training data. (c) A car in test data. (d) A human in test data.

For the accuracy assessment, the overall result is affected a lot by moving detection stage and compared with human eyes' result for classification. So, missed or wrong (not moving or bounding box bigger than twice of width or height of objects) objects by moving detection stage and objects unrecognizable even by human eyes without the whole frame are ignored. In each frame of videos, each bounding box is counted as an instance of its class. The proposal uses the contour of object as feature, so the object with changing poses will need to be reclassified. The speed is per object and calculated for feature extracting and classification stage only (single threaded), so if each frame has more than one object, the speed per frame will be slower. Certainly, the size of each bounding boxed object image also affects this speed. Table 1 is the experimental results.

The table shows the classification result for test data. There are 5685 correct and 145 incorrect (which are classified as car) results for total 5830 human images and for 2093 car images, they are 1981 correct and 112 incorrect (which are classified as human) images. The overall accuracy of both human and car classification is 96.8 % at 2.0 object per second speed. The proposal fails when objects appear with just a part of them because of stepping in or out of the scene, occlusion by something or bad

Table 1. The overall experimental result.

	Human	Car	Accuracy per class
Human (5,830)	5,685	145	97.5 %
Car (2,093)	112	1,981	94.6 %
Average speed	2.0 object/s	Overall accuracy	96.8 %

Table 2. The overall accuracy comparison on PETS 2001 (dataset 1 to 3).

	The proposal	Somasundaram [25]
Overall accuracy	96.8 %	95.7 %

bounding box (redundant or not enough). But this result is quite good with just a simple association of two features. The limitation on speed of NSCT prevents the experiment from more dataset but it is enough to show the promising of the proposal when compared to the proposal of Somasundaram et al. [25]. Table 2 shows the comparison.

Somasundaram et al. [25] uses area, velocity, DHOG and DCOV features which are combined by a Naive Bayes classifier. These features are also simple and show high accuracy when combined together. DHOG changes HOG to reflex rigid and non rigid motion between vehicles and humans. DCOV uses color, first order and second order gradients to differentiate human from vehicles. Somasundaram's approach bases much on object's appearance; this is a drawback because traffic videos are usually recorded from an average or far view. This results in low resolution of moving object images and less details of moving arms or legs or clothes colors. The proposal bases on contour which is less dependent on resolution than details inside the contour of the object. The more far view, the less details but the whole contour is usually still clear. That's why the proposal can reach a higher accuracy.

5 Conclusion

Object classification can be applied in many purposes such as security surveillance systems for airports, train stations, schools, or buildings of government etc. This task is not easy because they depend on context and environment. This paper proposes the algorithm to classify objects mainly based on their contour property which are represented by the amplitude of Zernike moment on NSCT binary contour image. This feature shows promising results by just a simple association with aspect ratio but gives high accuracy 96.8 %. The speed of NSCT is a bottle neck point for the whole algorithm which needs to be improved in the future to test on more data. And more object classes like scooter, bus, van or group of people, cars, etc. can be added to the experiment to exploit the effectiveness of the main feature in characterizing objects.

Acknowledgments. This research is funded by Ho Chi Minh City University of Technology, VNU-HCM under grant number TSĐH-2015-KHMT-07

References

1. Setitra, I.: Object classification in videos - an overview. J. Autom. Control Eng. **1**(1), 106–109 (2013)
2. Karasulu, B., Korukoglu, S.: Moving object detection and tracking by using annealed background subtraction method in videos: performance optimization. J. Expert Syst. Appl. **39**(1), 33–43 (2012)
3. Lin, D.T., Chen, Y.T.: Pedestrian and vehicle classification surveillance system for street-crossing safety. In: The 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition, pp. 564–570 (2011)
4. Elhoseiny, M., Bakry, A., Elgammal, A.: MultiClass object classification in video surveillance systems - experimental study. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 788–793 (2013)
5. Wang, R., Bunyak, F., Seetharaman, G., Palaniappan, K.: Static and moving object detection using flux tensor with split gaussian models. In: Proceedings of IEEE Workshop on Change Detection, pp. 420–424 (2014)
6. Vishnyakov, B.V., Malin, I.K., Vizilter, Y.V., Huang, S.C., Kuo, S.Y.: Fast human/car classification methods in the computer vision tasks. In: Proceedings of SPIE – The International Society for Optics and Photonics (2013)
7. Graps, A.: An introduction to wavelets. IEEE Comput. Sci. Eng. **2**(2), 50–61 (1995)
8. Ma, J., Plonka, G.: The curvelet transform a review of recent applications. IEEE Sig. Process. Mag. **27**(2), 118–133 (2010)
9. Candès, E.J., Donoho, D.L.: Ridgelets: a key to higher- dimensional intermittency? Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **357**(1760), 2495–2509 (1999)
10. Candès, E.J., Donoho, D.L.: Curvelets - a surprisingly effective nonadaptive representation for objects with edges. In: Cohen, A., Rabut, C., Larry, L. (eds.) Curves and Surface Fitting: Saint-Malo 1999, pp. 105–120. Vanderbilt University Press, Nashville (2000)
11. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2106 (2005)
12. Cunha, L.D., Zhou, J., Do, M.N.: The nonsubsampled contourlet transform: theory, design, and applications. IEEE Trans. Image Process. **15**(10), 3089–3101 (2006)
13. Prokop, R.J., Reeves, A.P.: A Survey of moment-based techniques for unoccluded object representation and recognition. CVGIP. Graph. Models Image Process. **54**(5), 438–460 (1992)
14. Bunyak, F., Palaniappan, K., Nath, S.K.: Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking. J. Multimedia **2**(4), 20–33 (2007)
15. Palaniappan, K., Ersoy, I., Seetharaman, G., Davis, S.R., Kumar, P., Rao, R.M., Linderman, R.: Parallel flux tensor analysis for efficient moving object detection. In: Proceedings of the 14th International Conference on Information Fusion, pp. 1–8 (2011)
16. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. **2**, 246–252 (1999)
17. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: two new techniques for image matching. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence, pp. 659–663 (1977)
18. http://www.mathworks.com/help/vision/ref/vision.blobanalysis-class.html. Accessed 1 August 2015
19. Mukundan, R.: A contour integration method for the computation of zernike moments of a binary image. In: USM-Penang: National Conference on Research and Development in Computer Science and Applications – REDECS 1997, pp. 188–192 (1997)

20. Tahmasbi, A., Saki, F., Shokouhi, S.B.: Classification of benign and malignant masses based on zernike moments. Int. J. Comput. Biol. Med. **41**(8), 726–735 (2011)
21. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. J. Ann. Stat. **28**(2), 337–407 (2000)
22. http://www.mathworks.com/help/stats/ensemble-methods.html. Accessed 1 August 2015
23. Cortes, C., Vapnik, V.: Support-vector networks. J. Mach. Learn. **20**(3), 273–297 (1995)
24. ftp.cs.rdg.ac.uk/pub/PETS2001. Accessed 1 August 2015
25. Somasundaram, G., Morellas, V., Papanikolopoulos, N., Bedros, S.: Object classification in traffic scenes using multiple spatio-temporal features. In: The 2012 20th Mediterranean Conference on Control and Automation (MED), pp. 1536–1541 (2012)