People Counting in Conference Scenes in Shearlet Domain

Nguyen Thanh Binh(✉)

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

Abstract. People counting is an important task in visual-based surveillance system. The task of people counting is not easy to solve problems. In this paper, the author has proposed a method for people counting which identify the objects present in a scene of conference into two classes: empty seat and non-empty seat. The proposed method based on saliency map and color smoothing in shearlet domain. The author uses shearlet transform and combine of adaboost with support vector machine for classifiers and people counting. The proposed method is simple but the accuracy of people counting is high.

Keywords: People counting · Shearlet transform · Saliency map · Color smoothing

1 Introduction

Today, computer vision is one of the important senses which helps people to receive information from the real world. Computer vision processing provides the methods and analyses images from the real world similar to the way people perform, to draw information to make the appropriate decision. Detecting the number of people in crowded scenes is an important task in visual-based surveillance system. Smart surveillance systems by image have been developed and proven effective in some specific areas such as human activity monitoring, traffic monitoring, etc. From the images obtained from various observations, the author can detect the movement of objects in the frame and identify the object that is people, vehicles, etc. Many systems have been researched and developed. For example, the problem of traffic monitoring can tell us the number of vehicles circulating through the ramp which is monitored and gives information on the speed of movement, and the path of the object to be tracked. However, the system still encountered some existences as the effectiveness of the observer always depend on the environmental conditions of observation, types of object motion or other objective reasons.

Estimating the number of people moving in crowded scenes used commonly techniques such as: detection the head information [1, 4], expectation maximization [2], low-level features and Bayesian regression [3], HOG features [5] and background subtraction [6]. Algorithm based on background subtraction utilizes the current image to compare it with the background image and detect the moving scene. Most of methods of

P.C. Vinh et al. (Eds.): ICTCC 2016, LNICST 168, pp. 363–372, 2016.
DOI: 10.1007/978-3-319-46909-6_33

background subtraction are median filter, mean filter, temporal median filter, Kalman filter, sequential kernel density approximation and eigen backgrounds. It is hard to model the background when the environment is complex. The optical flow method was used to solve this problem. However, the drawback of the optical flow is that it has high computational complexity and it is sensitive to noise.

Teixeira [7] used custom-built camera installed on the ceiling for localizing and counting people in indoor spaces. Chan [8] mapped feature statistics extracted from the blob to the count of people. Wang [9] built a spatio-temporal group context model to model the spatio-temporal relationships between groups to people counting. Zhang [10] proposed group tracking to compensate the weakness of multiple human segmentation which completes occlusion. Jun [11] designed a block-updating way to update the background model and used an improved k-means clustering for locating the position of each person. Wu [12] proposed to learn by boosting edgelet feature based weak classifiers for body part detectors. Kong [13] proposed a viewpoint invariant learning–based on the method from a single camera for people counting in crowds. However, most of these methods are complex as the object is occluded.

In this paper, the author proposes a method to implement for estimating the number of people in conference scenes based on people features in shearlet domain. The author uses shearlet transform and apply the combine of adaboost with support vector machine for classifiers to estimate the people counting. The proposed method was tested on the dataset which is picked up in conference scene. The rest of this paper is organized as follows: in Sect. 2, the author described the basic of shearlet transform, feature and its advantages for people counting. Also details of the propose method for people counting are presented in Sect. 3. In Sect. 4, results of the proposed method are given and conclusions are made in Sect. 5.

2 Shearlet Domain and Features Selection

2.1 Shearlet Transform

Shearlet is similar to curvelet in that both perform a multi-scale and multi-directional analysis. There are two different types of shearlet systems: band-limited shearlet systems and compactly supported shearlet systems [14]. The band-limited shearlet transform have higher computational complexity in frequency domain.

The digitization of discrete shearlet transform performed in the frequency domain. The discrete shearlet transform is the form [15]:

$$f \mapsto \langle f, \psi_n \rangle = \left\langle \hat{f}, \hat{\psi}_n \right\rangle = \left\langle \hat{f}, 2^{-j\frac{3}{2}} \hat{\psi}\left(s_k^T A_{4^{-j}}\right) e^{2\Pi i < A_{4^{-j}} s_k m, \ldots >} \right\rangle \tag{1}$$

where n = (j, k, m, i) indexes scale j, orientation k, position m, and cone i.

Shearlets perform a multiscale and multidirectional analysis. For images f(x) are C^2 everywhere, where f(x) is piecewise C^2, the approximation error of a reconstruction with the N-largest coefficients $\left(f_N(x)\right)$ in the shearlet expansion is given by [16]:

$$\left\|f - f_N\right\|_2^2 \leq B.N^{-2}(\log N)^3, \quad N \to \infty \tag{2}$$

The author has chosen shearlet transform because it not only has high directionality but also represents salient features (edges, curves and contours) of image in a better way compared with wavelet transform. Shearlet transform is useful for people counting due to its following properties [17]:

(i) Frame property: It is helpful to a stable reconstruction of an image.
(ii) Localization: Each of shearlet frame elements needs to be localized in both the space and the frequency domain.
(iii) Efficient implementation.
(iv) Sparse approximation: to provide sparse approximation comparable to the band-limited shearlets.

The shearlet transform will produce a highly redundant decomposition when implemented in an undecimated form [18]. Like the curvelet transform, the most essential information in the image is compressed into a few relatively large coefficients, which coincides with the area of major spatial activity in shearlet domain. On the other hand, noise is spread over all coefficients and at a typical noise level the important coefficients can be well recognized [19]. Thus setting the small coefficients to zero will not affect the major spatial activity of the image.

2.2 The Combine of Adaboost with Support Vector Machine for Classifiers

For a given feature set and a training set of positive and negative images, adaboost can be used in both of them to select a small set of features and to train the classifier. Viola [20] firstly used binary adaboost for their face detection system. Boosting is a method to improve the performance of any learning algorithm, generally consisting of sequential learning classifier [21]. Adaboost itself trains an ensemble to weak-learners to form a strong classifier which perform at least as well as an individual weak learner [22]. Adaboost ensembles a particular feature, where each feature represents observable quantity associated with target. In this proposed work, the author has used adaboost algorithm which is described by Viola [20].

Support vector machines (SVM) include associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis in machine learning. SVM can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces.

The idea of the combine of AdaBoost and SVM is that for the sequence of trained RBFSVM (SVM with the RBF kernel) component classifiers. The author starts with large s values. The s values reduce progressively as the Boosting iteration proceeds.

The steps of the combine of AdaBoost with Support Vector Machine for classifiers as below [23]:

(i) Consider a set of training samples with $\{(x_1,y_1), \ldots, (x_n,y_n)\}$. The initial value of σ is set to σ_{ini}; the minimal σ is set to σ_{min} and each step is set to σ_{step}.

(ii) The weights of training samples as:

$$w_i^1 = 1/N \text{ for all } i = 1, \ldots, n \tag{3}$$

(iii) While $(\sigma > \sigma_{min})$, the author trains a RBFSVM component classifiers, h_t, on the weighted training set. The training error of h_t calculate as: $\varepsilon_t = \sum_{i=1}^{N} wi_i^t$ and $y_i \neq h_t(x_i)$

(iv) If $\varepsilon_t < 0.5$, decrease σ value by σ step and goto (iii).
Set the weight of component classifier h_t as

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{4}$$

(v) Update the weights of training samples:

$$w_i^{t+1} = \frac{w_i^t exp\{-\alpha_t y_i h_t(x_i)\}}{C_t} \tag{5}$$

where C_t is a normalization constant, and

$$\sum_{i-1}^{N} w_i^{t+1} = 1 \tag{6}$$

(vi) This process continues until σ decrease to the given minimal value. The output of classifier is [24]:

$$f(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{7}$$

2.3 Feature Selection

Among most classification problems, it is not easy to learn good classifiers before removing these unwanted features due to the huge size of the data. The author can reduce the running time of the learning algorithms and a more general classifier by reducing the number of irrelevant features.

A general feature selection for classification is presented as Fig. 1:

In Fig. 1, the step of feature selection affects the training phase of classification. The features selection for classification will select a subset of features. The process data with the selected features will be sent to the learning algorithm. Any object classification algorithm is commonly divided into three important components:

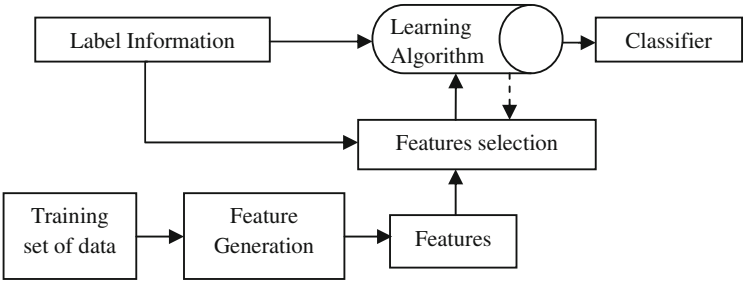

Fig. 1. Framework of feature selection for classification

extraction of features, selection of features and classification. Therefore, feature extraction and selection play an important role in object classification.

3 The Method for People Counting in Conference Scenes

In this section, the author describes a method for people counting in conference scenes in shearlet domain. People counting is hard work. In the past, there are many methods for this work. Every method has particular strengths and drawbacks depending on the scenes. The proposed method uses shearlet transform for feature evaluation and the combine of adaboost with support vector machine for identification. For experimentation, the author has considered two classes: empty seat and non-empty seat class. The empty seat class contains only objects of which seat and non-empty seat class contain people. The overall of the proposed method for object detection is described as Fig. 2.

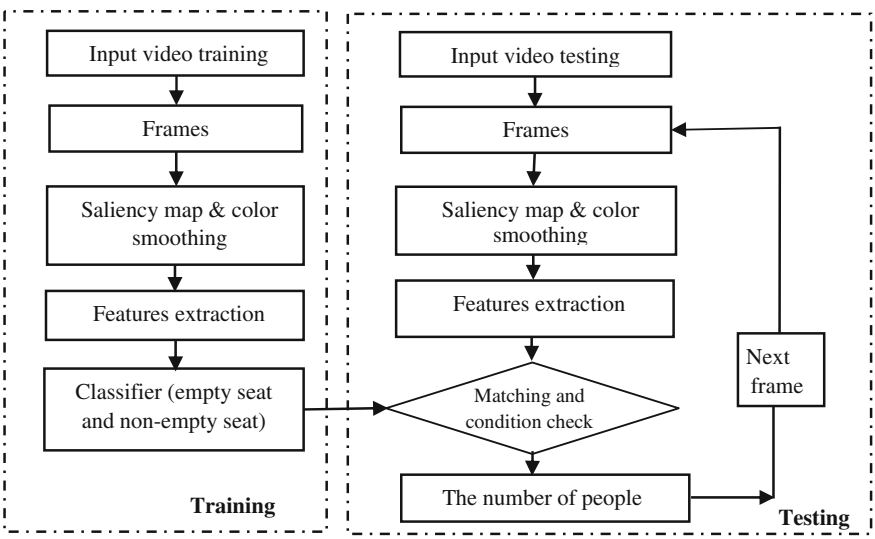

Fig. 2. The overall of the proposed method for people count.

In Fig. 2, a video sequence contains a series of frames. Each frame can be considered as an image. The proposed method includes two periods: Training and Testing. In the training periods, there are three steps:

Firstly, saliency map and color smoothing. In this step, the author uses histogram base contrast to generate the salient map. The salient map has noise problem. Therefore, the author uses color smoothing method to remove them. The histogram is used to find the saliency value for each color in an image based on the contrast method. The author uses the idea of linearly-varying smoothing weight to create the saliency map from the saliency value of each color. The saliency value of each color becomes more similar to neighbor colors and it helps grouping colors. The author also creates a binary mask to extract the object and binary the saliency maps. Defining threshold value k as average value of saliency values in the saliency map. If (the saliency value $>$ k) then assign value $= 255$ and 0 otherwise.

Secondly, feature extraction and create map. The author measures a threshold value from salient map to create a mask. In here, the author uses shearlet filter for computing searched area and detecting objects. These objects are saved as the blob. The feature regression-based method will be used to describe the relationship between the low-level features together. The author defines it as: the area is total number of pixels in the blob. The perimeter is the total number of pixels on the perimeter of the blob and the total edge pixels is the total number of edge pixels in the blob.

Thirdly, the author combines adaboost with support vector machine for identification. After feature extraction for positive and negative datasets, the author will train using the combine of adaboost with support vector machine for classifiers as presented in Subsect. 2.2. The author collects sample images for training and testing the classifier. The author has collected images for two classes: empty seat and non-empty seat from conference scenes. The author has assigned value '1' for non-empty seat data and value '-1' for empty seat object data.

In the testing period, there are three steps: the step 1 and step 2 are similar in the training period. The third step is matching and also a condition check. In this step, the author matches the result in step 3 of training period with the result in step 2 of testing period. If the results are true, the author has to count the number of people and go to next frames. This processing runs to the final frame.

4 Experimental Results

In this section, the author makes experiments to people counting in theatre scenes. Hard thresholding is applied to the coefficients after decomposition in shearlet domain. The author applies the same approach to the shearlet transform. For performance evaluation, this method has been done on many videos in the large video dataset. Here, the author reports the results on some video clips.

The proposed people counting technique has been tested on my own dataset created by the author this paper. The dataset consists of two classes: empty seat and non-empty seat. Empty seat class contain images of different types of seat whereas non-empty seat class contain images of other objects such as seat contain human. The proposed method

has been tested on this dataset. Some example images of both classes have been shown in Fig. 3. Some example images of conference scenes have been shown in Fig. 4.

(a). Empty seats

(b). Non-empty seats

Fig. 3. Sample images of empty seat and non-empty seats objects of my own dataset

Fig. 4. Sample conference scenes in my own dataset

My experiments are scene video clips with the frame size 288 by 352. The proposed method processes this video clips at 24 frames/second. In here, the author defines as:

The different performance metrics, such as Average Classification Accuracy (ACA), True Positive Rate (TPR) (Recall) and Predicted Positive Rate (PPR) (Precision), are depended on four values: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), where [25, 26]:

+ TP is the number of images, which are originally positive images and classified as positive images.

+ TN is the number of images, which are originally negative images and classified as negative images.

+ FP is the number of images, which are originally negative images and classified as positive images.

+ FN is the number of images, which are originally positive images and classified as negative images.

All above three performance metrics are defined in [25, 26]. In here, the author reviews parameters following:

+ ACA is defined as the proportion of the total number of prediction that was correct:

$$\text{ACA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

+ TPR is defined as the proportion of positive cases that were correctly classified as positive:

$$\text{TPR (Recall)} = \frac{\text{TP}}{\text{FP} + \text{FN}} \tag{9}$$

+ PPR is defined as the proportion of the predicted positive cases that were correct:

$$\text{PPR (Precision)} = \frac{\text{TP}}{\text{FP} + \text{TP}} \tag{10}$$

The accuracy of the proposed method is shown in Table 1.

Table 1. Values of performance measures of the proposed method

Video test	Time (second)	TPR (Recall) (%)	TNR (%)	FPR (%)	FNR (%)	PPR (Precision) (%)	Average accuracy (%)
Conference room 1	250	98	96	4	2	96.08	97
Conference room 2	300	98	98	2	2	98	98
Conference room 3	350	99	94	1	1	99	99
Conference room 4	400	100	100	0	0	100	100
Conference room 5	450	98	97	3	2	97	97.5

From Table 1, one can observe that the proposed method gives better performance results.

Besides, my experiments are also scene video clips with the frame size 288 by 352. The proposed method processes this video clips at 24 frames/second. In here, the author defines it as:

+ Real number presents the number of people who's sitting in conference room to scenarios.

+ Counting Number presents the number of people which the system counted.

The accuracy of the proposed method is shown in Table 2.

Table 2. The accuracy of people counting

Video test	Time (second)	Real people number in scenarios	People counting number in scenarios which the system counted	Accuracy (%)
Conference room 6	400	121	120	99.1
Conference room 7	450	59	58	98.3
Conference room 8	500	39	39	100.0
Conference room 9	550	102	101	99.0
Conference room 10	600	52	51	98.0

From Table 2, the proposed method also gives better performance results.

5 Conclusions and Future Works

People counting is an important task in visual-based surveillance system. In conference scenes, the objects are usually occlusion, blurring and noising because of low light, light changing, many color light, etc. The task of people counting is not easy to solve problems. In this paper, the author proposes a method to implement for people counting in conference scenes based on shearlet domain. The author uses shearlet transform and combine of adaboost with support vector machine for classification. The accuracy of people counting is high. However, if the conference scene is not clear the proposed method will be affected. The step of the proposed method is salient map and color smoothing. As mentioned above, the saliency value of each color becomes more similar to neighbor colors and it helps grouping colors. In the future work, the author will improve the color smoothing steps to reduce the impact of light change.

Acknowledgment. This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2015-20-08.

References

1. Cai, Z., Yu, Z.L., Liu, H., Zhang, K.: Counting people in crowded scenes by video analyzing. In: IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1841–1845 (2014)
2. Hou, Y.-L., Pang, G.K.: People counting and human detection in a challenging situation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **41**(1), 24–33 (2011)
3. Chan, A.B., Vasconcelos, N.: Counting people with low-level features and Bayesian regression. IEEE Trans. Image Process. **21**, 2160–2177 (2012)
4. Xu, H., Lv, P., Meng, L.: A people counting system based on head-shoulder detection and tracking in surveillance video. In: IEEE International Conference on Computer Design and Applications (ICCDA), vol. 1, pp. V1–V394 (2010)
5. Chen, L., Wu, H., Zhao, S., Gu, J.: Head-shoulder detection using joint hog features for people counting and video surveillance in library. In: IEEE Workshop on Electronics, Computer and Applications, pp. 429–432 (2014)
6. Bala Subburaman, V., Descamps, A., Carincotte, C.: Counting people in the crowd using a generic head detector. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2012)

7. Teixeira, T., Savvides, A.: Lightweight people counting and localizing in indoor spaces using camera sensor nodes. In: First ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2007, pp. 36–43 (2007)
8. Chan, A.B., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2008)
9. Wang, J., Fu, W., Liu, J., Lu, H.: Spatio-temporal group context for pedestrian counting. IEEE Trans. Circuits Syst. Video Technol., 1–11 (2014)
10. Zhang, E., Chen, F.: A fast and robust people counting method in video surveillance. In: International Conference on Computational Intelligence and Security, pp. 339–343. IEEE (2007)
11. Luo, J., Wang, J., Xu, H., Lu, H.: A real-time people counting approach in indoor environment. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015. LNCS, vol. 8935, pp. 214–223. Springer, Heidelberg (2015). doi:10.1007/978-3-319-14445-0_19
12. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int. J. Comput. Vis. **75**(2), 247–266 (2007)
13. Kong, D., Gray, D., Hat, T.: A viewpoint invariant approach for crowd counting. In: Proceedings of the International Conference on Pattern Recognition, pp. 1187–1190 (2006)
14. Kutyniok, G., Labate, D.: Shearlets: Multiscale Analysis for Multivariate Data. Applied and Numerical Harmonic Analysis. Birkhauser (2012)
15. Lim, W.-Q.: The discrete shearlet transform: a new directional transform and compactly supported shearlet frames. IEEE Trans. Image Process. **19**(5), 1166–1180 (2010)
16. Guo, K., Labate, D.: Optimally sparse multidimensional representation using shearlets. SIAM J. Math. Anal. **39**, 298–318 (2007)
17. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
18. Patel, V.M., Easley, G.R., Healy Jr., D.M.: Shearlet-based deconvolution. IEEE Trans. Image Process. **18**(12), 2673–2685 (2009)
19. Thanh Binh, N., Khare, A.: Object tracking of video sequences in curvelet domain. Int. J. Image Graph. **11**(1), 1–20 (2011)
20. Zivkovic, Z., Krose, B.: An EM-like algorithm for color histogram-based object tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, USA, vol. 1, pp. 798–803 (2004)
21. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**(4) (2006)
22. Wang, D.: Unsupervised video segmentation based on watersheds and temporal tracking. IEEE Trans. Circuits Syst. Video Technol. **8**(5), 539–546 (1998)
23. Li, X., Wang, L., Sung, E.: AdaBoost with SVM-based component classifiers. Eng. Appl. Artif. Intell. **21**, 785–795 (2008)
24. Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M.: Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE Trans. Med. Imaging **29**(1), 30–43 (2010)
25. Khare, M., Thanh Binh, N., Srivastava, R.K.: Dual tree complex wavelet transform based human object classification using support vector machine. J. Sci. Technol. **51**(4B), 134–142 (2013)
26. Thanh Binh, N.: Object classification based on contourlet transform in outdoor environment. Nat. Comput. Commun. **144**, 341–349 (2015)