# A Novel Hybrid Approach to Detect High Density Crowd Regions in Still Image Scenes

Bo LI*, Jiangqing WANG, Su ZHANG, Kai LIANG, Zece CHEN, and Li CHENG

College of Computer Science, South-Central University for Nationalities, Wuhan, Hubei, 430074, China
*Corresponding author, libo_hust@126.com

*Abstract* — **Crowd scene detection and crowd region location are important for public safety monitoring, because crowds often lead to emergency situations. A challenge is to judge high density crowd scenes which contain complex background and some non-crowd regions, such as trees, buildings, etc. This paper presents a novel detection approach for high density crowd scenes and crowd regions in still images. First, two important features which include Local Binary Pattern (LBP) and GIST are chosen and fused, whose principal components are extracted in advance before constructing a new feature vector. Then, crowd scenes are detected by Support Vector Machines (SVM) based on the fused features. Further, considering the orientation distribution difference between crowd and non-crowd regions, we use Gabor filter to enhance crowd regions, and remove non-crowd regions from detailed crowd regions. Experimental results demonstrate that the proposed detection approach is feasible and effective in variable conditions.**

***Keywords - crowd scene detection; LBP; GIST; Gabor filter; still image***

## I. INTRODUCTION

Nowadays, crowd phenomenon is very common with the increasing of human population, especially in public areas, such as commercial buildings, markets, and stadiums, etc. In recent years there are often some emergencies and incidents due to the high density gathering in crowd scenes. Therefore, it is important to enhance the safety monitoring and many governments invest plenty of funds to monitor crowd scenes by means of video surveillance. More and more researchers analyze the crowd scenes by computer vision and pattern recognition techniques.

In recent years, with the popularization of the smart phones, still images which are often emerging in many fields (e.g., social network) are gradually an important data source besides videos. However, video surveillance and computer vision solutions are not completely applicable to still images which are prevalent and can be found in all kinds of social network, such as Facebook, Friendster, etc. For this reason, this paper works on how to judge and recognize the crowd scenes and estimate the crowd density effectively based on still images. In fact, the crowd scene detection by still images is difficult because there is no relational information between neighboring frames unlike videos.

### A. *Related Work*

Crowd scenes detection and crowd density estimation have been paid plenty of attention in recent years. There are really strong relationships between the two issues. Recently there are many documents published in the field of crowd density estimation. Generally, the approaches of crowd density estimation can be divided into two classes, one is human detection based, and the other is feature analysis based. The first approaches count the number of people by means of segmenting and recognizing single person. For instance, head detector [1]. However, it is difficult to count the result because of complex background and occlusion conditions, especially in high dense crowd scenes. Gall proposed a Hough forest framework to recognize the action and track the people, which is robust to partial occlusions to some extent [2]. The second approaches analyze the crowd scenes by extracting several local and holistic features of crowd scenes. For instance, Hussain et al. presented a pixel-based crowd density estimation approach [3], however, high occlusion often leads to missed and false detection cases. Lots of researchers have paid high attention to texture based methods, and many texture features can be used to analyze and estimate crowd density, such as, Grey Level Dependence Matrices[4], Gabor filter[5], Local binary patterns[6][7], corner point based methods[8][9].

In terms of proposed literatures, most of the existing approaches analyze and estimate the people number based on low and medium density crowd scenes, but they are not suitable to high density crowd cases due to severe occlusion. In addition, some documents really consider the extreme crowd scenes, however their algorithms require video frames as input data and not appropriate for still images [5,10-13].

Some algorithms proposed estimate crowd density by establishing association between local features and estimation results, in still images of high density even extremely dense crowd scenes[14][15]. However, there are really many differences between crowd scenes detection and crowd density estimation.

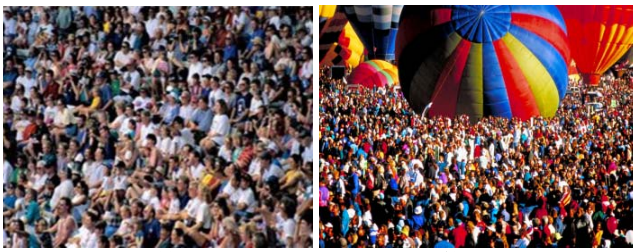


Figure 1. Crowd Scene Images (a) does not contain non-crowd regions, but b) contains non-crowd regions).

Another problem is how to detect the crowd regions in the crowd scene images. It's necessary to remove the non-crowd areas which often disturb the precision of estimation and analysis for entire still image. As Figure 1, image a) is complete full with crowd, but image b) contains some non-crowd regions in the upper region.

### B. *Overview of Our Approach and Contributions*

For high density crowd case, especially extremely dense crowd scene images, many exiting techniques such as, directly detecting heads, are not reliable to provide good count results because of occlusion. This paper proposes a detection algorithm for high dense crowd scenes in still images. The framework of our approach is shown as Figure 2.

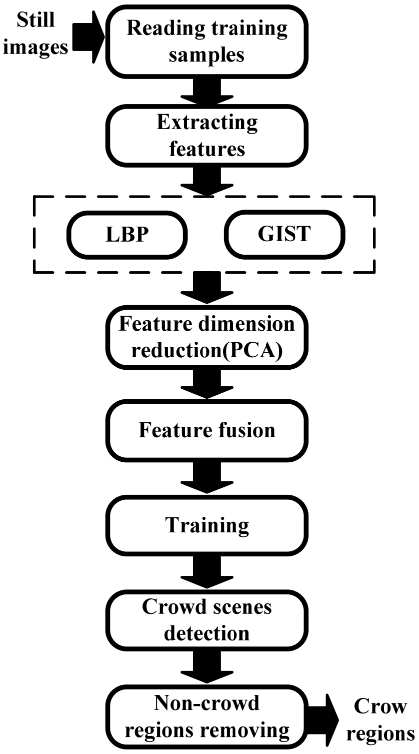


Figure 2. Flow Chart of Our Approach

Our contributions are presented in two folds.

1) A contribution in this paper is the construction of a novel feature based on local feature and holistic feature. Aim to all kinds of scenes, including crowd scenes, this paper can judge crowd scenes, and classify them based on their human population density in the still images using the new feature vector. We choose LBP and GIST because of their good describing ability for image texture.

2) Another contribution in this paper is that removing the non-crowd regions in entire still images using Gabor filter and locating the crowd regions. It is difficult to judge high dense crowd scene that may include some non-crowd areas, such as trees, buildings, etc. Many existing methods judge crowd areas by means of dividing the entire image into many blocks and judges the blocks respectively based on classifiers, which may be low effective. Although Ref.[5] adopted Gabor filter in estimation of crowd density, it did not consider removing the non-crowd regions based on the characteristics of high dense crowd regions.

The rest of the paper is organized as follows. Section II presents the method of a new feature construction. Section III introduces crowd scene detection and crowd regions locating. Section IV shows experimental results and demonstrates the accuracy of the proposed approach. Section V includes conclusions and prospects for future research.

## II. FEATURE VECTOR CONSTRUCTION

Color is often not a stable feature for scene detection, for example, different scenes can result in similar color histogram. Scenes can also be rapidly recognized without color feature, so does shape feature. Therefore, this paper omits color and shape feature in our experiments and focuses on texture as main basis for crowd scene detection.

### A. *Local Binary Patterns*

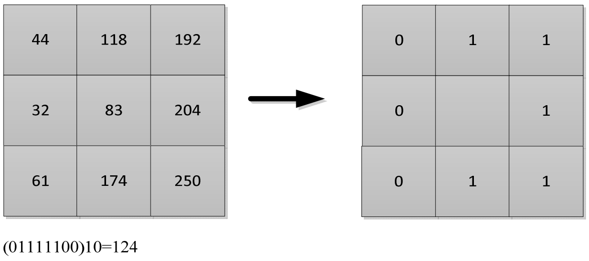
Local binary patterns (LBP) [16] is a kind of texture descriptor used for classification in pattern recognition which was first described in 1994. The LBP feature vector is created in the following steps:

Firstly, divide the entire image region into several cells(for example, 8×8 pixels for each cell). For each pixel in a cell, compare its value to each of its eight neighbors respectively. Then follow the pixels along a circle sequence (i.e. counter-clockwise or clockwise), the result is "0" if the neighbor's value is smaller than the center pixel value, otherwise the result is "1", as Figure 3. After traversing all the neighbors, an 8-digit binary number can be obtained. In addition, the LBP value can calculated by Formula(1) and (2).

$$LBP(x_c, y_c) = \sum_{p=0}^{N-1} 2^p s(g_p - g_c) \qquad (1)$$

$$s(x)=\begin{cases}1, & if\,(x\geq 0)\\ 0, & else\end{cases} \tag{2}$$

Where, $(x_c, y_c)$ is the center pixel in a cell, and N is the number of neighbor pixels. $g_c$ is the pixel value of center pixel, and $g_p$ is pixel value of neighbor pixels.



|  |  |  |
|---|---|---|
| 44 | 118 | 192 |
| 32 | 83 | 204 |
| 61 | 174 | 250 |

|  |  |  |
|---|---|---|
| 0 | 1 | 1 |
| 0 |  | 1 |
| 0 | 1 | 1 |

Figure 3. LBP Feature (left figure is original image pixel value, and right figure is LBP value).

Some improved circular LBP operators whose radius can be changed are shown as Figure 4.

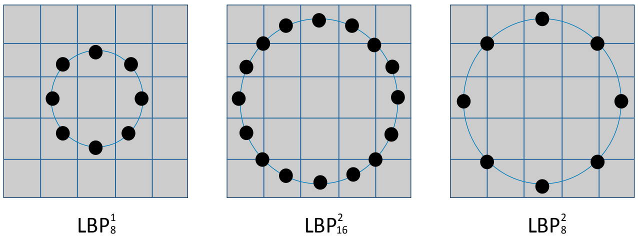


Figure 4. Circular LBP Operators

To obtain rotation invariance, it is feasible to get the minimum value of different patterns after rotation, as Figure 5.

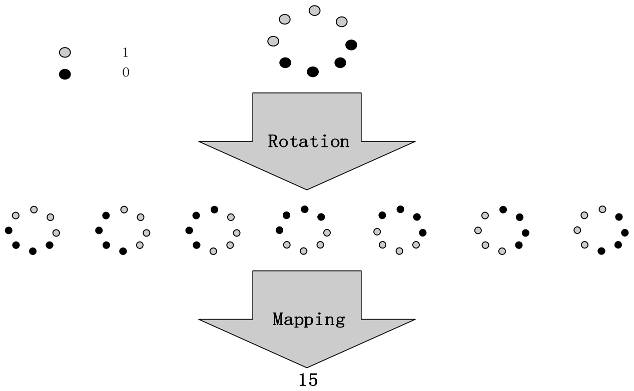


Figure 5. LBP Rotation Invariant Patterns

Then, the histogram is calculated, which counts the frequency of each "number" occurring in the cell. In fact, the histogram can be seen as a feature vector with 256 dimensions. Next, the histogram is normalized.

In the same way, histograms of all cells can be concatenated. Then, a feature vector for the entire image window can be obtained.

### B. *Gist Feature Vector*

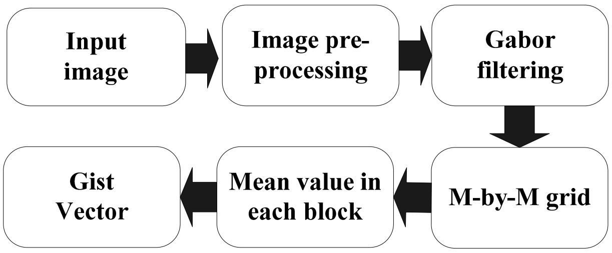


Figure 6. GIST Feature Extraction.

Gist is a global feature proposed in Ref. [17] and [18]. The feature can be computed by following steps. As Figure 6, the image sample is firstly converted into grayscale image, and is then processed by a whitening filter, so the dominant structural details of the image can be well preserved. Then the processed image in the last step is normalized with respect to local contrast. Next, the result is passed through a cascade of Gabor filters [19], which contain A different scales and B different filtering orientations. Each of these A×B images with specific orientation and scale, is then divided into an M-by-M grid. For each block on the grid, the average intensity in the block is computed. Finally, the extracted feature is a concatenated vector of A×B×M×M dimensions.

In the process of extracting GIST feature, Gabor filters are applied to an input image and then extracts a group universal textons. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function, as Formula (3)-(5).

$$g(x,y,\sigma,\theta)=\frac{1}{2\pi\sigma^2}e^{-(\frac{x'^2}{2\sigma^2}+\frac{y'^2}{2\sigma^2})}e^{j\frac{\pi x'}{\sigma}} \tag{3}$$

$$x' = x\cos\theta + y\sin\theta \tag{4}$$

$$y' = -x\sin\theta + y\cos\theta \tag{5}$$

Where, $\sigma$ and $\theta$ describes scale and orientation respectively. A set of Gabor filters with different scales and orientations may be helpful for extracting useful texture features from an image, as Figure 7.

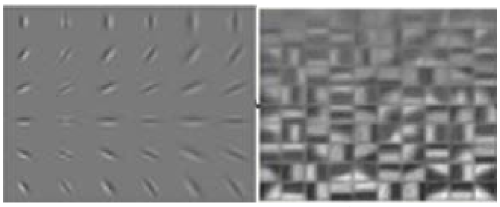


Figure 7. GIST Feature. a) Gabor filter, b)output textons.

### *C. Multi-features Fusion*

LBP are only a type of local feature which can effectively reflect the local texture pattern, however, LBP provides inadequate global spatial and statistical information for entire scene image detection. It is often not enough for LBP to recognize a large area of still scene images. Therefore, it is necessary to adopt global spatial relationships across the scene in texture analysis for scene detection.

GIST is a global feature which is suitable for scene detection and classification, so in this paper, GIST and LBP are combined as a new feature for scene detection.

Because there are plenty of feature dimensions in GIST feature, this paper chooses the 60 dimensions of vector based on uniform sampling. In order to improve computing efficiency, this paper performs principal components analysis for the LBP and GIST respectively in advance. A new feature vector is constructed by fusing the principal components of two features.

Table 1 shows a vector containing the percentage of the total variance explained by each principal component. For LBP, principal component contains 16 dimensions, and for GIST, 15 dimensions.

TABLE 1. PCA OF TWO FEATURES

| Feature | LBP | GIST |
|---|---|---|
| Feature dimensions | 38 | 60 |
| PC dimensions | 16 | 15 |
| PC Contribution(%) | 99.9 | 96.5 |

The new feature vector is constructed as Figure 8. The principal components of LBP and GIST are united as a new vector. The dimension of the new feature vector is the sum of the dimension of two principal components.

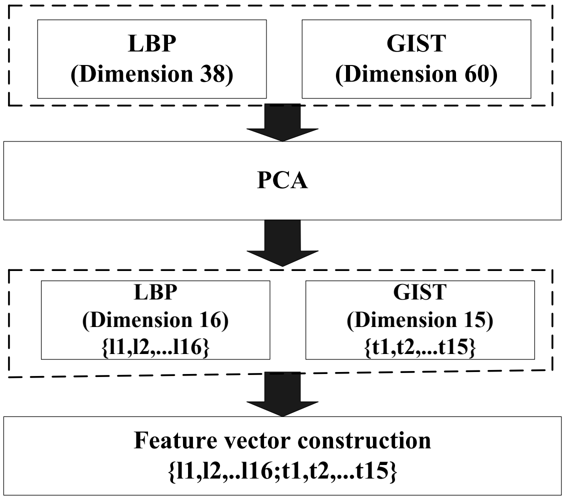


Figure 8. Two-feature Fusion

## III. CROWD SCENE DETECTION AND CROWD REGIONS LOCATING

### *A. Crowd Scene Detection*

After constructing the new feature vectors, they can be used to train classifiers to detect crowd scenes. This paper adopts SVM as the classifier. Because the crowd scene detection is just a two-class problem, SVM is very suitable for solving it.

### *B. Non-crowd Regions Removal*

The crowd scene image can be detected by SVM. Further, it is necessary to locate the crowd areas in the still image. In this paper, Gabor filter is used to perform non-crowd areas removal. In general, because crowd areas are full with plenty of randomly distributed blobs of faces and heads, they should generate a high response in every orientation angle. However, non-crowd regions, such as regular people, trees, and buildings, only have high responses orthogonal to their main orientations. A set of Gabor filters with different scales and orientations are very helpful for extracting the response in every direction from a crowd scene image. Figure 9 shows different scales and orientations of Gabor filters.

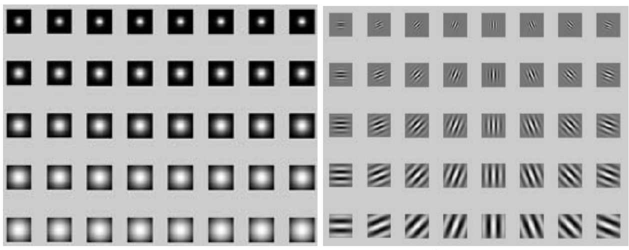


Figure 9. Gabor Filter. a)Magnitudes of Gabor filters, b)Real parts of Gabor filters.

Without loss of generality, a set of Gabor functions $g_{\sigma,\theta}(x,y)$ can be obtained by rotating and scaling $g(x,y)$, where, $g(x,y)$ is the Gabor mother generating function, $\sigma = 0,1,...,S-1$ and $\theta = 0,1,...,K-1$ . S and K are respectively the total number of scales and orientations.

Given an original image $I(x,y)$ , its Gabor filtered images are

$$R_{\sigma,\theta}(x,y) = \sum_{x1}\sum_{y1} I(x_1,y_1)g_{\sigma,\theta}(x-x_1,y-y_1) \quad (6)$$

Next, the magnitudes of the filtered images including all the orientations and scales are summed, as Formula (7).

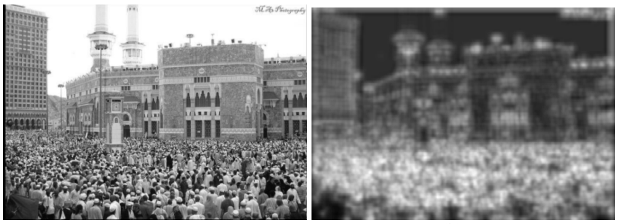
$$M = \sum_{\sigma=0}^{s-1}\sum_{\theta=0}^{k-1} | R_{\sigma,\theta} | \qquad (7)$$

Where, S and K are respectively the total number of scales and orientations. The crowd regions are enhanced to a great extent, after Gabor filtering and the operation of Formula (7) are performed. However, non-crowd regions only have high responses orthogonal to their main orientations, as Figure 10(b).

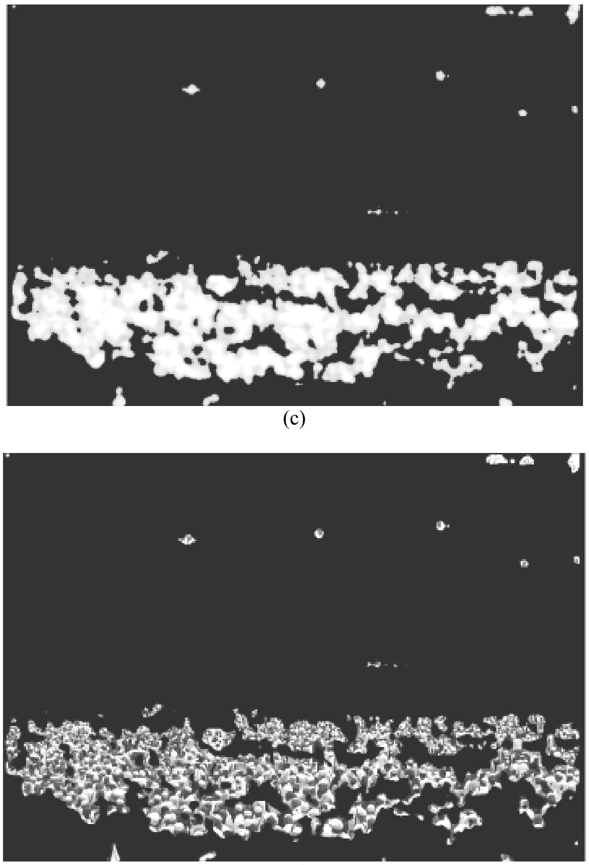
Then, the result in Figure 10(b) is converted into binary image based on a threshold, as Formula(8), and the result is as Figure 10 (c).

$$I = \begin{cases} 0, & if\,(M < threshlod) \\ original\ value, & otherwise \end{cases} \qquad (8)$$

Where, $M$ is the result of Formula (7). The range of crowd region can be mapped from the white areas of Figure 10(c) which is really crowd regions, as Figure 10(d). Therefore, the non-crowd regions are removed.



(a) (b)



(c)

(d)

Figure 10. Non-crowd Regions Removing. ( a) original image, b) Gabor filter, c) thresh holding, d) crowd regions).

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

### A. Experimental Datasets and Results

The data set used in this paper is from two resources. One is from the data set which is used in Multi-Source Multi-Scale Counting in Ref.[20]. We adopt 36 images in which the range of the crowd density is between 648 and 2540 with an average of 1609 individuals per image. The other sample source comes from publicly available web images.

In our experiments, samples are randomly divided into two groups. The training set includes 300 samples. The testing set includes 120 samples, which contain 80 positive samples and 40 negative samples. As Figure 11, the samples contain many sorts of scenes, for example, public areas, places of entertainments, natural scenes, etc. And some public scenes in these source images may belong to a diverse set of events, such as concerts, protests, stadiums, marathons, and pilgrimages. The samples are divided into two classes, including positive samples and negative samples, as Figure 11.

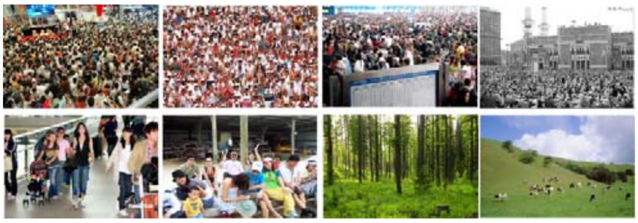


Figure 11. Experimental Samples (in the figure, the four samples in the above row are positive samples, and the other four samples in the below row are negative samples).

This paper adopts SVM to detect the crowd scenes based on three different features respectively, and the results are shown in Table 2. The testing set includes 120 samples, which contain 80 positive samples and 40 negative samples. From the Table 2, the conclusion can be obtained that GIST feature fusing LBP is better effective.

TABLE 2.THE RESULTS OF DETECTION RATE USING SVM

| Feature | LBP | GIST | GIST+ LBP |
|---|---|---|---|
| True positive | 66 | 69 | 73 |
| True negative | 34 | 36 | 38 |
| False positive | 6 | 4 | 2 |
| False negative | 14 | 11 | 7 |
| Precision | 91.7% | 94.5% | 97.3% |
| Accuracy | 83.3% | 87.5% | 92.5% |
| Recall | 82.5% | 86.3% | 91.3% |

Where, the precision rate, accuracy and recall are defined as formula (9), (10) and (11).

$$Precision\ rate = \frac{True\ positive}{True\ positive + False\ positive} \tag{9}$$

$$Recall\ rate = \frac{True\ positive}{True\ positive + False\ negative} \tag{10}$$

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

Table 3 shows the confusion matrix of SVM classier using GIST and LBP features.

TABLE 3. CONFUSION MATRIX OF SVM BASED ON GIST AND LBP

| Class | Class 1 | Class 0 |
|---|---|---|
| Crowd | TP 73 | FN 7 |
| Non-crowd | FP 2 | TN 38 |

### B. *Crowd Regions Locating*

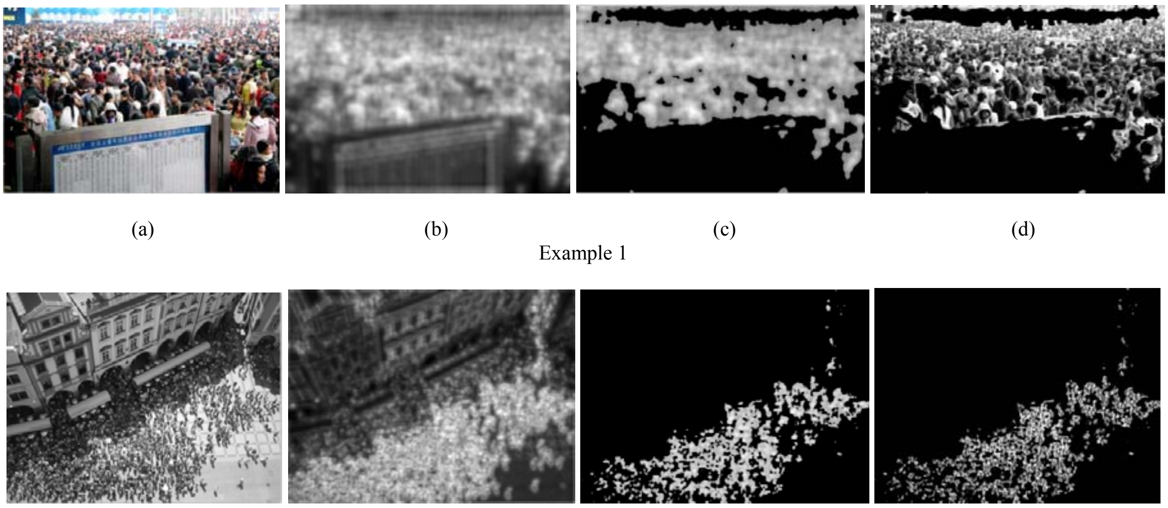
This paper adopts Gabor filtering to enhance crowd regions, and some results are shown in Figure 12. From Figure 12, it can be observed that our approach is effective. The crowd regions are enhanced, as Figure 12(b) of two examples, and the crowd regions after removing non-crowd regions are acceptable, as is shown in Figure 12(d). Therefore, our crowd regions locating approach is feasible for removing non-crowd regions and locating crowd regions.

### C. *Comparisons*

In this section, we compare SVM with BP Neural Network using three different features. Table 4 demonstrates the accuracy under different conditions. SVM is more effective than BP NN. Figure 13 is the column chart based on Table 4, which can intuitively illustrate the difference between three features and two classifiers.

TABLE 4. COMPARISONS OF ACCURACY（%）

| Classifier | BP NN | SVM |
|---|---|---|
| LBP | 81.6 | 83.3 |
| Gist | 85.4 | 87.5 |
| Gist+ LBP | 90.3 | 92.5 |



(a) (b) (c) (d)

Example 1

(a) (b) (c) (d)

Example 2

Figure 12. The Results of Crowd Region Detection.(a)original image, b)Gabor filter, c)thresholding, d)crowd region)

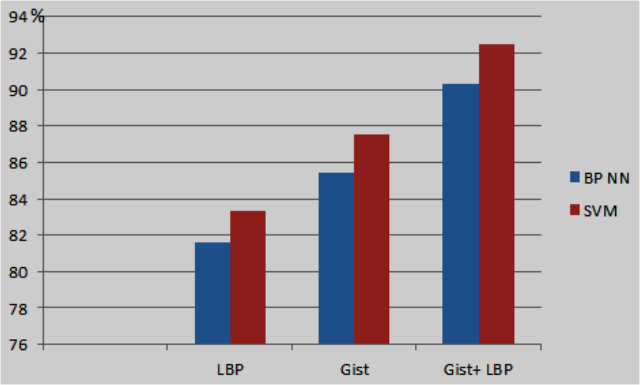


Figure 13. Comparisons of Accuracy

## V. CONCLUSION

This paper presents an approach for high dense crowd scene and crowd region detection in still images. Two important features which include LBP and GIST are chosen, fused and used for SVM classifier whose principal components are extracted and constructed a novel feature vector. Further, we remove and detect local crowd areas using Gabor filter. Our approach can effectively detect crowd regions instead of dividing image into plenty of blocks and classifying them. Experimental results demonstrate that the proposed detection approach is feasible and effective in variable conditions. In the future, it is necessary to improve the precision of the range of crowd regions.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

## ACKNOWLEDGMENT

This investigation is supported by the Project of State Ethnic Affairs Commission of China (No. 14ZNZ019), the Project of Natural Science Foundation of Hubei Province of China (No. 2012FFC13401), and the Fundamental Research Funds for the Central Universities(No. CZY16005).

## REFERENCES

[1] J.Xing, H.Ai, L.Liu, S.Lao, “Robust crowd counting using detection flow”. 18th IEEE International Conference on Image Processing (ICIP),pp. 2061-2064,2011.

[2] J.Gall, V.Lempitsky, “Class-specific hough forests for object detection”. Decision Forests for Computer Vision and Medical Image Analysis. Springer, pp.143-157,2013.

[3] N.Hussain, H.S.M.Yatim, N.L.Hussain, J.L.S.Yan, F.Haron. “CDES:apixel-based crowd density estimation system for masjid al-haram”. Safety Science,vol.49,no.6,pp.824-833, 2011.

[4] A.J.Thangam, P.T.Siva,B.Yogameena. “Crowd count in low resolution surveillance video using head detector and color based segementation for disaster management”. International Conference on Communications and Signal Processing (ICCSP), pp.1905-1909,2015.

[5] Thanh-Sach Le, Chi-Kien Huynh. “Human-Crowd Density Estimation Based on Gabor Filter and Cell Division”. International Conference on Advanced Computing and Applications (ACOMP), pp.157-161,2015.

[6] H.Fradi, X.Zhao, J.L.Dugelay. “Crowd density analysis using subspace learning on local binary pattern”. IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp.1-6,2013.

[7] Z.Wang, H.Liu, Y.Qian, T.Xu. “Crowd Density Estimation Based on Local Binary Pattern Co-Occurrence Matrix”. IEEE International

Conference on Multimedia and Expo Workshops (ICMEW), pp.372-377, 2012.

[8] A.Albiol, M.J.Silla, A.Albiol, J.M.Mossi. "Video analysis using corner motion statistics". IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp.31-38,2009.

[9] Conte, P.Foggia, G.Percannella, M.Vento. "Counting moving persons in crowded scenes". Machine Vision & Applications, vol.24, no.5,pp.1029-1042, 2013.

[10] L. Kratz and K. Nishino. "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models". CVPR, pp.1446-1453,2009.

[11] W. Li, V. Mahadevan, N. Vasconcelos. "Anomaly Detection and Localization in Crowded Scenes". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36,no.1, pp.18-32, 2014.

[12] Y. Yuan, J. Fang, and Q. Wang. "Online Anomaly Detection in Crowd Scenes via Structure Analysis". IEEE Transactions on Cybernetics,vol. 45, no.3,pp.562-575, 2015.

[13] N. Li, X. Wu, D. Xu, H. Guo, W. Feng. "Spatio-temporal context analysis within video volumes for anomalous-event detection and localization". Neurocomputing, vol.155,pp.309-319, 2015.

[14] H. Idrees, I. Saleemi, C. Seibert, M. Shah. "Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images". IEEE Conference on Computer Vision and Pattern Recognition, pp.2547-2554,2013.

[15] A. Bansal, and K. S. Venkatesh. "People Counting in High Density Crowds from Still Images", Computer Science,2015.

[16] T. Ojala, M. Pietikäinen, T. Mäenpää. "Multi resolution gray-scale and rotation invariant texture classification with Local Binary Patterns", IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.24,no.7,pp.971-987, 2002.

[17] A. Oliva, A. Torralba. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope". Intl J of Computer Vision, vol.42,no.3,pp.145-175, 2001.

[18] L. W. Renninger, J. Malik. "When is scene identification just texture recognition?" Vision Research, vol.44,pp.2301-2311, 2004.

[19] J. K. Kamarainen，V. Kyrki，H. KäLviäInen. "Invariance properties of gabor filter-based features--overview and applications". IEEE Transactions on Image Processing, vol,15,no.5,pp.1088-1099,2006.

[20] H. Idrees, I. Saleemi, C. Seibert, M. Shah. "Multi-source Multi-scale Counting in Extremely Dense Crowd Images". IEEE Conference on Computer Vision & Pattern Recognition, vol.9,no.4,pp.2547-2554, 2013.